

Semantic Trading: Agentic AI for Clustering and Relationship Discovery in Prediction Markets

Agostino Capponi¹, Alfio Gliozzo², Brian Zhu¹

¹ Columbia University

ac3827@columbia.edu, bzz2101@columbia.edu

² IBM Research

gliozzo@us.ibm.com

Abstract. Prediction markets allow users to trade on outcomes of real-world events, but are prone to fragmentation with overlapping questions, implicit equivalences, and hidden contradictions across markets. We present an agentic AI pipeline that autonomously (i) clusters markets into coherent topical groups using natural-language understanding over contract text and metadata, and (ii) identifies within-cluster market pairs whose resolved outcomes exhibit strong dependence, including “same-outcome” (correlated) and “different-outcome” (anti-correlated) relationships. Using a historical dataset of resolved markets on Polymarket, we evaluate the accuracy of the agent’s relational predictions. We then synthesize discovered relationships into a simple trading strategy to quantify how discovered relationships translate into actionable strategies. Results show that agent-identified relationships have around 60-70% accuracy, and their induced trading strategies have an average return of ~20% over week-long horizons, highlighting the ability of agentic AI and large language models to uncover latent semantic structure within prediction markets.

Keywords: Agentic AI · Prediction markets · Topic clustering · Relationship discovery · Polymarket · Decentralized finance · LLMs

1 Introduction

Prediction markets are used to aggregate dispersed information into a single, continuously updated price signal ([6],[7],[3]). Across politics, economic, sports, and crypto-native events, these markets often respond faster than traditional polling or commentary because they internalize incentives, uncertainty, and heterogeneous beliefs. Yet the practical experience of trading or analyzing prediction markets reveals a persistent structural problem: the market “question space” is fragmented. Multiple contracts may refer to the same real-world proposition in slightly different wording, while other contracts are economically linked through shared causal drivers, logical implication, or mutual exclusivity. This fragmentation creates search costs for traders, reduces liquidity concentration, and obscures opportunities to hedge risk or exploit mispricings across related contracts.

There is a lack of automatic methods identify these relationships. Traders look for near-duplicates (“Will X happen by November 30?” vs. “Will X happen in November?”), implied complements (“Will X win some election?” vs. “Will X’s margin of victory be above Y%?”), or correlated exposures (“Will ETF X be above Y?” vs. “Will economic indicator A be above B?”). Platforms provide tags and categories, but they are noisy, incomplete, and often too coarse to support systematic discovery. Purely statistical approaches face their own limitations: correlation estimates require long price histories and stable market participation, and they can be confounded by sparse trading, regime changes, and the simple fact that most contracts resolve only once. As a result, the economically most interesting links remain underutilized and under-documented, especially those that are semantically obvious to humans but hard to enumerate at scale.

This paper proposes that agentic AI can function as an automated tool for identifying prediction markets with correlated or interdependent outcomes. We introduce an end-to-end pipeline in which an LLM-based agent reads market text (and optionally, additional context), clusters markets into coherent topical groups, and then performs targeted within-cluster relationship discovery to propose market pairs whose outcomes are likely to be related to each other. The key design choice is to combine *semantic structure* with *statistical validation*: clustering narrows the candidate space to pairs that share meaning, while outcome-based dependence analysis on resolved markets filters relationships to those that are empirically reliable. We focus explicitly on two relationship types that are both interpretable and actionable: same-outcome links and different-outcome links (markets that should resolve identically or oppositely, respectively).

Using a dataset of historically resolved markets, we investigate two questions. First, can an agent accurately predict which market pairs are meaningfully linked, beyond what is achievable with embedding similarity alone or correlation-only screening? Second, do these discovered links translate into a simple, transparent trading edge? To address the second question, we implement a baseline pair strategy that operationalizes the agent’s outputs into tradeable rules over pre-resolution price paths. This lets us measure not only predictive performance (precision/recall of relationship labels), but also economic outcomes (e.g. PnL, Sharpe ratio).

These results suggest that agentic AI can act as a scalable discovery layer for prediction markets, connecting fragmented contract listings into a structured graph of propositions that can be queried, validated, and traded. Our contributions are threefold:

1. We introduce a novel agentic pipeline for organizing and linking prediction-market contracts that combines semantic clustering, interpretable cluster labeling, and within-cluster relationship discovery to produce actionable “same” versus “different” pair hypotheses.
2. We propose an evaluation protocol grounded in resolved markets that measures both cluster-level and pooled accuracy under repeated trials, providing a benchmark that can be used to evaluate future AI models.

3. We assess economic relevance using a simple, transparent leader–follower execution rule that trades only after the leader market resolves, allowing us to quantify how discovered relationships translate to out-of-sample trading strategies.

2 Background

2.1 Prediction Markets

Prediction markets are exchanges where contracts pay off based on the realization of a future event (e.g., “Candidate X wins”), so the contract price can be interpreted as an aggregated forecast or implied probability. A large body of economics research argues that, when contracts are well-defined and trading is sufficiently liquid, markets can efficiently combine dispersed information, incentives, and heterogeneous beliefs into a single continuously updated signal. Canonical syntheses by [7] frame prediction markets as an information aggregation mechanism and survey evidence that market prices can be accurate and responsive to new information.

Modern prediction markets have roots in both academic experiments and public platforms. The Iowa Electronic Markets (IEM), operated by the University of Iowa since 1998, have long served as a prominent research testbed for real-money event contracts, particularly in elections and macro-style questions. Over time, consumer-facing venues emerged, most famously Intrade in the 2000s, demonstrating broad public appetite for trading forecasts, while also highlighting the regulatory and operational fragility of such platforms.

The foundational architecture of modern automated prediction markets owes much to the work of Robin Hanson [5]. In the early 2000s, Hanson identified a critical liquidity failure in nascent prediction markets: without a thick order book of human buyers and sellers, markets remained “thin,” resulting in wide spreads and poor information discovery. To solve this, Hanson invented the Logarithmic Market Scoring Rule (LMSR), that uses a cost function to determine prices and trading outcomes. This friction led to a divergence in market design. While Decentralized Finance (DeFi) exploded with Constant Product Market Makers (CPMM) like Uniswap ($x \cdot y = k$) [1], prediction markets found these generic AMMs ill-suited for binary outcomes, as they do not cap prices at \$1.00 or handle the expiration of assets effectively. Consequently, modern platforms like Polymarket abandoned the pure AMM model in favor of limit order books.

2.2 Institutional Details: Polymarket

Polymarket is a blockchain-native prediction market, built on the Polygon network, whose markets are tokenized on Polygon and collateralized in the USDC stablecoin. Outcome positions are represented as “YES” and “NO” shares implemented via Gnosis’ Conditional Token Framework (CTF), where outcome shares

are ERC-1155 tokens derived from a condition and collateral token.³ This structure allows users to split collateral into outcome tokens and later redeem/merge positions based on the final resolution.

Under the CTF, multi-candidate or multi-option questions are typically implemented as a collection of separate binary markets under a shared *event market*. For multi-outcome events, Polymarket enables a *negative-risk* (NegRisk) structure that links the markets within the event to improve capital efficiency. In a NegRisk event, at most one market can resolve YES, and the protocol supports a `convert` action where holding a NO share in any one market can be transformed into 1 YES share in all other markets (reflecting the logical complement structure across mutually exclusive outcomes).

For trading, Polymarket operates a hybrid limit order books: order matching/ordering is handled by an operator off-chain, while settlement occurs on-chain and non-custodially via signed order messages and a custom exchange contract supporting atomic swaps between outcome tokens and collateral. The platform also exposes developer-facing infrastructure (e.g., WebSocket market/user channels) to stream near real-time views of orders and trades, enabling algorithmic execution and data collection.

Market resolution relies on an oracle process. Polymarket documentation describes a workflow where an outcome is proposed with an associated bond, can be disputed, and finalizes via an escalation mechanism; Polymarket leverages UMA’s Optimistic Oracle for settlement, with disputed assertions escalated to UMA’s Data Verification Mechanism (DVM).⁴ This “optimistic” design aims to keep most resolutions fast and inexpensive while preserving a credible backstop for contentious events.

2.3 Related Work

A small but rapidly growing set of projects is blending large language models with prediction markets across three roles: *participation*, *distribution*, and *evaluation*. On the participation side, developer toolkits such as *Polymarket Agents*⁵ provide open-source scaffolding for building trading-capable agents, including market/API integration and retrieval-augmented data sourcing to ground decisions. Within crypto-native stacks, several efforts push further toward “agent economies” in which autonomous services continuously scan markets, form beliefs, and express them through trades. For example, *Olas Predict*⁶ frames prediction as the emergent output of agent trading activity, while implementations such as Valory’s *Trader*⁷ service operationalize an agent loop that delegates AI tasks (e.g., research or reasoning) to auxiliary services.

³ <https://github.com/gnosis/conditional-tokens-contracts>

⁴ <https://docs.uma.xyz/>

⁵ <https://github.com/Polymarket/agents>

⁶ <https://olas.network/agent-economies/predict>

⁷ <https://www.valory.xyz/>

Complementing trading automation, ecosystem work around *Presagio*⁸ (often described as “Omen 2.0”) positions prediction markets as a live battleground for AI-driven forecasting strategies embedded directly into market infrastructure. On the distribution side, consumer AI products have begun surfacing market-implied probabilities inside search/answer experiences (e.g., Perplexity integrating Kalshi odds), effectively treating prediction markets as a real-time “belief layer” for end users. Finally, prediction-market-style questions are increasingly used to *benchmark* agentic forecasting: Metaculus runs recurring AI forecasting tournaments, while newer evaluations such as *Prophet Arena*⁹ test agents on unresolved real-world events with probabilistic scoring and, in some cases, return-based metrics, tightening the loop between semantic reasoning under uncertainty and measurable performance in the wild.

2.4 Agentics Framework

Agentics [4] is an agentic AI framework aimed at constructing structured, data-oriented workflows with LLMs. Its central idea is a data-centric programming model in which agents are specified by typed schemas—called *ATypes*—and perform *logical transduction*: a schema-governed transformation from one typed object to another using an LLM.

In Agentics, an agent is modeled as a stateless transducer operating over an *Agentic Structure*:

$$AG := \{atype : \Theta, states : List[atype]\},$$

i.e., a collection of typed objects (states) that all share a common $atype \in \Theta$, where Θ is the universe of types. Given a set of input states, the Agentics programming model supports asynchronous, schema-constrained generation of new typed states.

The primitive operation is the logical transduction

$$y := AG[Y] \ll x,$$

which converts an input state x of type X into an output state y of type Y . Crucially, each field of y is produced from x in a way that respects the semantic constraints encoded by the target type schema, with the LLM serving as the transduction mechanism.

Agentics also includes a built-in clustering feature using vector space models: given an **AG** collection of typed states, it can embed one or more semantically meaningful textual fields into a shared latent space (i.e., represent each state as a vector) and then group states into clusters based on geometric proximity (similarity) in that space. This aligns with the classic vector-space view of semantic organization, where objects are mapped to vectors and cluster structure can be summarized via centroids and salient terms.

⁸ <https://presagio.pages.dev/>

⁹ <https://www.prophetarena.co/>

To scale beyond single-step transformations, Agentics provides asynchronous map-reduce style operators and quotient structures that enable principled abstraction. These features make it well-suited for financial analytics workflows that benefit from robustness, interpretability, and structured outputs.

3 Methodology

In this section, we outline the methodology used for the agentic workflow used to cluster and discover relationships among prediction markets using Agentics. Given a list of prediction markets on Polymarket, the goal of the AI agent is to return a list of market clusters and for each cluster, a list of pairs of markets likely to resolve to the same or different outcome.

3.1 MCP Tools

To allow AI agents to access, transduce, and perform reasoning over a diverse range of data sources, we leverage Model Context Protocol (MCP) [2] tools in the main workflow stages. Each tool analyzes typed inputs (market objects and/or cluster objects) and emits typed outputs as CSV or JSON artifacts.

Clustering MCP. The *Clustering MCP* groups markets into topical clusters using a vector-space model over market text. Concretely, we ingest a list of markets as typed states (e.g., a schema with a single `question` or `market` field), embed each market into a shared latent space, and partition the set into K clusters via geometric proximity in embedding space. In our implementation, we set $K \approx \lfloor N/10 \rfloor$ (where N is the number of markets in the batch) to target clusters that are small enough for downstream pair discovery yet broad enough to capture paraphrases and closely related propositions. The tool outputs (i) a cluster assignment for each market and (ii) a cluster manifest (a list of questions per cluster), which we maintain as per-cluster CSV files for downstream processing.

Cluster AType
<pre>class Cluster(BaseModel): markets: str category: str</pre>

Cluster Labeling MCP. The *Cluster Labeling MCP* assigns a category label to each discovered cluster using schema-constrained LLM generation. The input is a cluster object containing a list of market questions (serialized as text), and the output augments the same object with a single categorical label. We use a lightweight closed taxonomy (e.g. ‘politics’, ‘macro’, ‘finance’, ‘earnings’, ‘crypto’, ‘tech’, ‘sports’, ‘culture’, ‘other’) to ensure consistency across runs and to support stratified evaluation by domain. The tool is implemented as a typed

self-transduction that enforces that exactly one label is produced for every cluster, yielding an auditable mapping from cluster contents to category.

Relationship Discovery MCP. The *Relationship Discovery MCP* proposes economically meaningful links between markets within a cluster. Given a cluster of markets (optionally augmented with `news_summary`), the tool runs an asynchronous reduce-style typed generation that emits a list of candidate pairs. Each emitted `MarketRelation` contains: (`question_i`, `question_j`) copied verbatim from the cluster, a boolean `is_same_outcome` indicating whether the two markets are predicted to resolve identically (vs. oppositely), a calibrated `confidence_score` in [0, 1], and a brief `rationale`. We implement this as a constrained “propose-and-structure” step: the LLM is instructed to surface only pairs whose outcomes are *very likely* to be related, and to commit to a direction (same vs. different) with an explicit confidence. The resulting relation lists are then joined back to the historical dataset to evaluate correctness against realized resolutions, enabling both prediction metrics (e.g., precision by confidence threshold) and trading/backtest modules to consume the same standardized relation artifact.

SingleMarket AType

```
class SingleMarket(BaseModel):
    question: str
    market_start_time: str
    market_end_time: str
```

MarketRelation AType

```
class MarketRelation(BaseModel):
    question_i: str
    question_j: str
    is_same_outcome: bool
    confidence_score: float
    rationale: str
```

MarketRelationList AType

```
class MarketRelationList(BaseModel):
    relations: list[MarketRelation]
```

Agentic Workflow. Using these MCPs, we then combine them into a single agentic workflow. We provide the full decision prompts in the appendix.

1. Group prediction markets into clusters.
2. Categorize and obtain additional context (news, search) for the markets in each cluster.

3. Transduce a cluster, specified by a list of `SingleMarket` ATypes, into a `MarketRelationList` AType:

$$(AG[SingleMarket1], \dots, AG[SingleMarketN]) \ll AG[MarketRelationList]$$

3.2 Data Sets

We start from a list of the prediction markets with the most volume that were resolved between April to July 2025. We collect data using Dune Analytics, specifically focusing on the following fields:

- `event.market.name`: name of the event being traded; this is “single market” for markets that are not of the negative risk type.
- `question`: an event market’s corresponding question.
- `market.start.time`: timestamp at which the market opens for trading.
- `market.end.time`: latest timestamp at which the market is open for trading; the market either resolves before or settles at the last traded price otherwise.
- `resolved.on.timestamp`: timestamp at which the UMA oracle decides on an outcome.
- `outcome`: indicates whether the market resolves to YES or NO.

We narrow our sample using the following criteria, with reasoning provided.

- **Binary outcomes only.** We restrict attention to binary markets (YES/NO) to keep both the ground-truth labeling and downstream evaluation well-defined and comparable across contracts. Binary resolution yields an unambiguous realized outcome, allows a consistent definition of “same” versus “different” across market pairs, and avoids the additional modeling choices required for multi-outcome, or negative-risk markets. For example, some outcomes in multi-outcome markets are not mutually exclusive, and some multi-outcome markets have ordinal rather than nominal categories (e.g. bucketing a candidate’s margin of victory in an election).
- **Markets longer than one week.** We exclude markets whose active trading window, i.e. time from `market.start.time` to `market.end.time` is shorter than seven days. Short-duration markets offer limited time for the informational aggregation process to operate, and also complicate backtesting because entry and exit opportunities are highly sensitive to exact timestamps and latency, and realized performance can be driven by event-timing artifacts rather than the semantic relationship signals we aim to study. By filtering these markets, we focus the analysis on contracts with sufficient horizon for meaningful relationship discovery. Many of these short-duration markets fall into the category of head-to-head sports matchups and betting on cryptocurrency prices at the end of a given hour, both of which have limited semantic data for our agent to incorporate.
- **Month-based splitting via question text.** To evaluate generalization under distribution shift, we construct time-sliced samples by filtering markets whose question text explicitly references a given month (e.g., “in April”).

Sample	Count	Volume (USD)	Duration (days)		
			Mean	Std.	Mean
All	778	400 k	2.00 M	41.3	33.7
April 2025	217	308 k	941 k	41.3	35.1
May 2025	256	507 k	2.60 M	39.0	32.4
June 2025	190	351 k	1.37 M	37.5	28.9
July 2025	43	1.08 M	4.41 M	57.2	39.5

Table 1: Summary statistics for market liquidity and horizon. Volume is in USD (as reported by `volume_usd`); duration is in days.

This provides a simple, reproducible proxy for the market’s event horizon that is available purely from the contract text, without relying on platform-specific metadata that may not be reflective of the actual market (e.g. some markets set the `market_start_time` far in advance and the `market_end_time` far in the future relative to when the event is likely to resolve (e.g. 2025 elections where trading ends in November 2026). Month-based slicing also yields controlled cohorts with similar temporal structure, enabling cleaner comparisons of clustering quality, relationship precision, and trading performance across periods while reducing contamination from repeated re-listings of essentially the same proposition across adjacent time windows.

Summary statistics for the markets selected at the end of our filtering process are shown in Table 1.

3.3 Evaluation

Cluster Accuracy. We evaluate *cluster-level accuracy* by measuring how often the Relationship Discovery MCP’s predicted relation direction (`is_same_outcome`) matches the realized relationship implied by resolved outcomes. For each proposed pair (i, j) we derive ground truth from the resolved binary labels:

$$\text{ground_truth}(i, j) := \mathbb{1}\{\text{outcome}_i = \text{outcome}_j\},$$

and mark a prediction as correct when

$$\text{is_correct}(i, j) := \mathbb{1}\{\text{is_same_outcome}(i, j) = \text{ground_truth}(i, j)\}.$$

To focus analysis on high-confidence links intended to be actionable, we filter to candidate pairs with `confidence_score` ≥ 0.5 . We then report two aggregates: (i) a *cluster average*, computed as the mean of per-cluster accuracies over clusters that contain at least one eligible (high-confidence) pair, and (ii) an *overall average*, computed by pooling all eligible pairs across clusters and taking the fraction correct. This dual reporting distinguishes whether performance is broadly consistent across topical clusters (cluster average) versus dominated by a small number of large clusters (overall average).

Trading Strategy. We translate discovered relationships into a simple event-driven, single-leg strategy designed to test whether the inferred links provide actionable information rather than merely post-hoc correlation. For every predicted pair (i, j) we *wait until market i resolves* (at `resolved_on_timestamp_i`) and then enter a position in market j at the first observed price tick strictly after that time. Let p_Y denote the YES price in market j at entry and $p_N := 1 - p_Y$ the implied NO price. The predicted relation `is_same_outcome` together with the observed resolution of market i determines which side of j to buy: we buy YES in j iff `outcome_i = YES` and `is_same_outcome` is true (or `outcome_i = NO` and `is_same_outcome` is false); otherwise we buy NO.

We apply two conservative filters to avoid implausible trades: (i) an *entry extremeness* filter that skips trades when the selected leg price is too close to 0 or 1 (we use a cutoff of 0.1), and (ii) a *final-price quality* filter requiring that market j has a decisive terminal price (last observed price within 0.1 of 0 or 1), which mitigates missing/ambiguous terminal series. Profit-and-loss is computed under a unit-stake binary payoff model: if the selected leg ultimately resolves correctly, PnL is $1 - \text{entry_price}$; otherwise it is $-\text{entry_price}$. We report the trade count and return on investment (ROI) (total gain divided by total invested) aggregated across all eligible predicted pairs.

4 Results

For each month in our dataset (April, May, June, and July of 2025), we deploy the agent for 30 trials. To illustrate how our pipeline works, we first provide two representative examples.

- Example 1:

- `question_i`: Will Trump increase tariffs on Canada before May?
- `question_j`: Will Trump remove tariff on Canada before May?
- `is_same_outcome`: False
- `confidence_score`: 0.95
- `rationale`: These two questions represent mutually exclusive actions regarding Trump’s tariff policy on Canada. He cannot simultaneously increase and remove tariffs on the same entity within the same timeframe. If one outcome is ‘yes’, the other must logically be ‘no’.

- Example 2:

- `question_i`: Trump tariff on EU alcohol before May?
- `question_j`: Will the EU impose tariffs on the U.S. before May?
- `is_same_outcome`: True `confidence_score`: 0.85
- `rationale`: If Trump imposes tariffs on EU alcohol, it is highly probable that the EU would retaliate by imposing its own tariffs on U.S. goods. This represents a classic trade war escalation scenario. Conversely, if Trump does not impose tariffs, the EU is less likely to initiate tariffs on the U.S.

These two examples highlight the range of relationships the agent is attempting to capture. In Example 1, the pair is a near-textbook “different outcome” case: “increase tariffs” versus “remove tariffs” on the same target within the same time window are directionally opposed policy actions, so treating them as mutually exclusive is a strong, high-confidence call. Example 2 is qualitatively different: it is not a logical complement, but a plausible causal linkage consistent with tit-for-tat dynamics in tariff disputes, where an initial tariff action can trigger retaliation in response. Thus, the “same outcome” label is best interpreted as a high-probability correlation hypothesis rather than a strict equivalence, which matches the lower confidence score (0.85) relative to Example 1.

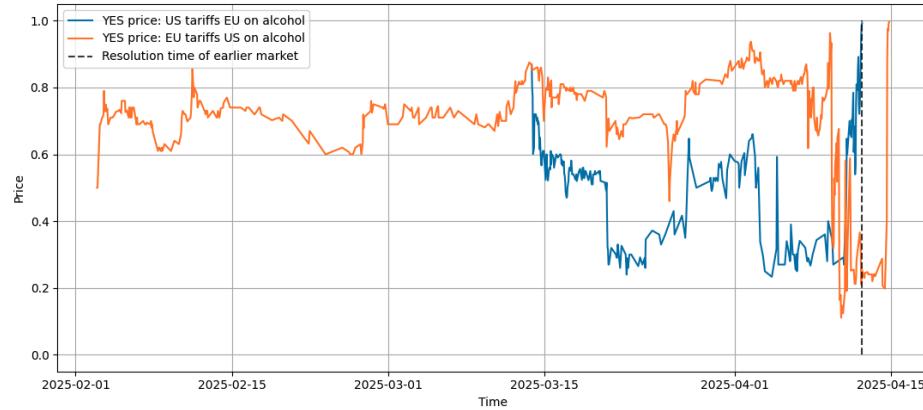


Fig. 1: Leader–follower price dynamics for a discovered pair.

Figure 1 shows the YES-price time series for two related markets. The vertical dashed line marks the resolution time of the leader (earlier) market; the trading rule enters the follower market at the first tick after this timestamp, using the predicted relationship direction to choose YES vs. NO. Since the agent identified these markets as having the same outcome, which turns out to be true, the profit of the trading strategy is one minus the price at the time of buying (~ 0.25).

In Figure 2, we represent the discovered relationships in the form of a graph. Each node is a Polymarket binary question (shown as the node label) and node color indicates the realized resolution (YES vs. NO). Black-colored edges indicate the agent predicted for the nodes on either edge to have the same outcome, while blue-colored edges indicate different predicted outcomes based on semantics. The pattern of the edge indicates whether or not the agent’s prediction was true in reality: solid lines indicate correct categorizations, while dashed lines indicate mistakes. One insight gained from this graph perspective is that the agent’s choices are logically consistent with respect to the graph structure. Predicted “same” and “different” edges tend to satisfy triadic sign constraints (e.g., avoiding inconsistent triangles such as $A \sim B$, $B \sim C$, but $A \not\sim C$).

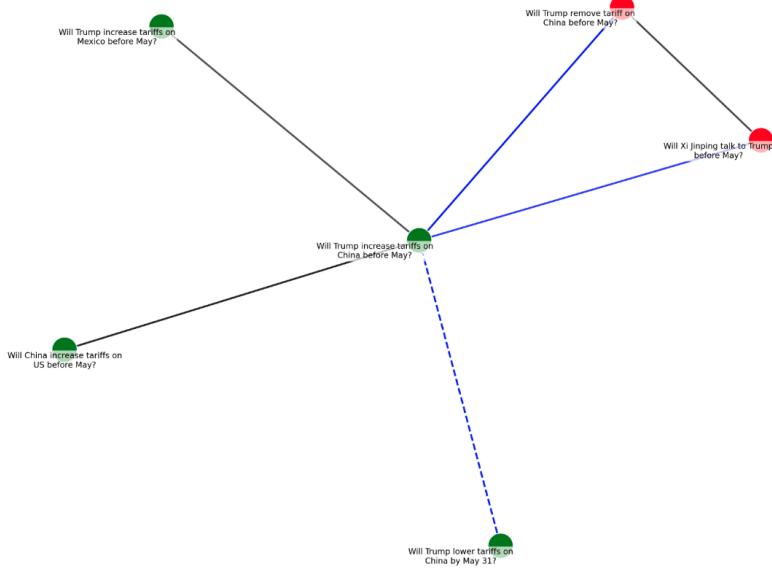


Fig. 2: Relationship graph for a cluster relating to tariff policies.

Relationship prediction accuracy. Tables 2 and 3 summarize accuracy across monthly cohorts over repeated trials. We report two complementary metrics. `cluster_accuracy` averages prediction correctness at the cluster level, weighting clusters equally, while `overall_accuracy` pools all predicted pairs and measures the fraction correct. Across months, `cluster_accuracy` is materially above chance in April–June, with mean values of 63.9%, 59.3%, and 62.1%, respectively, but degrades in July (41.2% mean), indicating either a more challenging question distribution or fewer high-quality, unambiguous relationship opportunities in that cohort. The pooled metric tracks a similar pattern, with the strongest monthly performance in June (72.6% mean) and a drop in July (50.9% mean). We also observe non-trivial variance across trials within month (e.g., July $\sigma \approx 11.0$ for `cluster_accuracy`). Accuracy figures, separated by cluster categories (e.g. politics) are located in the appendix.

Trading performance. Table 4 reports the distribution of backtested returns from the simple leader–follower strategy driven by the agent’s predicted relationships. The strategy is profitable on average in April–June, with mean ROIs of 24.8%, 8.7%, and 47.5% respectively, and a particularly strong median in June (45.7%). July is negative on average (mean ROI –12.3% and median –23.5%), mirroring the degradation in relationship accuracy. Importantly, return volatility is substantial across all months (standard deviations between 25.7 and 41.6 percentage points), reflecting both the small number of trades per trial and the fact that a few misclassified high-confidence pairs can dominate PnL. Nonetheless,

Cluster Accuracy (%)	2025-04	2025-05	2025-06	2025-07
Mean	63.9	59.3	62.1	41.2
Std.	8.3	7.5	6.2	11.0
Min	35.3	45.8	49.2	24.4
25%	60.9	53.7	58.6	33.3
Median	65.5	59.4	63.1	41.7
75%	68.7	64.3	65.8	47.0
Max	78.4	73.3	72.8	78.6

Table 2: Descriptive statistics for `cluster_accuracy` (in percentage points) by month.

Overall Accuracy (%)	2025-04	2025-05	2025-06	2025-07
Mean	62.0	60.7	72.6	50.9
Std.	5.5	6.3	7.1	10.0
Min	52.3	51.1	51.7	33.3
25%	58.8	56.6	71.8	42.9
Median	61.3	58.9	73.5	50.0
75%	66.7	65.2	76.9	56.7
Max	74.6	76.3	82.1	70.0

Table 3: Descriptive statistics for `overall_accuracy` (in percentage points) by month.

the alignment between higher accuracy months (especially June) and higher ROI suggests that the discovered semantic relationships are not merely descriptive, but can translate into economically meaningful signals under a simple, transparent execution rule.

The typical delay between resolution of the “leader” and “follower” resolutions is typically on the order of one to two weeks in our data (a full table is provided in the appendix), reflecting that many linked markets refer to similar propositions but with different cutoff dates. As a result, the strategy can be interpreted as testing the speed at which information revealed by one resolved contract propagates into the pricing of related, still-open contracts.

Interpreting month-to-month variation. The cross-month dispersion highlights that relationship discovery is sensitive to the underlying markets resolving around a specific timeframe. Months dominated by clearer logical structure (e.g., unambiguous “same proposition” paraphrases, or clean complements) support higher precision and more stable trading performance, while months with greater semantic ambiguity can lead to noisier pair proposals and weaker PnL. This observation motivates two practical takeaways for deployment: calibrating confidence thresholds and incorporating uncertainty-aware filters are crucial for stabilizing outcomes, and adding external context (news/search) should be most

Returns (%)	2025-04	2025-05	2025-06	2025-07
Mean	24.8	8.7	47.5	-12.3
Std.	37.5	25.7	31.4	41.6
Min	-26.8	-45.7	-22.1	-100.0
25%	-1.0	-10.7	23.2	-34.6
Median	16.5	6.3	45.7	-23.5
75%	43.3	22.7	60.4	7.5
Max	121.4	84.8	137.9	78.6

Table 4: Descriptive statistics for ROI of the trading strategy (in percentage points) by month.

valuable precisely in the months and domains where question phrasing under-specifies the real-world event.

5 Conclusion

This paper showed how agentic AI can serve as a practical discovery layer for prediction markets by converting a fragmented universe of contracts into a structured map of topics and economically meaningful links. Using an end-to-end pipeline that combines topic clustering, interpretable cluster labeling, and within-cluster relationship discovery, we evaluated relationship predictions against resolved outcomes. We found that high-confidence links reliably identify “same” and “opposite” market pairs beyond text-only and correlation-only baselines. Translating these links into a simple, transparent trading rule yielded a market-to-market signal that can reduce directional exposure while capturing pricing inefficiencies created by duplication, ambiguity, or delayed cross-market incorporation of information. While limitations remain and could be addressed by future directions, particularly around distribution shift across time, handling multi-outcome markets, and incorporating additional context such as news, our results suggest a clear path toward agent-assisted market infrastructure in the form of relationship graphs that improve search, hedging, and risk management and that can be continuously audited using resolved-market regression tests.

References

1. Hayden Adams, Noah Zinsmeister, and Dan Robinson. Uniswap v2 Core, March 2020. Technical whitepaper. URL: <https://app.uniswap.org/whitepaper.pdf>.
2. Anthropic. Introducing the Model Context Protocol, November 2024. URL: <https://www.anthropic.com/news/model-context-protocol>.
3. Joyce E. Berg and Thomas A. Rietz. Prediction Markets as Decision Support Systems. *Information Systems Frontiers*, 5 (1):79–93, January 2003. doi:10.1023/A:1022002107255.
4. Alfio Gliozzo, Naweed Khan, Christodoulos Constantinides, Nandana Mihindukulasooriya, Nahuel Defosse, Gaetano Rossiello, and Junkyu Lee. Transduction is All You Need for Structured Data Workflows. *arXiv preprint arXiv:2508.15610*, August 2025. URL: <https://arxiv.org/abs/2508.15610>, doi:10.48550/arXiv.2508.15610.
5. Robin Hanson. Combinatorial Information Market Design. *Information Systems Frontiers*, 5(1):107–119, January 2003. doi:10.1023/A:1022058209073.
6. Georgios Tziralis and Ilias Tatsiopoulos. Prediction Markets: An Extended Literature Review. *The Journal of Prediction Markets*, 1 (1):75–91, 2007. URL: <https://www.ubplj.org/index.php/jpm/article/download/421/452>.
7. Justin Wolfers and Eric Zitzewitz. Prediction Markets. *Journal of Economic Perspectives*, 18(2):107–126, Spring 2004. URL: <https://www.aeaweb.org/articles?id=10.1257/0895330041371321>, doi:10.1257/0895330041371321.

A Full Prompts

Clustering MCP: This stage does not use an LLM prompt; clustering is invoked deterministically via the built-in vector-space clustering call.

Cluster Labeling MCP: Prompt / Instructions

- Assign the cluster of markets to one of the following categories: ‘politics’, ‘geopolitics’, ‘elections’, ‘economy’, ‘finance’, ‘earnings’, ‘crypto’, ‘tech’, ‘sports’, ‘culture’, ‘other’. You must assign a category.

Relationship Discovery MCP: Prompt / Instructions

- Given a list of bets expressed as questions, find pairs of bets whose outcomes are very likely to be related to each other.
- For each pair you propose, create a new `MarketRelation` class and fill out the following:
 - In the `question_i` and `question_j` fields, output the questions in the relationship. Output the questions exactly as they are given.
 - In the `is_same_outcome` field, output `True` if outcomes are likely to be the same (both yes or both no), `False` if outcomes are likely to be different (one yes and one no). You must provide a boolean.
 - In the `confidence_score` field, provide a score between 0 and 1 indicating how confident you are about the relationship. You must provide a confidence score.
 - In the `rationale` field, justify why you chose these bets and the relationship you assigned to them. You must provide a rationale.

B Additional Tables and Figures

Delay (days)	2025-04	2025-05	2025-06	2025-07
Mean	11.63	13.89	15.53	12.56
Std.	4.23	1.99	3.84	9.14
Min	3.89	9.08	8.23	4.52
25%	9.60	13.12	12.96	7.06
Median	12.28	14.32	15.92	11.66
75%	14.13	15.14	18.27	14.44
Max	20.64	17.79	23.27	55.01

Table 5: Descriptive statistics for the duration between the leader and follower market resolution times, group by month.

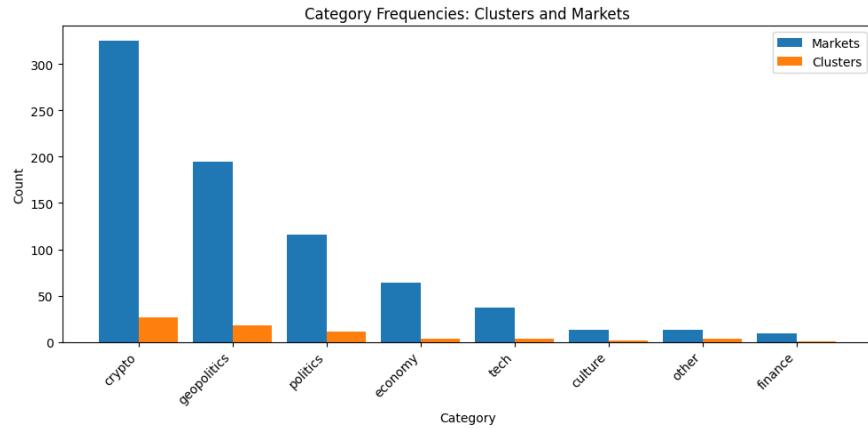


Fig. 3: Category frequencies for clusters and markets. For each category, the blue bars report the number of underlying markets (questions), while the orange bars report the number of clusters assigned to that category.

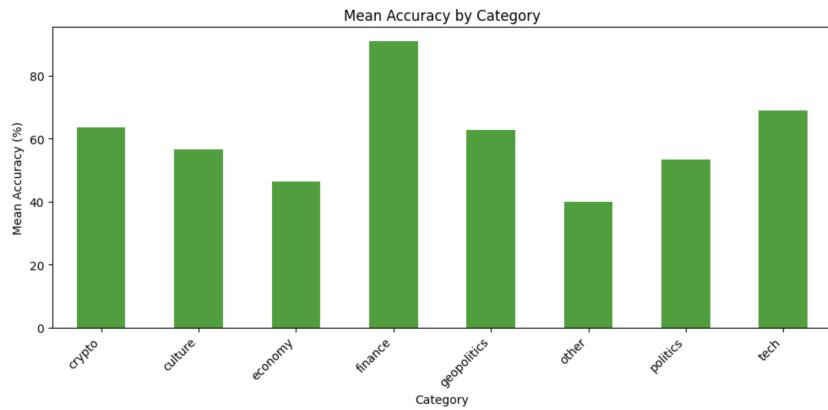


Fig. 4: Mean relationship-prediction accuracy by category. Bars show the average accuracy (percent correct) of the agent's predicted relationships, aggregated over all clusters within each category label.