PolySignal-BTC5: Multi-Agent Debate Prediction System for Bitcoin 5-Minute Prediction Market

Lance Yi

Abstract

Can multiple AI models argue with each other to get to a result beating the crowds' choice in short term prediction market? PolySingal-BTC5 is trying to testify this by deploying three AI agents: a momentum analyst, a contratian, and a judge decides who is more convicing on Bitcoin's short window prediction market.

After 191 consecutive predictions during a single day 16-hour running session, our system achieves **57.6% directional accuracy** and a **Brier score of 0.2474**. This has outperformed both the coin-flip baseline and the crowds' preference. A simulated account returned **+$13,321 (+13.32% ROI)** on a $100,000 account with $500 spent on each bet. The most intereting finding is: when the judge sides entirely with the momentum analyst, accuracy rises to 60.3%; when the contrarian wins, it falls to 25.0%. This shows that the debate's value not only depends on the debators but more important with the judging agent.

1. Introduction

1.1 Motivation

Short-term prediction markets are the most active markets with the most volatilities of the wide prediction market ecosystem, which have attracted millions of traders to invest. However, most of them rely heavily on instinct rather than rational analysis, especially for short windows. While some individual traders may win several rounds, but luck is not a repeatable strategy. This makes short-term markets both an interesting playground and target for AI: unlike image recognition or language translation, where the problem are statci, financial prediction markets actively adapt to any exploitable pattern.

In financial markets, at a 5-minute level, individual price moves are affected by different noises: order flow imbalances, liquidity gaps, and temporary momentum effects that decay within seconds. The singal cannot turn to ratio always, but the feedback loops are faster and can translate directly into real money.

But prediction markets offer a unique laboratory environment for this problem. Unlike traditional financial markets, prediction market provides only binary outcomes determined by objective oracle data. This has removied most of ambiguity about what "correct" means.

Polymarket's Bitcoin Up/Down 5-minute series resolves deterministically via the Chainlink oracle, creating a clean experimental setting where an AI agent's probability estimates can be rigorously evaluated against verifiable ground truth.

The main research question driving this work is: **Can a multi-agent adversarial LLM debate system extract directional signal from structured market data beyond what is already encoded in crowd odds?** If prediction market prices already aggregate all publicly available information — as the Efficient Market Hypothesis predicts (Fama 1970) — then no AI system should outperform the crowd. If LLMs can identify reasoning patterns invisible to the aggregate crowd, they should show measurable edge.

1.2 Research Questions

This paper addresses two major questions:

1. Does a multi-agent adversarial debate architecture achieve directional accuracy above the coin-flip and crowd-odds baselines on 5-minute BTC prediction markets?
2. Is the LLM's stated confidence score calibrated — do 55% confidence predictions succeed approximately 55% of the time?

2. Background

2.1 Prediction Markets

Prediction markets are exchange-traded contracts that pay out based on the result of a future event. Unlike traditional financial activities, prediction market contracts have a settled payoff which typically ranges from $0 to $1 for each share. The price of a prediction contract at any point in time represents the crowd's aggregate probability estimate and preference for the event to actually happen.

The theoretical foundation for prediction markets as information aggregation mechanisms comes from Hayek's (1945) dispersed knowledge framework and Galton's (1907) observation that crowd averages often outperform individual experts. Subsequent empirical work has validated prediction market accuracy across domains including elections (Wolfers and Zitzewitz 2004), sports outcomes (Spann and Skiera 2009), and corporate earnings (Chen and Plott 2002).

Polymarket is one of the largest and earliest decentralized prediction market platforms. It was operated on the Polygon blockchain (Polymarket 2024). It lists thousands of markets across politics, sports, finance, and entertainment. The platform uses an Automated Market Maker (AMM) model, where token prices adjust continuously as the moment when traders buy and sell.

2.2 Polymarket's Bitcoin 5-Minute Up/Down Markets

Polymarket's BTC Up/Down 5-minute market runs every five minutes, 24 hours a day, 7 days a week. And the question of that market is:

"Will the Bitcoin price at [window_end] be higher than at [window_start]?"

Market mechanics: - The opening reference price is the Bitcoin price at the start of the 5-minute window - The closing price is the Chainlink oracle price at window end - If `close_price >= open_price`, the market resolves UP; otherwise DOWN

Why this market: 1. **Binary outcomes** — no ambiguity about what "correct" means 2. **High frequency** — 288 markets per day provide dense feedback loops 3. **Crowd signal** — token prices directly encode the aggregate probability estimate 4. **Real stakes** — live market participants with financial incentives ensure the crowd price is informative

3. Related Work

LLM financial forecasting: Lopez-Lira and Tang (2023) and Xie et al. (2023) show LLMs carry predictive signal in equity markets and financial QA tasks, but both operate at daily or longer horizons where news and fundamentals drive returns. At 5-minute resolution, only price action and crowd odds are available — a signal environment their work does not address.

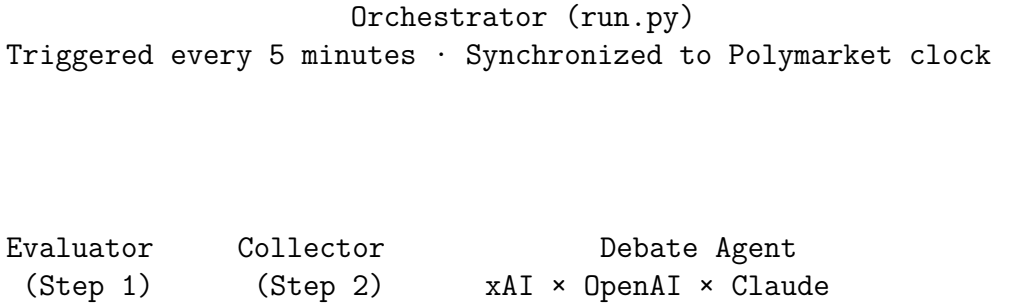
Prediction market calibration: Tetlock (2005) and Mellers et al. (2015) establish that a forecaster's stated confidence should match their realized accuracy, a standard applied extensively to human forecasters but rarely to LLMs in live markets. This work applies that same standard directly to LLM confidence scores using Brier score and calibration bins.

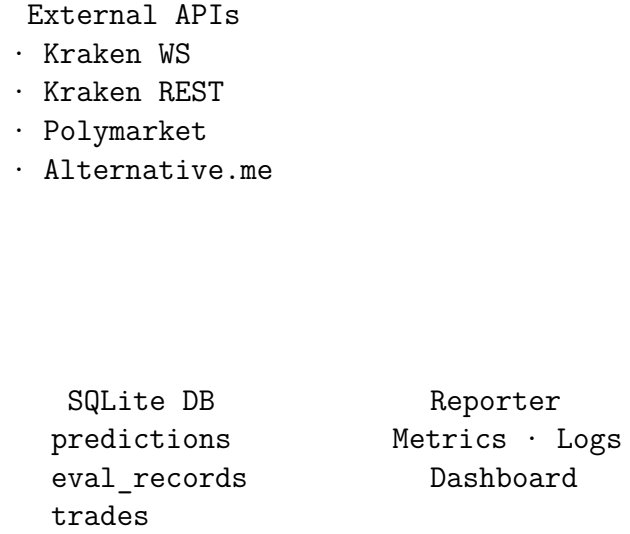
Multi-agent debate and ensembles: Du et al. (2023) and Chan et al. (2023) show multi-agent debate improves LLM reasoning, while ensemble methods (Brown et al. 2020) reduce variance by combining identical models. PolySignal-BTC5 differs by assigning heterogeneous models opposing analytical roles — momentum and contrarian — and evaluating them against live market ground truth that resolves every five minutes.

4. System Architecture

4.1 Overview

PolySignal-BTC5 is a multi-agent system that operates on a fixed 5-minute session to Polymarket's BTC Up/Down market windows. Each cycle consists of four sequential stages: evaluation, collection, prediction, and reporting. They are orchestrated by a central coordinator. The system is designed to be stateless between cycles: all intermediate results are persisted to a SQLite database, allowing the pipeline to resume cleanly after interruption.





4.2 Data Collection Agent

The BTCCollector constructs a `MarketWindow`: a structured status table of market conditions at the moment each 5-minute window opens. It draws from four external sources concurrently:

Source	Data	Latency
Kraken WebSocket	Live BTC/USD tick price	~100ms
Kraken REST	1-minute OHLCV candles	~1s
Polymarket Gamma API	Active 5-min market slug, UP/DOWN odds	~2s
Alternative.me	Fear & Greed Index (0–100)	~1s

The agent maintains a background WebSocket thread tracking a 20-minute rolling BTC price history, from which it derives momentum features at 1-minute, 5-minute, and 15-minute horizons.

The `MarketWindow` output is as shown below:

```
btc_price_now       Current BTC/USD spot price
price_change_1m     % change over past 1 minute  (primary momentum signal)
price_change_5m     % change over past 5 minutes
price_change_15m    % change over past 15 minutes
volatility_5m       Realized volatility (StdDev of last 5 closes)
rsi_14              RSI over last 14 one-minute candles
volume_5m           Aggregate BTC volume over last 5 minutes
up_price            Polymarket implied P(UP)
down_price          Polymarket implied P(DOWN)
```

```
fear_greed_score     Daily sentiment index (0-100)
```

4.3 Debate System

The DebateAgent is the core of our prediction engine. It implements a structured 3-round adversarial debate between two heterogeneous LLMs with opposing analytical frameworks, mediated by the third judge model. The architecture is adapted from the multiagent debate framework of Du et al. (2023) and the structured multi-round protocol of Chan et al. (2023).

Round 1 — Independent Analysis (parallel, ~3–5s)

Two analysts independently analyze the `MarketWindow` without knowledge of each other's position:

Analyst	Model	Role	Primary Signal
Momentum Trader	xAI Grok-3-mini	Exploit price velocity	1m/5m price change, RSI extremes
Contrarian	OpenAI GPT-4o-mini	Spot crowd mispricing	Crowd odds vs momentum alignment

Each analyst returns structured JSON:

```
{"direction": "UP", "confidence": 0.55, "reasoning": "xxx", "key_factors": ["xxx"]}
```

Round 2 — Cross-Examination (parallel, ~3–5s)

Each analyst reads the other's Round 1 position and writes a rebuttal, either defending their original position or conceding if the opponent reveals a missed signal. Rebuttals are required to be short so we can prevent verbose rationalization.

Round 3 — Judgment (single call, ~3–5s)

A judge model (Claude Sonnet) reads all four outputs (both initial positions and both rebuttals) and delivers a final verdict. The judge is explicitly instructed to evaluate argument quality rather than simply adopt the majority position. Rules enforced in the system prompt include:

- Penalize overconfidence: claims above 60% without multiple aligned signals are discounted
- Max final confidence at 0.65 maximum to prevent false precision
- If both analysts agree after rebuttals, slight confidence boost
- If they still disagree, max it at 0.55 (weak signal)

Total latency: 10–15 seconds per cycle — fine within the 5-minute prediction window when it just begins.

4.4 Evaluator

Runs at the start of each cycle, before new predictions are made. For each prediction where the window has closed:

1. Fetches the actual BTC close price from Kraken REST at `window_end`
2. Determines outcome: UP if `close >= open`, else DOWN
3. Computes directional correctness and Brier score
4. Saves `EvalRecord` to SQLite
5. Records the profit/loss in the simluated account

This design guarantees zero look-ahead bias: no prediction is scored until its window has fully closed, and no new prediction is made until all prior closed windows are scored.

4.5 Reporter

Aggregates all `EvalRecord`s and generates a structured accuracy report at the end of each cycle:

- Overall directional accuracy with 95% confidence interval
- Brier score vs random (0.25) baseline
- Confidence calibration by bin (50–55%, 55–60%, 60–65%, 65%+)
- Four-baseline comparison (coin flip, always-UP, always-DOWN, crowd odds)
- Per-model accuracy: xAI (Grok), OpenAI (GPT-4o-mini), Claude (Sonnet as Judge)
- Debate winner analysis (xAI / OpenAI / synthesis)

Reports are saved as timestamped Markdown files and rendered to terminal via `rich`.

4.6 Paper Trading Simulator

A simulated $100,000 account places $500 bet trades on each prediction using live Polymarket odds. Winning tokens pay $1.00; losing tokens lose everything. The simulator is strictly follows the polymarket mechanism.

4.7 Data Persistence

All agent outputs are saved to a local SQLite database:

Table	Contents
`predictions`	All predictions: direction, confidence, reasoning, window timestamps, model votes
`eval_records`	Scored outcomes: correctness, Brier score, actual BTC prices
`trades`	Simulated trades: bet size, odds, payout, P&L, running balance
`account`	Current simulated account balance

5. Data and Evaluation

5.1 Evaluation Metrics

5.1.1 Brier Score

The Brier score measures the accuracy of probabilistic predictions:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2$$

$f_i \in [0,1]$ is the predicted probability for prediction i, and $o_i \in \{0,1\}$ is the actual outcome (1 if correct direction, otherwise 0).

Brier Score	Interpretation
0.00	Always correct at 100% confidence
0.25	Pure random: equivalent to always predicting exact 50%
1.00	Always wrong at 100% confidence

The Brier score has a critial property in analyzing accuracy which is it penalizes overconfidence quadratically. A model that correctly predicts UP at 99% confidence scores $(0.99 - 1.0)^2 = 0.0001$. The same model predicting UP incorrectly at 99% scores $(0.99 - 0.0)^2 = 0.9801$. This incentivizes the model to express genuine uncertainty rather than inflated confidence.

5.1.2 Confidence Calibration

$$P(\text{correct} \mid \text{confidence} = c) \approx c \quad \forall c \in [0,1]$$

We measure calibration by seperating predictions into confidence ranges and computing the calibration gap:

$$\text{gap}_b = \left| \frac{1}{|B_b|}\sum_{i \in B_b} \mathbf{1}[\text{correct}_i] - \frac{1}{|B_b|}\sum_{i \in B_b} f_i \right|$$

where B_b is the set of predictions in confidence range b.

Gap	Calibration Rating
$< 10\%$	Well calibrated
10–20%	Slightly off
$> 20\%$	Poorly calibrated

5.1.3 RSI (Relative Strength Index)

We computed over the last 14 sessions:

$$RSI = 100 - \frac{100}{1 + RS}, \quad RS = \frac{\overline{G}_{14}}{\overline{L}_{14}}$$

where $\overline{G}_{14}$ is the average gain and $\overline{L}_{14}$ the average loss over the 14 most recent price changes.

5.1.4 5-Minute Realized Volatility

$$\sigma_{5m} = \sqrt{\frac{1}{5}\sum_{i=1}^{5}(c_i - \bar{c})^2}$$

where c_i are the 5 most recent close prices and $\bar{c}$ is their mean. High volatility windows indicate high uncertainty and decays the model's confidence.

5.2 Data Collection and Experiment Design

5.2.1 Data Collection Protocol

Data was collected in a single continuous session lasting 16 hours, with 192 consecutive 5-minute Polymarket windows during both day and night to simulate a complete automated trading day.

5.2.2 Baselines

Four baselines are computed over the same 191 scored predictions:

Baseline	Method
Coin flip	50.0% accuracy, Brier 0.2500 (theoretical)
Always UP	Predict UP for every window at 100% confidence
Always DOWN	Predict DOWN for every window at 100% confidence
Crowd odds	Predict the direction favored by Polymarket odds

6. Models and Technologies

6.1 LLM Configuration

Lite models are used for the analyst roles to minimize latency in a 5-minute prediction window.

Model	Provider	Role	Temperature
`grok-3-mini`	xAI	Momentum analyst	0.3
`gpt-4o-mini`	OpenAI	Contrarian analyst	0.3

Model	Provider	Role	Temperature
`claude-sonnet-4-6`	Anthropic	Debate judge	default

Temperature 0.3 is used for the debaters to maintain analytical consistency while preserving enough variability to avoid degenerate agreement. The judge uses the model's default to allow full reasoning flexibility.

6.2 Analytical Personas

xAI Grok (Momentum Trader): Instructed to weight 1-minute and 5-minute price change as the primary signal, use RSI extremes as a secondary signal, and treat crowd odds as background noise.

OpenAI GPT-4o-mini (Contrarian): Instructed to weight crowd odds skew vs momentum alignment as the primary signal, look for RSI mean-reversion setups, and treat momentum as confirmation rather than the primary driver.

Claude Sonnet 4.6 (Judge): Instructed to evaluate argument quality rather than vote count, penalize overconfidence, and synthesize a calibrated final verdict.

7. Responsible AI Considerations

7.1 Financial Risk Disclosure

PolySignal-BTC5 is a **research instrument**, not a trading system. The paper trading results reflect simulated performance under controlled conditions and do not guarantee future returns. Cryptocurrency markets are volatile and prediction accuracy at any timescale is not guaranteed to persist (Sebastião and Godinho 2021). Nothing in this paper constitutes financial advice.

7.2 Hallucination Mitigation

LLM hallucination in financial prediction can be a critical fatal error — the model asserting a direction with high certainty when the evidence is weak. PolySignal-BTC5 mitigates this through:

1. **Hard confidence cap at 0.65** — prevents any single model from overconfidence
2. **Structured JSON output with explicit confidence rules** — the system prompt quantifies what evidence is required for each confidence level
3. **Adversarial rebuttal** — a second model explicitly challenges the first's reasoning, surfacing errors before the judge rules
4. **Brier score evaluation** — overconfident wrong calls are penalized quadratically, creating a feedback signal for prompt refinement

7.3 Bias and Fairness

The system potentially shows a directional bias: 66% UP calls (126/191) against a 49.7% actual UP rate. This UP bias is consistent with our momentum analyst's preference to interpret flat or slightly positive signals as UP. The bias did not harm the overall accuracy, but it suggests a potential issue for future works.

8. Findings and Discussion

8.1 Overall Performance

Over 191 scored predictions:

Metric	Value
Directional accuracy	**57.6% ± 7.0%** (95% CI)
95% CI lower bound	**50.6%** (above coin flip)
Avg Brier score	**0.2474**
Correct / Total	110 / 191
UP calls	126 (66%)
DOWN calls	65 (34%)
Actual UP rate	49.7%
Session duration	16 hours (2026-04-24)

8.2 Baseline Comparison

Strategy	Accuracy	Brier Score	vs PolySignal-BTC5
PolySignal-BTC5 (ours)	**57.6%**	**0.2474**	—
Crowd (Polymarket odds)	55.0%	0.2475	−2.6% accuracy
Coin flip	50.0%	0.2500	−7.6% accuracy
Always DOWN	50.3%	0.4974	−7.3% accuracy
Always UP	49.7%	0.5026	−7.9% accuracy

PolySignal-BTC5 outperforms all four baselines on directional accuracy. On Brier score, PolySignal-BTC5 is better than the crowds' preference with identical probabilistic calibration but higher directional accuracy.

The outperformance over crowd odds is worthy noticed: Polymarket prices are set by financially incentivized traders with full access to the same technical data. Beating this baseline suggests the debate architecture surfaces signal that the aggregate crowd underweights.

8.3 Confidence Calibration

Bin	Number	Correct	Actual %	Avg Conf	Gap	Rating
50–55%	**180**	103	57.2%	52.8%	**4.4%**	**Well calibrated**
55–60%	10	7	70.0%	55.0%	15.0%	Slightly off
60–65%	1	0	0.0%	61.3%	61.3%	Insufficient data

The model correctly expresses genuine uncertainty: it never claims high confidence, and its low-confidence calls succeed at a rate (57.2%) that meaningfully exceeds the stated confidence level (52.8%).

8.4 Per-Model Accuracy

Model	Role	Votes	Accuracy	Brier
xAI (Grok-3-mini)	Momentum analyst	188	**59.6% ± 7.0%**	**0.2436**
Claude (Sonnet 4.6)	Final verdict	188	**57.4% ± 7.1%**	**0.2469**
OpenAI (GPT-4o-mini)	Contrarian analyst	188	49.5% ± 7.1%	0.2532

xAI Grok is the biggest and best analyst with 59.6% accuracy and achieves the highest Brier score (0.2436). Claude's final verdict (57.4%) tracks closely with Grok's position, suggesting the judge frequently agrees to the momentum analyst's argument.

OpenAI's contrarian achieves 49.5% accuracy. The contrarian role, as for this case, does not provide good enough directional signal at 5-minute Bitcoin timescales.

8.5 Debate Winner Analysis

Judge's ruling	Count	Correct	Accuracy
xAI argument wins	63	38	**60.3%**
Synthesis (both)	121	69	57.0%
OpenAI argument wins	4	1	**25.0%**

When the Claude judge determines that xAI's momentum argument alone is more persuasive and without synthesis, the prediction achieves 60.3% accuracy. When OpenAI's contrarian argument wins alone, accuracy drops to 25.0% (the sample size is small). Synthesis predictions achieve 57.0%, intermediate between the two extremes.

This decomposition reveals that the debate's value lies in its ability to identify when the momentum signal is unambiguous enough to override the contrarian view. When Claude synthesizes both positions, it incorporates noise from the contrarian, slightly diluting the signal.

8.6 Simulated Trading Performance

Metric	Value
Starting balance	$100,000.00
Final balance	**$113,321.74**
Total P&L	**+$13,321.74 (+13.32%)**
Trades	191
Win / Loss	110W / 81L
Win rate	57.6%
Avg win P&L	+$525.16 per trade
Avg loss P&L	−$500.00 per trade

The positive ROI is mathematically consistent with the directional accuracy. At near-even Polymarket odds (~50.5 cents per token) and 57.6% win rate, expected value per trade is positive.

8.7 Limitations

8.7.1 Sample Size and Temporal Coverage

191 predictions across 16 hours is sufficient for directional accuracy claims (95% CI width ±7.0%) but insufficient for long-run calibration stability.

The evaluation window covers only a single market session. Performance during high-volatility events may differ substantially from the results reported here.

8.7.2 Single Session Evaluation

All 191 predictions were made in a single 16-hour window on the same day. Market conditions during this session were relatively calm, with an actual UP rate of 49.7% (close to the expected 50/50). A more rigorous evaluation would span multiple sessions across different market regimes.

8.7.3 Directional Bias

The system called UP in 66% of predictions (126/191) against an actual UP rate of 49.7%. During this session the actual UP rate happened to be close to 50/50, which means the bias did not hurt accuracy too much. In a session where the market trends strongly downward, however, a system that defaults toward UP would underperform significantly. The bias originates in the momentum analyst's prompt design: flat or weakly positive short-term price changes are interpreted as UP signals.

8.7.7 Price Reference Timing

`btc_price_at_call` and the Chainlink oracle opening price used for market resolution may differ by up to a minute due to data collection latency. This introduces a small timing mismatch between the reference price used for prediction and the price used for outcome determination.

9. Conclusion and Future Work

9.1 Conclusion

PolySignal-BTC5 shows that a multi-agent LLM debate system can beat the triditional methods. During 191 predictions in a 16-hour session, the system gave a 57.6% directional accuracy which was above the 55.0% crowd preference with a Brier score of 0.2474 and a 13.32% simulated profit.

The key finding is that the momentum analyst drove most of the system's edge. When the judge sided with Grok alone, accuracy hit 60.3%. When the contrarian won, it dropped to 25.0%. At 5-minute timescales, price continuation is a stronger signal than mean reversion, and a judge model can reliably tell the difference.

9.2 Future Work

9.2.1 Model Upgrades

The most direct improvement could be upgrading the debater models. From what we have now to their upgraded version to get a better reasoning and evaluation. However, bigger models come with higher latencies which should also be taken into concern.

9.2.2 Continuous 24/7 Deployment

The current implementation runs in a limited time sessions. Production deployment on AWS EC2 with a systemd service running `run.py --forever` would provide continuous data collection and a live public dashboard reflecting real-time predictions and outcomes.

9.2.3 Debate Architecture Variants

The current 2 analyst debate with 1 judge decide system could be extended to 3 debaters with different timescale specializations, or to a sequential debate where additional rounds are triggered when debaters fail to reach consensus after certain rounds.

9.2.4 Ensemble-Debate Hybrid

Since xAI Grok is the best analyst we had during the trail, we can make a hybrid architecture: run three Grok instances in parallel as an ensemble for directional consensus, then trigger the full debate only when the ensemble is split.

References

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, et al. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–1901. arXiv:2005.14165.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1–3.

Chan, Chi-Min, Weize Chen, Yiyuan Li, Zhiyuan Liu, and Maosong Sun. 2023. "ChatEval: Towards Better LLM-Based Evaluators through Multi-Agent Debate." arXiv:2308.07201.

Chen, Kay-Yut, and Charles R. Plott. 2002. "Information Aggregation Mechanisms: Concept, Design, and Implementation for a Sales Forecasting Problem." Working Paper 1131. Social Science Working Paper Series. Pasadena, CA: California Institute of Technology.

Du, Yilun, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. "Improving Factuality and Reasoning in Language Models through Multiagent Debate." arXiv:2305.14325.

Fama, Eugene F. 1970. "Efficient Capital Markets: A Review of Theory and Evidence." *Journal of Finance* 25 (2): 383–417.

Galton, Francis. 1907. "Vox Populi." *Nature* 75 (1949): 450–451.

Hayek, Friedrich A. 1945. "The Use of Knowledge in Society." *American Economic Review* 35 (4): 519–530.

Lopez-Lira, Alejandro, and Yuehua Tang. 2023. "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models." arXiv:2304.07619.

Mellers, Barbara, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz, Lyle Ungar, and Philip Tetlock. 2015. "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions." *Perspectives on Psychological Science* 10 (3): 267–281.

Polymarket. 2024. "Polymarket Platform Documentation and Market Resolution Rules." Accessed April 2026. https://polymarket.com.

Sebastião, Helder, and Pedro Godinho. 2021. "Forecasting and Trading Cryptocurrencies with Machine Learning under Changing Market Conditions." *Financial Innovation* 7 (1): 1–30.

Spann, Martin, and Bernd Skiera. 2009. "Sports Forecasting: A Comparison of the Forecast Accuracy of Prediction Markets, Betting Odds and Tipsters." *Journal of Forecasting* 28 (1): 55–72.

Tetlock, Philip E. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.

Wilder, J. Welles. 1978. *New Concepts in Technical Trading Systems*. Greensboro, NC: Trend Research.

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

Xie, Qianqian, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. "PIXIU: A Large Language Model, Instruction Data, and Evaluation Benchmark for Finance." arXiv:2306.05443.