Polysingal-BTC5: Multi-Agent Debate System in Short-Term Bitcoin Prediction Market

Lance Yi

"A large group of diverse individuals will come up with better and more robust forecasts and make more intelligent decisions than even the [experts]."

— James Surowiecki, *The Wisdom of Crowds* (p.41)

If crowds outperform experts,
can structured AI outperform the crowd?

Can *MAD* beat *Crowd* ?

MAD: Multi-agent debate

Three AI agents — a momentum analyst, a contrarian, and a judge — debate Bitcoin's next 5-minute price move on prediction market.

Directional accuracy

57.6 %

(crowd: 55.0%)

Brier Score

0.24 74

(crowd:0.2475)

ROI (in 16 hours, 191 trades)

13.3 %

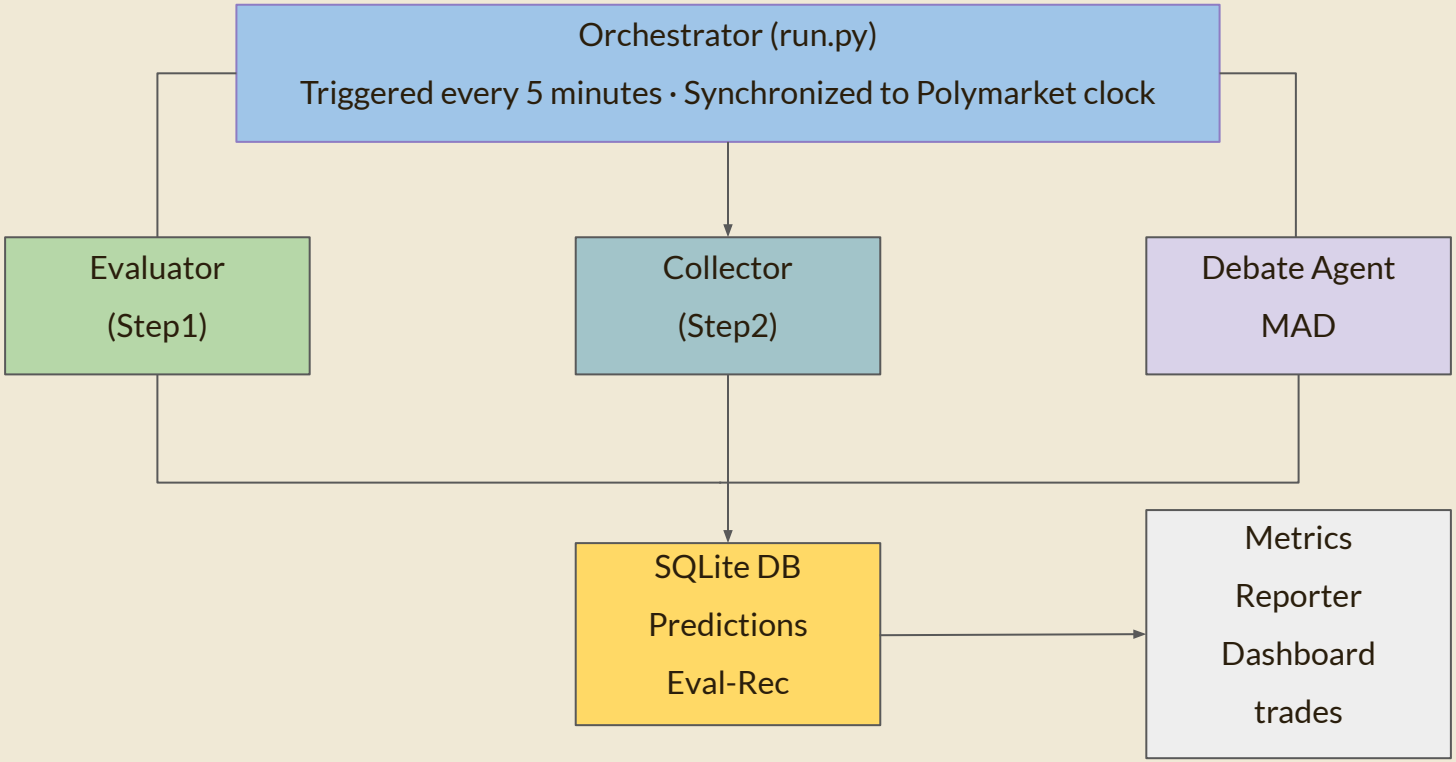
Motivation

Why 5-minute Bitcoin prediction markets?

- 288 markets per day → dense, fast feedback loops
- Binary outcomes resolved by Chainlink oracle → no ambiguity about ground truth
- Crowd odds (Polymarket AMM) encode aggregate public information → strong baseline
- If prices are fully efficient (EMH), no AI should beat the crowd

Can a multi-agent LLM debate system extract directional signal beyond what the crowd already knows?

System Architecture



Data Collection

10 features per market window from 4 live sources:

Source	Signal	Latency
Kraken WebSocket	Live BTC price	~100ms
Kraken REST OHLCV	RSI(14), volatility, 1/5/15m momentum	~1s
Polymarket Gamma API	Crowd UP/DOWN odds	~2s
Alternative.me	Fear & Greed Index (0–100)	~1s

MAD: Multi-Agent Debate

3 rounds: ~8–14 seconds total

Round 1 — Independent (parallel)

- xAI Grok → Momentum

- OpenAI GPT-4o-mini → Contrarian

Round 2 — Cross-examination (parallel)

- Grok reads GPT's argument → Rebuttal

- GPT reads Grok's argument → Rebuttal

Round 3 — Judgment (single call)

- Claude Sonnet reads all 4 outputs

- Delivers: direction + confidence + reasoning summary

```
Round 1 (Independent - Diversity)
 Grok  →  Position A
 GPT   →  Position B

Round 2 (Debate - Conflict)
 Grok  →  Rebuttal A (reads B)
 GPT   →  Rebuttal B (reads A)

Round 3 (Judge - Aggregation)
 Claude → Final verdict
          direction + confidence
```

Evaluation Metrics

Directional Accuracy and Brier Score

$$DA = \frac{k}{N}$$

Directional Accuracy

- Did the model predict the correct movement?
- Treats 51% and 99% predictions equally
- Answers: *Would this signal make money?*

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

Brier Score

- How confident and calibrated was the prediction?
- Heavily penalizes overconfident errors
- Answers: *Can you trust the probability?*

In prediction markets, direction tells you *if* you'd profit — calibration tells you *how much* to bet. Both are required for a trustworthy forecasting system.

Results

191 Trades in *16* Hours

Model	Votes	Correct	Accuracy	Avg Brier
Claude (Judge)	190	110	57.9%	0.2466
Haiku (Momentum)	3	1	33.3%	0.2799
OpenAI (GPT-4o-mini)	190	94	49.5%	0.2532
Sonnet (Technical)	3	2	66.7%	0.2722
Sonnet 4.6 (Contrarian)	2	1	50.0%	0.3074
xAI (Grok)	190	114	60.0%	0.243

Results

191 Trades in *16* Hours

Metric	PolySignal	Baseline
Directional accuracy	57.6%	55.0% (crowd)
95% CI lower bound	50.6%	50.0% (coin flip)
Brier score	0.2474	0.2475 (crowd)
Simulated ROI	+13.32%	—

Per-Model Accuracy

Model	Role	Accuracy	Note
xAI Grok	Momentum analyst	*59.6%*	Strongest single model
OpenAI GPT-4o-mini	Contrarian	49.5%	~random at 5-min
Claude Sonnet	Judge	*57.6% (same as the system accuracy)*	Final verdict only
Judge sides with Grok	—	60.3%	Highest observed
Judge sides with GPT	—	25.0%	Lowest observed

Momentum dominates short-term prediction signal

Limitation

- **Statistical power:** 191 trades from a single 16-hour session — wide confidence intervals, sample size a bit small
- **No live capital:** Simulated P&L only; actual market is restricted to US users
- **Weak contrarian:** GPT-4o-mini at 49.5% may be too small for the contrarian role
- **Single calm day:** No evaluation across different market regimes (high/low volatility)
- **Model Performance:** Lite models used only, reasoning ability not most updated

Future Work

- **Longer evaluation:** 7+ days, 2,000+ trades for tighter confidence intervals, expand it to months even years later.
- **Stronger models:** GPT-5.4 or Claude Opus 4.7 as contrarian instead of GPT-4o-mini and Sonnet 4.6
- **Adaptive sizing:** Dynamic position sizing instead of flat $500/trade
- **Multi-horizon:** Test different market windows like 15-min, 30-min
- **Live deployment:** Real capital on Polymarket to validate simulated results

Demo

Conclusion

- **Multi-agent debate achieves a modest improvement over the crowd baseline**
  (57.6% vs 55.0% directional accuracy)
- **Structured disagreement + aggregation provides incremental predictive signal**
  beyond market consensus
- **Results are preliminary**
  limited by *sample size, short horizon, and simulated trading environment*