Alive and Predicting: A Live Evaluation of Multi-Step Forecasting Agents

Will Wu [1] **Hui Dai** [1 2] **Mengye Ren** [1]

Abstract

Large language models are increasingly capable forecasters, yet most of this capability has been measured by retrospective backtest on already-resolved questions, and recent live benchmarks score only the agent's final probability. We present a live, multi-step forecasting agent that operates autonomously on a major prediction market, with every intermediate forecast, retrieved evidence item, and tool invocation recorded and published in real time. Scoring forecasts by their information coefficient against the subsequent market movement, the agent beats a zero-shot baseline at every horizon from one day to one month after the forecast, with the largest margin in the first two weeks before the market converges toward the agent's earlier view. Because every reasoning step is recorded, we can identify which pipeline stages and tools contribute most to forecasting accuracy and surface agent pipeline design lessons.

1. Introduction

Large language models (LLMs) have rapidly become capable forecasters, both as retrieval-augmented agents (Halawi et al., 2024; Hsieh et al., 2024; Murphy, 2026; Alur et al., 2025) and as fine-tuned predictors (Turtel et al., 2025a;b; 2026; Chandak et al., 2025). However, how these systems are evaluated has lagged behind how they are built. Most are graded via a historical backtest on resolved questions, filtering to questions resolved after the model's training cutoff. This setup raises two concerns: contamination filtering is imperfect when models retrieve from the internet (Paleka et al., 2025), and evaluators commonly substitute degraded tooling for production, such as static news corpora in place of live web search (Halawi et al., 2024; Dai et al., 2025; Chandak et al., 2025). Recent benchmarks such as ForecastBench (Karger et al., 2025) and Prophet Arena (Yang et al., 2026) have introduced live data, but their primary unit of evaluation is still the final forecast accuracy. There is less insight into where the forecasting skill is distributed across the particular stages and tool calls used on each live scan.

This distinction matters because a multi-step agent can improve its final score for several different reasons: it may retrieve genuinely new information, average away idiosyncratic model errors, calibrate against market prices, or simply become more conservative. A final probability alone cannot distinguish these mechanisms. To design better forecasting agents, we need evaluations that preserve the intermediate decisions that produced the forecast, rather than collapsing the entire pipeline into one number.

We address the question: *how do stages and tools of a multi-step LLM forecasting agent contribute to the final forecast accuracy?* For each forecasting question, the agent records six forecasts: a zero-shot baseline, three independent forecasts, their average, and a final aggregated forecast. All instances of tool calls and retrieved evidence are published live on a public website.[1] We evaluate forecasts using the information coefficient, the correlation between the agent's forecast and the market's movement in the following days. This design allows us to decompose the agent's performance by pipeline stages and tool usage.

We deploy our agent on a major prediction market for 4+ weeks, and the agent beats both a zero-shot baseline and a mean-reversion baseline at every horizon. Decomposing performance stage by stage and tool by tool, we observe that: (i) most of the multi-step gain comes from the final post-critic stage, but the trace shows the critic mainly moving the forecast toward the market price, so that gain looks more like calibration than added reasoning; (ii) of the ten retrieval tools, only a Wikipedia base-rate lookup and the Kalshi orderbook raise the information coefficient when used, while several of the data tools' contributions are negative.

[1]New York University, New York, NY, USA [2]The University of Chicago, Chicago, IL, USA. Correspondence to: Mengye Ren <mengye@nyu.edu>, Will Wu <willwu@stern.nyu.edu>.

Accepted at the ICML 2026 Workshop on Forecasting as a New Frontier of Intelligence.

[1]https://forecast.agenticlearning.ai

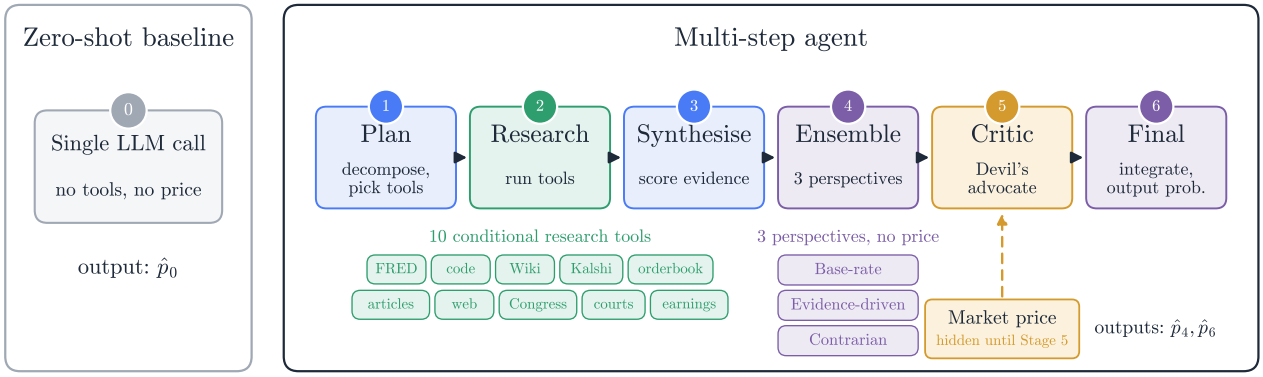

Figure 1. **The live forecasting agent.** The *zero-shot baseline* (left) is a single LLM call with no tools and no market price. The *multi-step agent* (right) decomposes the question, retrieves evidence through ten tools, runs three ensemble perspectives without showing them the market price, and then integrates the result through an adversarial critic that sees the price for the first time at Stage 5. Every numbered stage's probability is recorded per scan, yielding the six forecasts $\{\hat{p}_0, \hat{p}_{4,a}, \hat{p}_{4,b}, \hat{p}_{4,c}, \hat{p}_4, \hat{p}_6\}$ used throughout the paper. Full prompts and tool inventory in Appendix B.

2. Agent Pipeline Methodology

Our multi-step forecasting agent comprises seven sequential stages (see Fig. 1).

1. **Stage 0:** A zero-shot baseline produces a single-call forecast on the question alone.
2. **Stage 1:** A planning model classifies the question, decomposes it into weighted sub-questions, and selects which of ten research tools to invoke.
3. **Stage 2:** A research executor calls the selected tools in parallel.
4. **Stage 3:** A synthesis model labels each retrieved evidence item with categorical strength, continuous credibility, direction, and whether it is already priced in.
5. **Stage 4:** Three parallel forecasters read the same evidence under three different prompts, labelled base-rate, evidence-driven, and contrarian, without seeing the market price.
6. **Stage 5:** A Devil's Advocate critic is then shown the market price and asked to challenge the ensemble forecast.
7. **Stage 6:** A final integrator merges the critique with the market price and emits the final probability.

In our experiments, Claude Sonnet 4.6 (Anthropic, 2026b) handles planning and evidence synthesis (Stages 1 and 3), and Claude Opus 4.7 (Anthropic, 2026a) handles the zero-shot baseline, the three forecasters, the Devil's Advocate critique, and the final integration (Stages 0, 4, 5, and 6).

For each scan, the agent runs the zero-shot baseline and the multi-step pipeline on the same question with the same evidence cutoff, and stores the full evidence trace, every intermediate probability, and price follow-ups. This paper analyzes 335 scans covering 269 unique Kalshi markets, drawn from an 4-week rolling window after the pipeline was finalized.

Snapshot	Value
Scans (follow-ups at every day to $t{+}30$)	335
Mean tools per scan	5.4

Table 1. Snapshot dataset summary. All 335 scans have a matured price follow-up at every day from $t{+}1$ to $t{+}30$.

Six forecasts per scan. Each scan records the zero-shot probability $\hat{p}_0$ from Stage 0, three perspective probabilities $\hat{p}_{4,a}, \hat{p}_{4,b}, \hat{p}_{4,c}$ from Stage 4, their ensemble forecast $\hat{p}_4$, and the final post-critic probability $\hat{p}_6$ from Stage 6. Computing the information coefficient at each gives a stage-level ablation from one set of inference passes.

Snapshot dataset. The snapshot dataset (from our live evaluation) contains 335 scans covering 269 unique markets across eight categories: politics, economics, science, geopolitics, tech, legal, general, and culture. The scans began on 2026-04-07, the day the pipeline was finalized. The full schema and per-scan fields are in Appendix A.

Metric. Most markets in the snapshot are unresolved at analysis time, so we evaluate against the market movement in the following days. Let $p_m^{(t)}$ denote the market mid-price at the moment of the scan and $p_m^{(t+K)}$ its mid-price K days later. We evaluate forecasting performance using the **information coefficient** (IC), the Pearson correlation between the agent's edge $\hat{p}_i - p_m^{(t)}$ and the subsequent price move $p_m^{(t+K)} - p_m^{(t)}$ at horizon K. IC evaluates whether the market subsequently moves toward the agent's forecast, capturing both direction and magnitude. A higher positive IC indicates that markets tend to move more closely toward the agent's forecast.

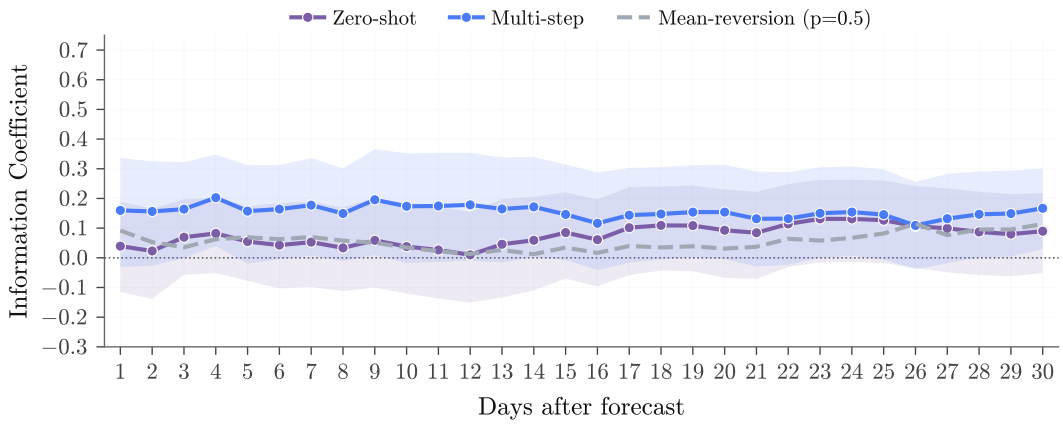

Figure 2. **Agent forecasting performance across forecast horizons.** Information coefficient by day after forecast. The multi-step agent (blue) beats the zero-shot baseline (purple) and a mean-reversion baseline (gray) at every horizon from $t{+}1$ to $t{+}30$. Mean IC over $t{+}1$ to $t{+}14$: multi-step $+0.17$, zero-shot $+0.05$, mean-reversion $+0.05$. Shaded bands are 95% bootstrap CIs. Every horizon has 335 scans; the multi-step margin is largest over the first two weeks and narrows toward $t{+}30$ as the market converges to the agent's view.

Stage	Mean IC	95% CI
Zero-shot	$+0.05$	$[-0.09,\ +0.17]$
Ensemble (3-persp. avg)	$+0.14$	$[-0.01,\ +0.27]$
Final (post-critic)	$\mathbf{+0.17}$	$[\mathbf{+0.00},\ \mathbf{+0.33}]$

Table 2. Per-stage mean IC ($t{+}1$ to $t{+}14$) with 95% bootstrap CIs (2000 resamples). Only the post-critic forecast has a non-negative lower bound. Individual perspectives (base-rate, evidence-driven, contrarian) are within ± 0.03 of the ensemble average.

3. Results

Main result. Fig. 2 and Table 2 report the core result on forecasting agent pipelines. The final forecasts show positive market correlation at all horizons, and the full multi-step agent beats the zero-shot baseline and a simple mean-reversion baseline at every horizon, so the gain is not explained by a mechanical extreme-price strategy. Mean IC rises across the pipeline, from $+0.05$ for the zero-shot baseline to $+0.14$ for the three-perspective ensemble average and $+0.17$ for the post-critic final (Table 2). Among the stages, only the post-critic final has a non-negative lower bound in its 95% bootstrap interval. The agent's edge over zero-shot is established within the first two weeks, and the market then moves toward the position it had already taken: the edge appears early and the price later confirms it. This suggests that the agent's forecast contains information the market has not yet priced. An IC of $+0.17$ is a significant difference on a live evaluation whose markets are mostly still open, where the price already aggregates the views of many active participants. The win is greatest when the agent commits capital, on its high-conviction calls (Fig. 4).

Stage-level decomposition. Table 2 and Fig. 3 show how forecast skill changes across the pipeline. The three Stage 4 perspectives provide only limited diversity: on 38% of scans

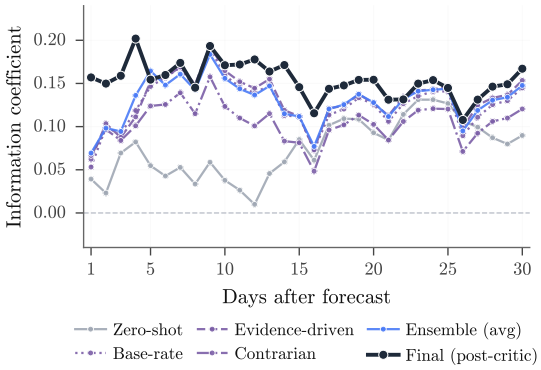

Figure 3. **Per-stage IC across horizons.** Same scans, different snapshots within each scan. The final forecast (black) sits above every other line at every horizon. Ensemble average and its three perspectives cluster near $+0.14$; zero-shot (gray) is lowest.

their probabilities differ by less than 0.02, so the individual perspectives and their simple average have similar ICs (Appendix D). The main additional gain appears after the market price is revealed. The post-critic final forecast achieves a mean IC of $+0.17$, compared with $+0.14$ for the price-blind ensemble, and remains the top-performing stage at every horizon from $t{+}1$ to $t{+}30$. This suggests that the final stage improves performance by reconciling the model's independent estimates with the market price, rather than by generating a substantially new forecast from scratch.

High-conviction calls are crucial. The IC averages high and low conviction forecasts together. Splitting the scans by conviction, IC increases with the size of the agent's edge over the market price: only calls with an edge of 15% or more hold a clearly positive IC at every horizon, while lower-conviction calls have IC near zero or below. This concentration is especially useful in practice, since

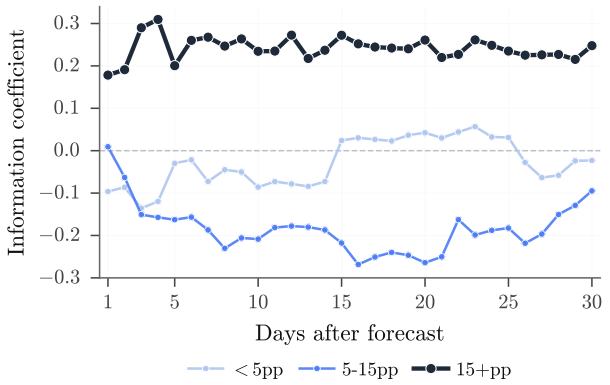

Figure 4. **IC by conviction across horizons.** Scans split by conviction, the absolute edge of the agent's forecast over the market $|\hat{p} - p_m^{(t)}|$, into <5pp (n=141), 5–15pp (n=115), and 15+pp (n=79).

Tool	% used	Mean ΔIC	SD
Always invoked (no contrast)			
article search	99.7	—	—
kalshi data	99.7	—	—
web search	96.1	—	—
Conditional (ranked by ΔIC)			
wikipedia lookup	88.1	+0.25	0.17
kalshi orderbook	57.9	+0.13	0.13
earnings data	7.5	+0.05	0.39
court docket	5.4	−0.12	0.22
code execution	70.1	−0.24	0.17
fred data	11.0	−0.26	0.21
congress bills	6.3	−0.27	0.22

Table 3. Per-tool change in IC, with bootstrap standard deviation (2000 resamples). The change is the IC on scans where the planner used the tool minus the IC on scans where it did not.

the trading agent typically takes a position only on high-conviction calls.

Tool-level decomposition. We compared the tools with conditional usages (Table 3). The Wikipedia lookup stands out: used on most scans, it is the most consistent help to the forecast. The Kalshi orderbook tool and earnings data tool provide a subtle increase in IC. The remaining four conditional tools, namely code execution, FRED, Congress bills, and court docket, are tied to worse forecasts.

We do not read a tool's lower IC as evidence that its data source is unhelpful: it could mean that the tool returned little of value, or that the specific sub-set of questions is inherently harder to forecast. Therefore, the takeaway is that the selection or usage of these tools could be misaligned in the current pipeline, not that the data itself is uninformative in principle.

4. Conclusion

Forecasting is becoming a central test of machine intelligence, yet the agents built for it are still judged mostly by their final accuracy, with little insight into which steps/tools are most useful. In this work, we evaluated a live, multi-step forecasting agent on a real prediction market, with every intermediate forecast recorded. Scored by information coefficient against subsequent market movement, the agent shows forecasting skill, beating a zero-shot baseline at every horizon from one day to one month. Because the whole trace is stored in a database, we can also locate where that gain comes from. We release the agent, its trace database, and the full analysis, so the same live record can test new designs as forecasting markets continue to change.

Limitations. There are several limitations of our current release: 1) most markets are still unresolved; 2) the data is based on a single platform; and 3) the evaluation is based on a single family of base models. We plan to continue developing our online platform with a more flexible design of agentic architecture and a wider evaluation set.

Impact Statement

This work studies an LLM forecasting agent on live prediction markets. The broader trajectory of LLM-driven forecasting carries both promise and risk. On the positive side, capable forecasting agents could broaden access to probabilistic reasoning historically concentrated in well-resourced quantitative trading firms, support better-calibrated public discourse on uncertain events, and help institutions price risk more accurately. On the negative side, the same systems could concentrate market-making advantage among those able to deploy them at scale, encourage opaque automated trading that is harder for regulators and counterparties to audit, and amplify the impact of correlated model errors when many participants run similar pipelines.

To mitigate this risk, we embrace open science in this work. The agent code, trace database, analysis scripts, and the live dashboard at `https://forecast.agenticlearning.ai` are public and experiments are reproducible from public data. Temporarily, the full reasoning trace dataset is locked behind an API key to prevent commercial misuse of our agents, but researchers can request full data access by contacting us. Releasing the full reasoning trace lets external researchers identify failure modes and disagree with our design choices on equal footing. All trading results reported in the paper are paper trades with no real capital deployed; the live agent does not place real-money orders.

Acknowledgment

This work was supported in part by Visko AI, a Google TPU Award, the NYU-KAIST Award A25-0081-002, NSF BCS Award 2545541, and the Institute of Information & Communications Technology Planning Evaluation (IITP) under grant RS-2024-00469482, funded by the Ministry of Science and ICT (MSIT) of the Republic of Korea in connection with the Global AI Frontier Lab International Collaborative Research. The compute is supported by the NYU High Performance Computing resources, services, and staff expertise.

References

Alur, R., Stadie, B. C., Kang, D., Chen, R., McManus, M., Rickert, M., Lee, T., Federici, M., Zhu, R., Fogerty, D., Williamson, H., Lozinski, N., Linsky, A., and Sekhon, J. S. AIA forecaster: Technical report. *arXiv preprint arXiv:2511.07678*, 2025. Bridgewater AIA Labs.

Anthropic. Introducing claude opus 4.7. https://www.anthropic.com/news/claude-opus-4-7, April 2026a. Accessed: 2026-05-13.

Anthropic. Introducing claude sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6, February 2026b. Accessed: 2026-05-13.

Chandak, N., Goel, S., Prabhu, A., Hardt, M., and Geiping, J. Scaling open-ended reasoning to predict the future. *arXiv preprint arXiv:2512.25070*, 2025.

Dai, H., Teehan, R., and Ren, M. Are LLMs prescient? a continuous evaluation using daily news as the oracle. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models. *Advances in Neural Information Processing Systems*, 2024.

Hsieh, E., Fu, P., and Chen, J. Reasoning and tools for human-level forecasting. *arXiv preprint arXiv:2408.12036*, 2024.

Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. E. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations (ICLR)*, 2025.

Murphy, K. Agentic forecasting using sequential bayesian updating of linguistic beliefs. *arXiv preprint arXiv:2604.18576*, 2026.

Paleka, D., Goel, S., Geiping, J., and Tramèr, F. Pitfalls in evaluating language model forecasters. *arXiv preprint arXiv:2506.00723*, 2025.

Turtel, B., Franklin, D., and Schoenegger, P. Llms can teach themselves to better predict the future. *arXiv preprint arXiv:2502.05253*, 2025a.

Turtel, B., Franklin, D., Skotheim, K., Hewitt, L., and Schoenegger, P. Outcome-based reinforcement learning to predict the future. *arXiv preprint arXiv:2505.17989*, 2025b.

Turtel, B., Wilczewski, P., Franklin, D., and Skothiem, K. Future-as-label: Scalable supervision from real-world outcomes. *arXiv preprint arXiv:2601.06336*, 2026.

Yang, Q., Mahns, S., Li, S., Gu, A., Wu, J., and Xu, H. LLM-as-a-prophet: Understanding predictive intelligence with prophet arena. In *International Conference on Learning Representations (ICLR)*, 2026.

A. Reproducibility

Release. The agent code, the trace database, and the analysis scripts that produce every figure and number reported here are public. The live agent, the full reasoning traces, and a one-click data export are available at https://forecast.agenticlearning.ai. The dataset, methodology, and prompts are documented below, and the agent specification (Section 2, Appendix B) together with the dataset summary (Table 4) is sufficient to reproduce the results.

Snapshot definition. Multi-step scans with `scanned_at` between 2026-04-07 (the date the agent pipeline was finalised) and 2026-05-11, operator-archived flag unset. The official release includes both this static snapshot and up-to-date versions.

Metric	Value
Multi-step scans	335
Zero-shot baselines	335
Ensemble forecasts (3× scans)	1005
Scans with $t+1$ price follow-up	335
Scans with $t+3$ price follow-up	335
Scans with $t+7$ price follow-up	335
Scans with $t+14$ price follow-up	335
Unique markets covered	269
Categories (politics, economics, etc.)	8
Mean tools used per scan	5.4

Table 4. Snapshot summary. All main-paper numbers are computed on this sample.

.

B. Full Pipeline

Stage descriptions. Below we describe each stage in our agent design. Frontier-tier and fast-tier refer to the two model classes used; specific model versions are documented in the released code.

- **Stage 0, Zero-shot baseline** (frontier). Single call on title, description, resolution date, current date. No tools, no market price.
- **Stage 1, Planning** (fast). Classifies the question, decomposes it into 2–5 weighted binary sub-questions, selects which of 10 tools to invoke.
- **Stage 2, Research** (no LLM). Executes selected tools in parallel.
- **Stage 3, Synthesis** (fast). Labels each evidence item with categorical strength, continuous credibility (1–100), direction, and whether already priced in.
- **Stage 4, Ensemble forecaster** (3× frontier). Three perspectives (base-rate, evidence-driven, contrarian) on the same evidence. None see the market price. Point estimate is the simple average $\hat{p}_4$.
- **Stage 5, Adversarial review** (frontier). Market price revealed. Critic challenges the ensemble for reasoning flaws and math errors.
- **Stage 6, Final forecast** (frontier). Integrates the critique and the market price; produces final probability $\hat{p}_6$ and confidence.

Ensemble perspective prompts. **Base-rate.** "You anchor heavily on historical frequencies and reference classes before considering specific evidence. Start from 'how often does this type of event happen?' and only deviate with strong evidence."

Evidence-driven. "You weight recent, specific evidence most heavily. Focus on what has CHANGED recently: new information, breaking developments, and shifts that make this situation different from historical base rates."

Contrarian. "You actively look for reasons the consensus might be wrong. What are people overlooking? What scenario would surprise most observers? Weight evidence against the popular narrative more heavily."

C. Tool Inventory

Tool	Source	Use case
article_search	Custom endpoint	News articles
web_search	LLM provider API	Breaking news
kalshi_data	Kalshi API	Price history, trends
kalshi_orderbook	Kalshi API	Bid/ask depth
fred_data	FRED API	GDP, unemployment, CPI
congress_bills	Congress.gov API	Bill tracking
court_docket	CourtListener API	Court cases
code_execution	Python sandbox	Statistical analysis
wikipedia_lookup	Wikipedia API	Base rates, facts
earnings_data	Yahoo Finance	Company financials

Table 5. Research tools available to the planning agent. Selection is conditional per question.

D. Additional Charts

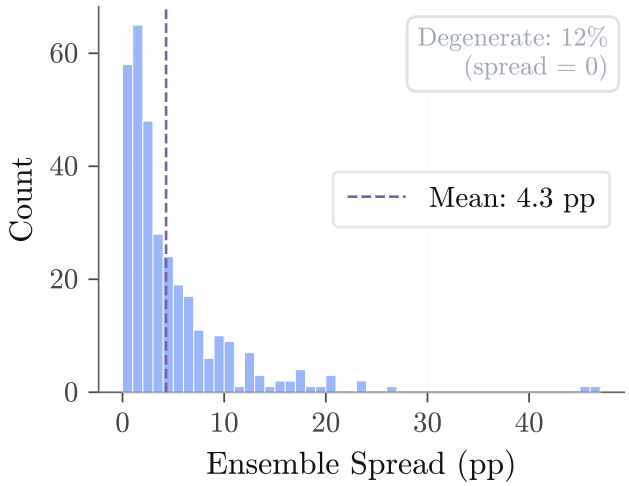

Figure 5. **Diversity collapse.** Histogram of ensemble spread. On 38% of scans the spread is below 0.02: the three prompts converge on nearly the same forecast. This is the mechanism behind the ensemble-average step's flat IC contribution.

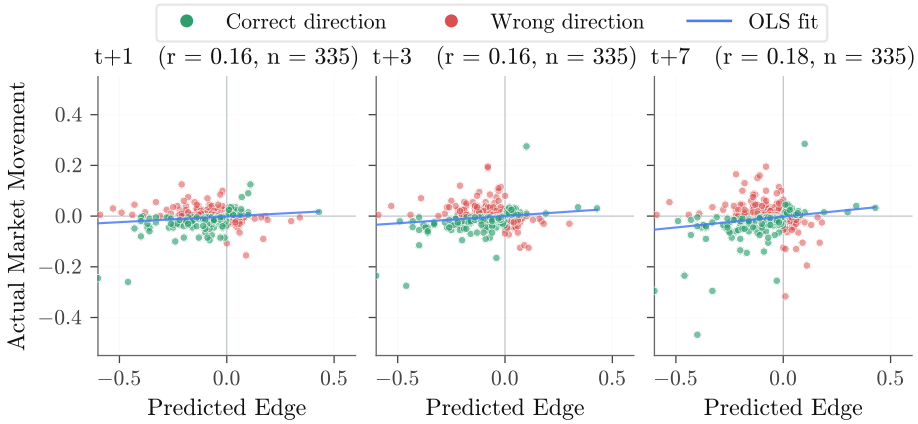

Figure 6. Independent ensemble edge ($\hat{p}_{\text{ens}} - p_m$) vs. subsequent price movement, with $t = \{1, 3, 7\}$. Correlation grows with the forecast horizon.

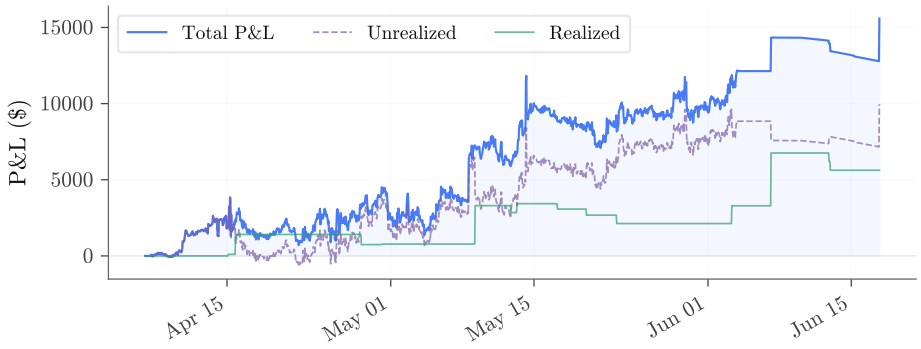

Figure 7. **Cumulative paper-trading P&L.** Positions sized by the final forecast at the prevailing market price; all trades are simulated with realistic bid/ask spreads from Kalshi's orderbook, although some markets may not support enough liquidity in real trading.