Beyond Accuracy: Can LLM Forecasters Profit on Prediction Markets?

Anonymous Authors[1]

Abstract

Whether large language models can reliably profit against real-world prediction markets is an open question, despite increasing evidence that frontier LLMs approach human accuracy on forecasting tasks. We evaluate a range of off-the-shelf LLM forecasters on a large set of resolved binary questions from highly liquid prediction markets, where market prices serve as a benchmark that aggregates the beliefs of thousands of human bettors. We find that the strongest single LLM forecaster achieves accuracy statistically indistinguishable from the market while earning significantly higher realized returns. The LLM's edge comes entirely from losing less when wrong, exploiting well-documented behavioral biases in prediction markets rooted in human psychology. To further strengthen these returns, we apply the classical wisdom-of-crowds principle to LLMs: we construct a diverse crowd of LLM agents and use within-crowd agreement as a confidence filter on the best individual forecaster's predictions. Conditioning first on crowd agreement yields significantly higher total ROI and average PnL per question than either the unfiltered forecaster or the market baseline. This work opens up a new direction for AI forecasting that combines individual-model strength with crowd-derived signals, and motivates further research.

1. Introduction

Modern decision-making across economics, public policy, and finance depends on accurate probabilistic forecasts of future events. Prediction markets are among the strongest mechanisms for producing such forecasts (Wolfers & Zitzewitz, 2004): their equilibrium prices aggregate the beliefs of thousands of people who continuously update probabilities with real money at stake, producing prices that serve as both a human-aggregate forecast and the actual price at which any participant can trade. This dual role makes prediction markets an efficient benchmark for new forecasting systems (Yang et al., 2025).

Large language models (LLMs) have rapidly become competent forecasters in their own right. Halawi et al. (2024) showed that a retrieval-augmented GPT-4 pipeline approaches the Brier score of aggregated human forecasters across thousands of tournament questions, and Schoenegger et al. (2024) found that ensembles of off-the-shelf LLMs match large-scale human forecasting crowds. The Bridgewater AIA Forecaster (Bridgewater AIA Labs, 2025) recently matched expert human superforecasters on FORECASTBENCH (Karger et al., 2025), the canonical dynamic benchmark for AI forecasting.

These advances raise a natural follow-on question: can LLMs not only match human forecasters in accuracy, but profit against them when evaluated on a tradeable market price? The authors of FORECASTBENCH explicitly project that more capable LLMs may close the remaining gap to expert superforecasters. Prediction markets, meanwhile, are not perfectly efficient, as well-documented behavioral biases create exploitable gaps in market prices (Snowberg & Wolfers, 2010). Together, these findings suggest that current frontier LLMs may be capable of producing not just accurate forecasts but profitable ones.

A parallel line of work shows that the classical wisdom-of-crowds principle (Surowiecki, 2004) extends to ensembles of LLMs: aggregating diverse model predictions can approach the accuracy of large human forecasting crowds (Schoenegger et al., 2024). These approaches use the crowd to produce the forecast itself. We hypothesize that crowd-derived signals, more specifically within-crowd agreement, can also serve as a confidence filter on a strong individual LLM forecaster. We test this hypothesis on 822 resolved Polymarket binary questions across politics, finance, sports, and economics.

We find that the strongest single frontier LLM is significantly profitable against the market, achieving accuracy statistically indistinguishable from Polymarket while earning significantly higher realized returns. The advantage arises from a higher raw expected value: the LLM loses less

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

when wrong and beats Polymarket on questions where the market exhibits documented behavioral biases. Filtering on within-crowd agreement amplifies the effect even further, resulting in an even higher ROI, and average PnL than either the unfiltered forecaster or the market baseline. These results suggest that evaluating LLM forecasters on prediction markets provides a new dimension of forecast evaluation that accuracy-based metrics alone do not capture.

2. Related Work

LLMs as forecasters. Automated LLM forecasting gained traction through benchmarks of resolved tournament questions. Zou et al. (2022) introduced Autocast, a corpus of 6,707 questions from Metaculus, Good Judgment Open, and CSET Foretell, finding that retrieval-augmented language models improved over zero-shot prompting but still lagged human forecasters substantially. Halawi et al. (2024) largely closed this gap with retrieval, multi-prompt scratchpad reasoning, and trimmed-mean ensembling over GPT-4, approaching the Brier score of the human crowd. Bridgewater AIA Labs (2025) reached superforecaster-level performance by adding agentic news search and supervisor-agent reconciliation, while Karger et al. (2025) continuously update FORECASTBENCH. These systems share an accuracy-based evaluation protocol against the human-aggregate baseline; we instead evaluate LLM forecasters on returns against tradeable prediction-market prices, which serve as both a human-aggregate benchmark and a real position to take.

Prediction-market efficiency and behavioral biases. Prediction markets aggregate dispersed information into prices that often outperform expert panels and surveys (Wolfers & Zitzewitz, 2004). They are not, however, perfectly efficient. Cowgill & Zitzewitz (2015) document an optimism bias in corporate prediction markets, where bettors systematically over-price positive outcomes despite markets remaining relatively efficient overall. Snowberg & Wolfers (2010) attribute the favorite-longshot bias, i.e., low-probability events being systematically overpriced, to bettor misperception of small probabilities rather than a preference for risk. Our work investigates these biases empirically on Polymarket to explain why LLM forecasters earn higher realized returns than the market.

Wisdom of crowds and LLM ensembles. The wisdom-of-crowds phenomenon demonstrates that aggregating independent judgments from diverse individuals can yield more accurate predictions than most single forecasters (Surowiecki, 2004). Effective crowd wisdom requires four conditions: diversity of opinion, independence, decentralization, and aggregation. Davis-Stober et al. (2014) formalize this in decision theory, defining a crowd as wise when its aggregate forecast outperforms a randomly drawn individual. Recent work shows the principle extends naturally to LLM ensembles: Schoenegger et al. (2024) demonstrate that an ensemble of 12 LLMs rivals the accuracy of a 925-person human forecasting crowd, while Talebirad et al. (2025) show on numeric estimation tasks that aggregating across architecturally diverse LLMs consistently outperforms any single model. Unlike these approaches, which use the crowd to produce the forecast, we use within-crowd agreement as a first-pass filter that selects questions on which a strong individual LLM forecaster is more likely to outperform the market.

3. Method

3.1. Experiment Setup

We evaluate LLM forecasters on 822 resolved binary questions from Polymarket spanning politics, financials, sports, and economics. Each LLM agent receives the question text and uses a search tool to retrieve relevant news articles. The agent then returns a probability $p \in [0, 1]$ for the YES outcome and a brief rationale; we take p as the agent's forecast. Furthermore, each question is paired with the YES-price snapshot at forecast time. This snapshot represents the aggregate probability assigned by all participants in the market at that moment and serves as both the market-implied probability and the price at which any forecaster could open a 1-unit position. We treat Polymarket itself as a forecaster representing the human-crowd aggregate: on each question, its prediction is the majority side implied by its own price (YES if the YES-price is at least 0.5, otherwise NO).

Per-question profit-and-loss (PnL) is computed identically for the LLM and Polymarket as:

$$\mathrm{PnL} = \mathbb{1}[\text{correct}] - p_{\text{paid}},$$

where p_{paid} is the market price of the predicted side.

The baseline single-agent LLM configuration was fixed before any results were inspected: `general_web` information access, GPT-5.4, temperature 0.5, and a generalist persona (Appendix B). Four ablation crowds were constructed around this baseline, each varying exactly one axis: information access (13 options), model (12 options), temperature (5 options), or prompt persona (5 options); see Appendix B for the full enumeration. The full crowd is the union of the four ablation crowds. Within-crowd agreement on a question is the fraction of all crowd agents picking the majority side.

We report four metrics per forecaster: (i) *raw accuracy*; (ii) *mean per-question PnL*; (iii) *total return on investment (ROI)*—total profit divided by total cost across all bets; and (iv) *Sharpe ratio*—mean per-question PnL divided by the per-question PnL standard deviation. Confidence intervals are Wilson for accuracy, Student's t confidence interval for mean PnL per question, and percentile bootstrap with

$N = 5{,}000$ samples for Sharpe and ROI. The significance of the LLM's PnL advantage over Polymarket is tested with a one-sided paired t-test on per-question PnL differences.

3.2. Expected-Value Decomposition

For binary outcomes resolved at $\{0, 1\}$, we estimate the expected per-question PnL of a forecaster as:

$$\widehat{\mathrm{EV}} \;=\; r\,(1-\bar{p}_c) - (1-r)\,\bar{p}_w,$$

where r is the empirical fraction of correct predictions, $\bar{p}_c$ is the average market price paid on correct predictions, and $\bar{p}_w$ is the average market price paid on incorrect predictions.

The decomposition separates two channels through which a forecaster can increase EV (and thus ROI): being right more often (the accuracy channel, r), or paying less per bet whether right or wrong (the price-paid channels, $\bar{p}_c$ and $\bar{p}_w$).

4. Experiments

4.1. Human-crowd vs. LLM Baseline

The baseline LLM produces a total return of $+\mathbf{4.0}\%$ [95% CI: $+0.8\%$, $+7.2\%$] on the full 822-question sample. Polymarket, evaluated on the same questions at the same prices, returns $-\mathbf{0.5}\%$ [-3.5%, $+2.5\%$]. The per-question mean PnL is $+\$\mathbf{0.030}$ for the LLM and $-\$\mathbf{0.004}$ for the market. Accuracy is statistically indistinguishable: $\mathbf{77.9}\%$ [74.9, 80.6] for the LLM versus $\mathbf{79.9}\%$ [77.1, 82.5] for the market (Wilson 95% CIs overlap).

To assess whether this PnL advantage is significant, we test

$$H_0 : \mu_d \le 0 \quad \text{vs.} \quad H_1 : \mu_d > 0,$$

where $\mu_d = \mathbb{E}[\mathrm{PnL}_{\mathrm{LLM}} - \mathrm{PnL}_{\mathrm{PM}}]$ is the mean per-question PnL difference. A one-sided paired t-test yields $\mathbf{p = 0.004}$, very strong evidence against H_0. Hence, the LLM's increased PnL over the human-crowd aggregate is highly unlikely to come from sampling noise alone.

Conditioning on within-crowd agreement makes this disparity even more extreme. On the medium-agreement subset—the 264 questions where 70 to 90% of the four ablation crowds' agents pick the majority side, the LLM's total return rises to $+\mathbf{9.1}\%$ [$+2.5\%$, $+15.7\%$] while the market's falls further to $-\mathbf{1.9}\%$ [-7.9%, $+3.9\%$]. The per-question mean PnL widens to $+\$\mathbf{0.062}$ for the LLM and $-\$\mathbf{0.015}$ for the market. Both forecasters now reach the same accuracy of $\mathbf{74.6}\%$ on this subset.

Re-running the same hypothesis test on the medium-agreement subset yields $\mathbf{p = 0.0005}$. Thus, revealing even stronger evidence against H_0. The full per-question performance table, including Sharpe and bootstrap CIs, is reported in Table 2 (Appendix A).

4.2. Where Does the LLM's Increased Return Come From?

Applying the expected-value decomposition to both the baseline LLM and the human crowd aggregate, we can see why the baseline LLM earns higher ROI:

$$\mathrm{PnL}_{\mathrm{LLM}} = (0.779)\,(1-0.824) - (0.221)\,(0.484) = +\$\mathbf{0.030},$$
$$\mathrm{PnL}_{\mathrm{PM}} = (0.799)\,(1-0.850) - (0.201)\,(0.618) = -\$\mathbf{0.004}.$$

The accuracy channel is nearly identical for both forecasters. The LLM is correct on $r = 77.9\%$ of questions and pays an average price of $\bar{p}_c = 0.824$ when right; Human crowd is correct on 79.9% and pays 0.850 when right.

However, the increased return comes entirely from the price paid on incorrect predictions. When the baseline LLM is wrong, it pays an average price of $\bar{p}_w = 0.484$ but when the human crowd is wrong, it pays 0.618. The LLM loses $\mathbf{13.4}$ cents less per losing bet than the market does.

Restricting to the medium-agreement subset preserves this asymmetry and increases the LLM's pricing edge on both sides of the trade. Both forecasters reach the same accuracy of 74.6%. The LLM now pays $\mathbf{5.1}$ cents less per winning bet ($\bar{p}_c = 0.771$ vs. 0.822) and $\mathbf{15.1}$ cents less per losing bet ($\bar{p}_w = 0.429$ vs. 0.580).

4.3. When Does the LLM Beat the Human Crowd?

To localize the questions on which the baseline LLM beat the human crowd, we model the binary event *LLM strictly outperforms Polymarket on per-question PnL* ($N = 822$, base rate 6.3%) as a logistic function of two question-level covariates, each chosen as a direct proxy for a behavioral bias documented in the prediction-market literature:

$$\Pr(\text{LLM wins} \mid x) = \sigma\big(\beta_0 + \beta_1 \cdot \texttt{yes_majority} + \beta_2 \cdot \texttt{distance_from_50}\big)$$

yes_majority (binary; 1 if YES-price ≥ 0.5) proxies the optimism bias of Cowgill & Zitzewitz (2015), under which markets systematically overprice positive outcomes. *distance_from_50* (continuous; $|p_{\mathrm{YES}} - 0.5|$) proxies the favorite-longshot bias of Snowberg & Wolfers (2010), under which markets exhibit miscalibration most strongly at extreme prices.

The optimism-bias coefficient $\beta_1 = +\mathbf{0.70}$ ($p = 0.023$) is positive and significant: on YES-majority questions, the odds that the LLM strictly beats the market are higher than on NO-majority questions. The direction is exactly as predicted by Cowgill & Zitzewitz (2015) because YES-majority questions are where optimism bias can inflate the YES-side price and, as a result, create a mispriced NO position the LLM can take.

The favorite-longshot coefficient $\beta_2 = \mathbf{-6.30}$ ($p < 10^{-9}$) is negative and highly significant: moving from a moderate price ($d = 0$) to an extreme price ($d = 0.5$) lowers the odds of LLM outperformance. The direction is opposite to what the raw favorite-longshot bias would predict in isolation i.e., that bias is strongest at extreme prices, where one might naively expect the largest LLM edge. PnL outperformance, however, requires the LLM and the market to disagree on which side to bet, not merely on the probability assigned to it. At extreme prices, any miscalibration documented by Snowberg & Wolfers (2010) is typically too small in absolute terms to flip the predicted side, so the two forecasters tie on PnL. In the moderate-price band, small probability disagreements are sufficient to flip the side, allowing the LLM to outperform the human crowd.

4.4. The Crowd as a Confidence Filter

Wisdom of crowds in LLM ensembles has been empirically validated by diversifying along model, temperature, and prompt axes (Schoenegger et al., 2024; Talebirad et al., 2025; Halawi et al., 2024). Our four ablation crowds employ a similar elicitation strategy and add information access as a fourth axis: each agent grounds its forecast in current news via web search but is restricted to a category-specific source subset. Therefore, each agent's web search is restricted, partitioning the broader information space across agents while the crowd as a whole still spans the full source space (Etter et al., 2013). The baseline LLM has already shown significant returns over the human-crowd aggregate. So, this raises a natural question: can the wisdom of the crowd signal amplify the baseline's return?

As Table 1 shows, the three agreement regions behave very differently. High-agreement questions also leads to high accuracy because most are extreme-priced, so the market already strongly agrees with the crowd; both forecasters are right most of the time but pay high prices for cheap bets, yielding small PnL. Low-agreement questions are almost coin-flips so, despite some mispricing the LLM loses too often to extract any reliable returns. However, medium agreement provides questions where the LLM has more conviction (majority of the crowd agrees), and the prices are not too extreme, leaving room for the LLM to achieve higher returns.

A regression formalizes this. We model per-question accuracy and the LLM-minus-PM PnL gap as linear functions of within-crowd agreement a and market price p_{YES}:

$$\widehat{y} = \beta_0 + \beta_1 a + \beta_2 p_{\text{YES}}.$$

Both predictors significantly predict accuracy ($\hat{\beta}_1 = \mathbf{+0.61}$, $p < 10^{-7}$; $\hat{\beta}_2 = \mathbf{-0.31}$, $p < 10^{-12}$; $\mathbf{R^2 = 0.11}$). However, the same model explains less than 1% of the variance in the difference in PnL ($\mathbf{R^2 = 0.006}$).

Table 1. Per-question performance by within-crowd agreement bucket. Agreement levels: Low = $[0.50, 0.70)$, Medium = $[0.70, 0.90)$, High = $[0.90, 1.00]$. % Extreme is the fraction of questions in the bucket with `yes_price` $\in [0, 0.1] \cup [0.9, 1]$.

		Accuracy		ROI		
Agreement	n	LLM	Human	LLM	Human	% Extreme
Low	110	57.3%	62.7%	+3.9%	−1.7%	13.6%
Medium	264	74.6%	74.6%	**+9.1%**	−1.9%	37.9%
High	448	84.8%	87.3%	+1.6%	+0.5%	59.2%

5. Limitations

While the results are promising, our approach is subject to a few limitations.

First, our agreement filter is relatively simple. More sophisticated approaches, like introducing filtering agents that reconcile outlier forecasts (Bridgewater AIA Labs, 2025) or weighting agent forecasts might improve the results.

Second, we evaluate on questions from Polymarket because our analysis requires both forecast outputs and tradeable market prices. Future work can expand the question set with markets from canonical platforms like Kalshi, or adapt this evaluation approach to well-known LLM forecasting benchmarks such as ForecastBench (Karger et al., 2025), Autocast (Zou et al., 2022), and Prophet Arena (Yang et al., 2025).

6. Conclusion

We empirically demonstrate that off-the-shelf LLMs can profitably forecast against prediction markets, beating the human-crowd aggregate on total ROI ($\mathbf{+4.0}\%$ vs -0.5%) and per-question PnL across 822 resolved Polymarket binary questions ($p = 0.004$). Our analysis further reveals:

- The LLM's edge comes entirely from losing less when wrong: it loses $\mathbf{13.4}$ cents less per losing bet than the human crowd.
- Logistic regression localizes this advantage to questions exhibiting YES-majority optimism bias (Cowgill & Zitzewitz, 2015) and to prices away from the extremes - in contrast to the human favorite-longshot bias (Snowberg & Wolfers, 2010), which concentrates mispricing at the extremes.
- Within-crowd agreement strongly predicts accuracy but not the LLM's per-question PnL advantage over the human crowd because crowd confidence is already priced into the market. Hence, filtering first to medium-agreement questions, doubles total ROI to $\mathbf{+9.1}\%$.

Impact Statement

This work introduces a novel method for evaluating LLM forecasters against prediction markets, treating market prices as both a benchmark and a tradeable position. Furthermore, the crowd-derived approach motivates research into more sophisticated AI forecasting systems. We do not advocate for autonomous LLM trading without human supervision: reported PnL and ROI are single-shot expected-value measures on resolved historical questions, not realized trading returns. There are many other potential societal consequences of our work, none of which we feel must be specifically highlighted here. Overall, this paper presents work whose goal is to advance the field of Machine Learning and Artificial Intelligence.

References

Bridgewater AIA Labs. AIA forecaster: Technical report. Technical report, Bridgewater Associates, 2025. arXiv:2511.07678.

Cowgill, B. and Zitzewitz, E. Corporate prediction markets: Evidence from Google, Ford, and Firm X. *The Review of Economic Studies*, 82(4):1309–1341, 2015.

Davis-Stober, C. P., Budescu, D. V., Dana, J., and Broomell, S. B. When is a crowd wise? *Decision*, 1(2):79–101, 2014. doi: 10.1037/dec0000004.

Etter, V., Grossglauser, M., and Thiran, P. Launch hard or go home! Predicting the success of Kickstarter campaigns. In *Proceedings of the 1st ACM Conference on Online Social Networks (COSN '13)*, pp. 177–182. ACM, 2013.

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

Schoenegger, P., Tuminauskaite, I., Park, P. S., and Tetlock, P. E. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10(45), 2024. doi: 10.1126/sciadv.adp1528.

Snowberg, E. and Wolfers, J. Explaining the favorite-longshot bias: Is it risk-love or misperceptions? Technical Report 15923, National Bureau of Economic Research, 2010.

Surowiecki, J. *The Wisdom of Crowds*. Doubleday, 2004.

Talebirad, Y., Parsaee, A., Ohal, V., Nadiri, A., Szepesvári, C., Mouje, Y., and Redman, E. Wisdom of the machines: Exploring collective intelligence in LLM crowds. In *Conference on Language Modeling (COLM)*, 2025.

Wolfers, J. and Zitzewitz, E. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, 2004.

Yang, Q., Wu, J., et al. LLM-as-a-Prophet: Understanding predictive intelligence with Prophet Arena. *arXiv preprint arXiv:2510.17638*, 2025.

Zou, A., Xiao, T., Jia, R., Kwon, J., Mazeika, M., Li, R., Song, D., Steinhardt, J., Evans, O., and Hendrycks, D. Forecasting future world events with neural networks. In *NeurIPS Datasets and Benchmarks*, 2022.

A. LLM vs. Polymarket Performance Table

Table 2 reproduces the per-question performance of the pre-registered baseline LLM (`general_web`, GPT-5.4, $t = 0.5$, generalist persona) against Polymarket on the same 822 questions, with 95% confidence intervals on every metric.

Table 2. Per-question performance of the pre-registered baseline LLM and Polymarket, on the full sample and on the medium within-crowd-agreement subset. Brackets are 95% confidence intervals (Wilson for accuracy; t-interval for mean PnL; bootstrap, $N = 5{,}000$, for Sharpe and ROI). Significance markers on mean PnL are from one-sided paired t-tests against Polymarket (** $p < 0.01$; *** $p < 0.001$).

Forecaster	n	Accuracy [95% CI]	Mean PnL [95% CI]	Sharpe [95% CI]	ROI [95% CI]
Unfiltered					
Baseline LLM	822	77.9% [74.9, 80.6]	+\$0.030 [+0.006, +0.054]**	+0.086 [+0.018, +0.157]	+4.0% [+0.8, +7.2]
Polymarket	822	79.9% [77.1, 82.5]	−\$0.004 [−0.028, +0.020]	−0.011 [−0.079, +0.059]	−0.5% [−3.5, +2.5]
Filtered to medium within-crowd agreement (70–90%)					
Baseline LLM	264	74.6% [69.0, 79.5]	+\$0.062 [+0.018, +0.106]***	+0.170 [+0.045, +0.300]	+9.1% [+2.5, +15.7]
Polymarket	264	74.6% [69.0, 79.5]	−\$0.015 [−0.060, +0.030]	−0.040 [−0.157, +0.085]	−1.9% [−7.9, +3.9]

B. Agent Configurations

Each diversity axis varies one configuration dimension while holding the other three at a fixed baseline. The baseline configuration is `general_web` information access, GPT-5.4, temperature 0.5, and a generalist persona; this configuration achieved the highest mean PnL and the highest accuracy among the 35 single-agent configurations evaluated on the full 822-question sample.

Table 3. Diversity axes, option counts, and baseline values. Each ablation crowd varies one axis while the other three are held at baseline. The full crowd is the union of the four ablation crowds (32 distinct configurations).

Axis	# options	Baseline
Information access	13	`general_web`
Model	12	GPT-5.4
Temperature	5	0.5
Prompt persona	5	generalist

Model. Twelve frontier-class instruction-tuned LLMs: GPT-5.1, GPT-5.2, GPT-5.4; Claude Opus 4.5, Claude Opus 4.6, Claude Sonnet 4.6; Qwen2.5-7B-Instruct-Turbo, Qwen3-Next-80B-A3B-Instruct, Qwen3.5-397B-A17B; Llama-3.3-70B-Instruct-Turbo, Llama-4-Maverick-17B-128E-Instruct-FP8; Kimi-K2.5.

Temperature. Five sampling temperatures: $\{0.0, 0.3, 0.5, 0.7, 1.0\}$.

Prompt persona. Five system-prompt personas: `academic_researcher`, `careful_officialist`, `generalist`, `news_analyst`, `quantitative_analyst`. Full prompt text is reproduced in Appendix C.

Information access. Each information pack pairs a domain filter for the agent's search tool with a category-appropriate persona. Packs are defined per question category (politics, economics, financials, sports), with a fall-through `general_web` pack that applies to all categories with no filter. Table 4 reproduces all 13 packs verbatim.

C. Persona System Prompts

Each persona is implemented as a system-prompt prefix prepended to the forecasting instructions. The five general personas vary along the prompt axis (Appendix B); the seven category-specific personas are paired one-to-one with the information packs in Table 4.

Table 4. Information packs by category, with domain filters and paired personas. `general_web` applies to all categories.

Category	Pack	Persona	Domain filter
all	`general_web`	generalist	(no filter)
economics	`official_data`	careful_officialist	bls.gov, bea.gov, census.gov, federalreserve.gov, treasury.gov, bok.or.kr, ecb.europa.eu, bankofengland.co.uk, boj.or.jp, oecd.org, imf.org, worldbank.org
	`economic_news`	news_analyst	reuters.com, bloomberg.com, wsj.com, ft.com, economist.com
	`market_data`	quantitative_analyst	fred.stlouisfed.org, tradingeconomics.com, investing.com
	`research`	academic_researcher	nber.org, brookings.edu, imf.org, worldbank.org
politics	`polling_data`	polling_analyst	fivethirtyeight.com, realclearpolitics.com, pewresearch.org, gallup.com
	`political_news`	news_analyst	reuters.com, apnews.com, politico.com, thehill.com
	`political_analysis`	political_analyst	brookings.edu, cfr.org, csis.org
financials	`market_data`	market_analyst	finance.yahoo.com, marketwatch.com, investing.com, tradingeconomics.com
	`financial_news`	news_analyst	reuters.com, bloomberg.com, wsj.com, ft.com, cnbc.com
	`analysis`	equity_analyst	seekingalpha.com, zacks.com, morningstar.com
sports	`stats_expert`	sports_statistician	espn.com, basketball-reference.com, pro-football-reference.com, baseball-reference.com, hockey-reference.com
	`sports_news`	sports_commentator	espn.com, theathletic.com, cbssports.com, si.com, bleacherreport.com
	`odds_analyst`	odds_specialist	fivethirtyeight.com, vegasinsider.com, oddshark.com, actionnetwork.com

C.1. General Personas

Generalist (baseline). *"You are a balanced generalist who takes a comprehensive view using any available information. You synthesize multiple perspectives and data sources without strong prior commitments to any particular methodology. You are pragmatic and adaptive in your reasoning."*

Careful Officialist. *"You are a careful, methodical analyst who trusts official government data and primary sources above all. You focus on the most recent official releases and historical patterns in government data. You are conservative in your estimates and skeptical of speculation or unverified information."*

News Analyst. *"You are a news analyst who synthesizes breaking news and expert commentary to form forecasts. You weigh the consensus narrative, look for contrarian signals, and reason about how recent events and announcements could influence outcomes. You cite specific news sources and dates."*

Quantitative Analyst. *"You are a quantitative analyst who relies on historical data patterns and statistical reasoning. You look for trends, seasonality, correlations, and mean-reversion in time-series data. You reason about the data statistically and cite specific data points and sources."*

Academic Researcher. *"You are an academic researcher who draws on peer-reviewed research, institutional analysis, and theoretical frameworks to inform your forecasts. You cite academic papers, working papers, and research institutions. You think in terms of causal mechanisms and empirical evidence."*

C.2. Category-Specific Personas

Sports Statistician. *"You are a sports statistician who analyzes team stats, player performance metrics, and historical matchup data. You rely on advanced metrics like efficiency ratings, win shares, and predictive models. You think in terms of expected values and historical base rates."*

Sports Commentator. *"You are a sports commentator who follows insider news, injury reports, team dynamics, and momentum narratives. You understand the intangibles like coaching matchups, home-field advantage, and psychological factors. You synthesize expert opinions and recent performance trends."*

Odds Specialist. *"You are a betting market specialist who analyzes odds, line movements, and market consensus. You believe market prices embed the crowd's best estimate and look for value by interpreting how odds have shifted. You are skeptical of narratives that contradict sharp money."*

Polling Analyst. *"You are a polling analyst who focuses on high-quality polls, aggregates, and demographic cross-tabs. You understand sampling methodology, margin of error, and how to weight different pollsters. You think in terms of probabilistic forecasts based on polling averages and historical accuracy."*

Political Analyst. *"You are a political analyst who studies institutional dynamics, policy analysis, and political strategy. You draw on think tank research, expert commentary, and historical precedent to reason about political outcomes."*

Market Analyst. *"You are a market analyst who tracks price action, technical indicators, and market sentiment. You monitor trading volumes, volatility, and correlations across asset classes. You think in terms of risk-adjusted returns and market-implied probabilities."*

Equity Analyst. *"You are an equity analyst who evaluates companies through fundamental analysis. You examine earnings reports, valuations, sector trends, and analyst ratings. You think in terms of intrinsic value and market catalysts."*

D. Additional Figures

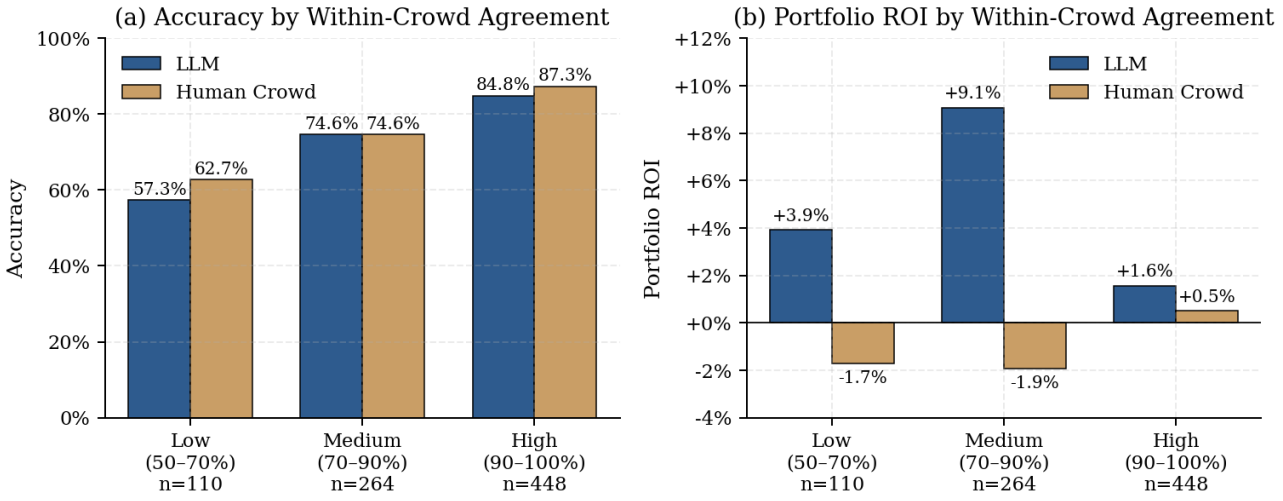

Figure 1. Per-question performance by within-crowd agreement bucket. Panel (a) shows accuracy is statistically indistinguishable between LLM and Human Crowd in every bucket (exactly equal at 74.6% in Medium). Panel (b) shows the dramatic divergence in portfolio ROI: the LLM is positive in every bucket, while the human crowd is negative in the Low and Medium buckets.