When do Prophets Profit in Prediction Markets?

Anonymous Authors[1]

Abstract

Prediction markets aggregate dispersed beliefs into prices that act as probabilistic forecasts of uncertain events. Classical theory establishes a clean equivalence between forecasting accuracy and trading profit, but only for a specific automated market maker (AMM) design. However, the largest exchanges today are based on central limit order books in which informed forecasters routinely lose money while uninformed strategies can profit using simple heuristics. We resolve this discrepancy by establishing a formal equivalence between predictive accuracy and profitability. For any strictly proper scoring rule S, we exhibit a "proper" betting strategy that depends only on the forecaster's prediction $\mathbf{p}$ and the market price $\mathbf{q}$ and earns positive expected profit whenever $\mathbf{p}$ outperforms $\mathbf{q}$ under S and the market has sufficient liquidity. The proof rests on a decomposition of expected profit that strictly generalizes the classical AMM guarantee and also explains how strategies can profit without an accuracy edge. Empirically, across thousands of forecasts from AI models, proper betting is the only strategy that reliably converts accuracy into profit. We further identify systematic forecasting personas and show how the optimal proper strategy varies across them. A month-long live deployment achieves $+80.33\%$ return on investment with a Sharpe ratio of 3.35.

1. Introduction

Prediction markets let people trade contracts whose prices reflect the market's estimated probability of an event, aggregating dispersed beliefs in the spirit of Hayek (1945)'s "marvel of the price system." Empirical studies confirm that real prediction-market prices are well calibrated and often outperform polls and expert panels (Berg et al., 2008; Rothschild, 2009; Wolfers & Zitzewitz, 2004; Arrow et al., 2008). The market scoring rule literature (Hanson, 2003; 2007; Chen & Pennock, 2007; Abernethy et al., 2013) formalizes the underlying incentive: an informed forecaster profits by trading against the market, and that trade pushes the price closer to the truth, so information aggregation and individual reward go hand in hand.

This picture, however, was developed for one specific market design: an automated market maker (AMM) (Hanson, 2003), in which an informed forecaster ensures profit by moving the market price exactly to their own forecast. The largest prediction markets today (e.g., Kalshi, Polymarket) instead use central limit order books (Kalshi, 2025; Polymarket, 2023; Ng et al., 2026), favored over AMMs for scalability, adaptability, and regulatory fit (Appendix C). Prices arise from matching opposing limit orders, and a forecaster can pick any bet size but faces whatever price impact the order book happens to deliver, plus bid-ask spreads, finite liquidity, and platform fees. Recent empirical work makes the resulting gap visible: forecasters who outperform the market on average routinely fail to convert that edge into profit (Della Vedova, 2026; Jang et al., 2025), and every agent on the largest AI forecasting benchmark loses money (Yang et al., 2025) despite several beating the market under proper scoring rules. Even in the idealized frictionless setting, intuitive heuristics produce negative profit on accurate forecasts, while simple strategies can earn positive profit without any accuracy edge (see Examples B.1 and B.2 in Appendix B). Translating accuracy into profit is therefore a problem in its own right.

Main contributions. This paper establishes the link between predictive accuracy and trading profitability in general prediction markets with an arbitrary price-impact function. (i) We exhibit a "proper" betting strategy that converts an accuracy edge under any strictly proper scoring rule S into positive expected profit (Theorem 3.2). The expected profit decomposes into three explicit terms—score gap, Bregman divergence, and liquidity loss—which together explain how strategies can profit *without* an accuracy edge (the Bregman term alone suffices) and recover the classical AMM guarantee as the boundary case where slippage exactly absorbs the Bregman bonus. (ii) We give a new characterization of proper scoring rules: S is proper iff some forecaster-and-price-only betting strategy profits whenever the fore-

[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.

caster outperforms the market under S (Theorem 3.5). We further extend both results to multi-event sequences, bid-ask spreads, and long horizons (Section 3.3). (iii) Across thousands of AI-generated forecasts, proper betting is the only strategy that reliably converts accuracy into ROI. We identify systematic forecasting personas and show how the optimal proper strategy varies across them. A month-long live deployment on Kalshi achieves $+80.33\%$ ROI with a Sharpe ratio of 3.35.

2. Preliminaries

Notation. Let $[K] := \{1, \dots, K\}$, $\Delta_{[K]} := \{\mathbf{p} \in [0,1]^K : \sum_k p_k = 1\}$ the probability simplex, $\mathbf{1}_y \in \{0,1\}^K$ the one-hot vector for outcome y, ∇G the gradient of G, and $\|\cdot\|$ the Euclidean norm.

Prediction markets. Consider an event with K disjoint outcomes and *ground truth* $\mathbf{p}^* \in \Delta_{[K]}$. For each outcome k, a contract pays 1 if k occurs and 0 otherwise; let $\mathbf{q} \in \Delta_{[K]}$ denote the price vector. Any arbitrage-free market satisfies $\sum_k q_k = 1$; bid-ask spreads are handled in Appendix E.2. A *betting strategy* is a position vector $\mathbf{s} \in \mathbb{R}^K$ where $s_k > 0$ buys and $s_k < 0$ sells outcome k; on realized y the payout is $\mathbf{s} \cdot \mathbf{1}_y = s_y$.

Finite liquidity is captured by a *price-impact function* $\rho : \mathbb{R}^K \to \Delta_{[K]}$ giving the post-trade spot price upon executing $\mathbf{s}$, with $\rho(\mathbf{0}) = \mathbf{q}$ and the monotonicity $(\rho(\mathbf{s}) - \rho(\mathbf{s}')) \cdot (\mathbf{s} - \mathbf{s}') \geq 0$. Executing $\mathbf{s}$ costs $\int_0^1 \rho(t\mathbf{s}) \cdot \mathbf{s}\, dt$, which we split into the spot-price piece $\mathbf{s} \cdot \mathbf{q}$ plus a non-negative *liquidity loss*

$$L_\rho(\mathbf{s}; \mathbf{q}) \;:=\; \int_0^1 \rho(t\mathbf{s}) \cdot \mathbf{s}\, dt - \mathbf{s} \cdot \mathbf{q} \;\geq\; 0, \quad \text{(Liquidity Loss)}$$

so the expected profit on realized $y \sim \mathbf{p}^*$ is

$$\pi(\mathbf{s}, \mathbf{p}^*) \;:=\; \mathbf{s} \cdot (\mathbf{p}^* - \mathbf{q}) - L_\rho(\mathbf{s}; \mathbf{q}). \quad \text{(Expected Profits)}$$

This abstraction covers both AMMs ($\rho = \nabla C$ for a cost function C) and central limit order books (ρ piecewise constant, read off the book); Appendix C unpacks both as special cases.

Scoring rules. Given a forecast $\mathbf{p} \in \Delta_{[K]}$ and realized $y \in [K]$, a *scoring rule* $S(\mathbf{p}, y)$ assigns a reward, with expected score $S(\mathbf{p}; \mathbf{p}^*) := \mathbb{E}_{y \sim \mathbf{p}^*}[S(\mathbf{p}, y)]$.

Definition 2.1 (Proper scoring rule)**.** S is *proper* if for all $\mathbf{p}, \mathbf{p}'$, $S(\mathbf{p}; \mathbf{p}) \geq S(\mathbf{p}'; \mathbf{p})$; *strictly proper* if strict for $\mathbf{p}' \neq \mathbf{p}$.

We say the forecaster *outperforms* the market under S if $S(\mathbf{p}; \mathbf{p}^*) > S(\mathbf{q}; \mathbf{p}^*)$. Whether this holds depends on $\mathbf{p}^*$, which is unobservable. Every proper scoring rule corresponds to a convex *potential* G:

Table 1. Common proper scoring rules with associated functions.

	$S(\mathbf{p}, y)$	$G(\mathbf{p})$	$D_G(\mathbf{q}, \mathbf{p})$	$\mathbf{s}_G(\mathbf{p}, \mathbf{q})$
Brier	$-\|\mathbf{1}_y - \mathbf{p}\|^2$	$\|\mathbf{p}\|^2 - 1$	$\|\mathbf{q} - \mathbf{p}\|^2$	$2(\mathbf{p} - \mathbf{q})$
Log	$\log p_y$	$\sum_k p_k \log p_k$	$\sum_k q_k \log \frac{q_k}{p_k}$	$\log \mathbf{p} - \log \mathbf{q}$
Spherical	$\frac{p_y}{\|\mathbf{p}\|}$	$\|\mathbf{p}\|$	$\|\mathbf{q}\| - \frac{\mathbf{p} \cdot \mathbf{q}}{\|\mathbf{p}\|}$	$\frac{\mathbf{p}}{\|\mathbf{p}\|} - \frac{\mathbf{q}}{\|\mathbf{q}\|}$

Proposition 2.2 (McCarthy, 1956a)**.** *S is (strictly) proper iff there is a (strictly) convex G with $S(\mathbf{p}, y) = G(\mathbf{p}) + \nabla G(\mathbf{p}) \cdot (\mathbf{1}_y - \mathbf{p})$.*

The associated *Bregman divergence* is $D_G(\mathbf{q}, \mathbf{p}) := G(\mathbf{q}) - G(\mathbf{p}) - \nabla G(\mathbf{p}) \cdot (\mathbf{q} - \mathbf{p}) \geq 0$, with equality iff $\mathbf{p} = \mathbf{q}$. Table 1 lists canonical examples.

3. A theory of proper betting strategies

Several natural betting strategies—highest-margin, inverse-margin, and Kelly (Kelly, 1956)—fail to convert an accuracy edge into profit (Section 4): an accurate forecaster can lose under highest-margin betting, while an inaccurate one can profit by betting in the right direction (Examples B.1 and B.2 in Appendix B). Betting optimization is thus a problem in its own right. We resolve this by defining, for every proper scoring rule S, a corresponding *proper* betting strategy.

Definition 3.1 (Proper betting strategy)**.** Under a proper scoring rule S with potential G, the *proper* betting strategy for forecast $\mathbf{p}$ and price $\mathbf{q}$ is the position vector $\mathbf{s}_G(\mathbf{p}, \mathbf{q}) := \nabla G(\mathbf{p}) - \nabla G(\mathbf{q})$.

Table 1 lists $\mathbf{s}_G$ for the three canonical rules; Figure 1 in Appendix B illustrates the Brier case geometrically.

3.1. Proper betting and profitability

Our first main result shows that the proper bet from Definition 3.1 converts an accuracy edge under S into positive expected profit, with an explicit three-term profit decomposition.

Theorem 3.2 (Forecaster's profitability guarantee)**.** *For any ground truth $\mathbf{p}^*$, forecast $\mathbf{p}$, price $\mathbf{q}$, and strictly proper S with potential G, the proper bet $\mathbf{s}^* = \mathbf{s}_G(\mathbf{p}, \mathbf{q})$ earns expected profit*

$$\pi(\mathbf{s}^*, \mathbf{p}^*) \;=\; \underbrace{S(\mathbf{p}; \mathbf{p}^*) - S(\mathbf{q}; \mathbf{p}^*)}_{\textit{score gap}} + \underbrace{D_G(\mathbf{q}, \mathbf{p})}_{\textit{Bregman divergence}} - \underbrace{L_\rho(\mathbf{s}^*; \mathbf{q})}_{\textit{liquidity loss}}.$$

In particular, if $S(\mathbf{p}; \mathbf{p}^*) > S(\mathbf{q}; \mathbf{p}^*)$, then $\pi(\mathbf{s}^*, \mathbf{p}^*) > 0$ whenever the market has sufficient liquidity, $L_\rho(\mathbf{s}^*; \mathbf{q}) < S(\mathbf{p}; \mathbf{p}^*) - S(\mathbf{q}; \mathbf{p}^*) + D_G(\mathbf{q}, \mathbf{p})$.

We defer the full proof to Appendix D.1; the key observation

Table 2. ROI (%) of baselines and the proposed Proper (Brier) strategy on the standardized 200-event shared subset. ΔS is the Brier score gap. Bold marks the best ROI per row. Full results in Appendix F.2.

		ROI (%)					
Model	ΔS	**Proper (Brier)**	**Max (Mkt)**	**Max (Grp)**	**Inv-Margin**	**Kelly-Alike**	**Kelly**
Claude Opus 4.6	$+0.0016$	**+22.1**	+5.0	-14.0	-3.5	+10.9	-99.9
Gemini 3	$+0.0008$	**+8.1**	+1.3	-0.1	+0.7	+1.8	-42.7
GPT-5.2 (Base)	-0.0347	**+4.5**	+2.4	-14.9	+1.9	-0.1	-99.9
LLaMA 4 Maverick	-0.0450	-13.7	**-4.1**	-15.5	-4.0	-10.0	-99.9
Grok 4.1 Fast	-0.0462	-11.6	-7.8	-26.4	-7.0	**-6.3**	-99.9

is a pointwise identity connecting the realized score gap and the realized payout of the proper bet:

Lemma 3.3 (Profit decomposition)**.** *For any realized y, forecast $\mathbf{p}$, price $\mathbf{q}$, and proper S,*

$$\mathbf{s}_G(\mathbf{p},\mathbf{q})\cdot(\mathbf{1}_y-\mathbf{q}) \;=\; [S(\mathbf{p},y)-S(\mathbf{q},y)]+D_G(\mathbf{q},\mathbf{p}).$$

The decomposition directly explains how an inaccurate forecaster can still profit: when $\mathbf{p}$ is far from $\mathbf{q}$, a large Bregman divergence can outweigh a negative score gap, revealing a tension between being *accurate* and being *different from the market*.

Remark 3.4 (Strict generalization of the AMM guarantee). Theorem 3.2 strictly generalizes the classical profitability guarantee for AMMs (Hanson, 2003; 2007): under an AMM with cost dual to G, the proper bet is precisely the trade that moves the spot price from $\mathbf{q}$ to $\mathbf{p}$, and its liquidity loss $L_\rho(\mathbf{s}_G;\mathbf{q}) = D_G(\mathbf{q},\mathbf{p})$ exactly absorbs the Bregman term, leaving $\pi = S(\mathbf{p};\mathbf{p}^*) - S(\mathbf{q};\mathbf{p}^*)$. In a frictionless CLOB, by contrast, the Bregman bonus survives, so the trader does *better* than the AMM under the same forecast. See Appendix C.

3.2. Properness is necessary for profitability

The proper bet $\mathbf{s}_G$ is well-defined for *any* G, not only convex G corresponding to proper rules. Among forecaster-and-price-only strategies, the proper bet is in fact the unique direction that ensures positive expected profit whenever the forecaster outperforms the market (Appendix D.2). Furthermore, properness of S is necessary for *any* such strategy to robustly profit:

Theorem 3.5 (Characterization of proper scoring rules)**.** *S is proper iff there exists a betting strategy $\mathbf{s}(\mathbf{p},\mathbf{q})$ that is nowhere locally constant in $\mathbf{q}$, depends only on $\mathbf{p},\mathbf{q}$, and has positive expected profit* for all $\mathbf{p}^*$ *for which* $S(\mathbf{p};\mathbf{p}^*) > S(\mathbf{q};\mathbf{p}^*)$.

Theorem 3.5, proven in Appendix D.3, characterizes proper scoring rules as precisely those rules that admit a profitable forecaster-and-price-only betting strategy: properness is not just sufficient but *necessary* for the accuracy-profit link.

3.3. Proper betting in real prediction markets

Three practical extensions of Theorem 3.2 are developed in Appendix E and used in our experiments.

Multiple events. With n events and empirical scores $\hat{S}_F, \hat{S}_M$, the sequential proper bet $\mathbf{s}_i := \mathbf{s}_G(\mathbf{p}_i,\mathbf{q}_i)$ has strictly positive realized return whenever $\hat{S}_F > \hat{S}_M$, bounded below by the average Bregman divergence (Corollary E.1, Appendix E.1). Two consequences: bet sizes should vary across events (proportional to $\nabla G(\mathbf{p}_i) - \nabla G(\mathbf{q}_i)$, not uniform), and positive realized return does not require beating the market—the profit can come entirely from the Bregman divergence.

Bid-ask spreads. When buy/sell prices satisfy $q_k^+ + q_k^- \geq 1$, decomposing each outcome into two binary events and applying Lemma 3.3 coordinate-wise recovers Theorem 3.2 under a strictly weaker accuracy condition that absorbs the spread cost (Appendix E.2).

Long horizons. A *fundamental* strategy holds proper bets to resolution; a *momentum* strategy rebalances each round. Both inherit the decomposition of Theorem 3.2 summed across rounds, but the momentum guarantee depends on an operationally observable per-round edge against the next-step price $\mathbf{q}^{t+1}$ rather than the unobservable $\mathbf{p}^*$ (Appendix E.3); we exploit this in our live deployment.

4. Experiments

Testing betting strategies requires data pairing forecasts with contemporaneous market prices and realized outcomes. Recent AI forecasting benchmarks built on real-money prediction markets—most notably Prophet Arena (Yang et al., 2025)—provide such data at scale. We evaluate proper betting in two stages: an offline backtest on archived Kalshi markets, and a live deployment under real-market frictions. Implementation details and full per-model results are in Appendices F.1, F.2, F.4, F.6 and F.7.

4.1. How good are proper betting strategies?

We use forecasts on 2,418 Kalshi markets (Kalshi, 2025) collected through Prophet Arena, spanning sports, politics,

Table 3. ROI (%) decomposition of proper betting strategies. ΔS is the aggregate score gap between forecast and market; D is total Bregman divergence. All quantities are summed across bets, normalized by total cost staked under each rule, and multiplied by 100, so $\Delta S + D = \text{ROI}$.

	Brier			**Log**			**Spherical**		
Model	ΔS	D	**ROI**	ΔS	D	**ROI**	ΔS	D	**ROI**
Claude Opus 4.6	$+3.2$	$+17.8$	$+21.0$	$+4.5$	$+16.9$	$+21.4$	-1.3	$+10.1$	$+8.8$
Gemini 3	$+4.0$	$+5.0$	$+9.0$	$+3.9$	$+5.0$	$+8.9$	-2.4	$+2.9$	$+0.4$
GPT-5.2 (Base)	-108.4	$+112.7$	$+4.3$	-5.8	$+15.9$	$+10.1$	-52.7	$+46.7$	-6.0
LLaMA 4 Maverick	-123.7	$+110.6$	-13.1	-21.0	$+20.2$	-0.8	-53.3	$+38.5$	-14.8
Grok 4.1 Fast	-137.4	$+128.3$	-9.1	-20.3	$+23.8$	$+3.5$	-89.3	$+74.3$	-15.0

economics, and crypto. For each model and strategy, we apply the strategy to historical forecast–price pairs and compute realized ROI. We compare proper betting strategies against four heuristics: **Max-Margin** (concentrates capital on the largest disagreements), **Inverse-Margin**, **Kelly-Alike**, and **Kelly Criterion** (Kelly, 1956); see Appendix F.2 for heuristic strategy details.

Table 2 shows that heuristic allocations fail to yield consistent positive returns, while Brier-weighted proper betting is the only strategy that reliably produces positive ROI for stronger models. Kelly is particularly unstable under miscalibrated probabilities (Appendix F.2). These results show that *how* probabilities are translated into position sizes is a key determinant of returns; predictive accuracy alone is insufficient.

4.2. How to choose a proper betting strategy?

The decomposition in Lemma 3.3 reveals that profitability depends not only on the score gap ΔS but also on the divergence term D, which is rule-dependent. Table 3 reports the two terms summed across bets and normalized by total spend so $\Delta S + D = \text{ROI}$. Models with worse accuracy can still achieve higher ROI when they generate larger divergence (GPT-5.2 (Base) beats Gemini 3 under Log), and a model's score under a rule does not determine which rule yields the best returns (LLaMA 4 Maverick: Brier has a worse gap than Spherical but higher ROI, driven by a larger divergence term).

Forecaster personas. To identify when each proper rule is preferable, we construct four synthetic "personas" that share the same Brier score on the same markets but differ in (i) the share of small-margin bets ($|p-q| \leq 0.15$) and (ii) the slope of directional win rate against margin: **Conservative** (small margins, flat accuracy), **Aggressive** (fewer small-margin bets, flat accuracy), **Dispersed** (broadly spread, gradual decline), and **Brittle** (small margins but sharply declining accuracy). Evaluated under ± 0.05 Brier gaps (full results in Appendices F.4 and F.5), *Brier* is optimal when the model outperforms (positive gaps imply expected returns are positive across margins, so linear scaling in $|p-q|$ pays off), while *Log* is optimal when the model underperforms (Log's sublinear weighting attenuates costly high-margin mistakes). Mapping real models to personas via small-margin share and win-rate slope (Appendix F.6) cleanly recovers Brittle, Dispersed, and Conservative groupings, letting us pick a proper rule per model from observable behavior alone.

4.3. Live deployment

We further evaluate proper betting in a live portfolio setting using the momentum-driven strategy from Section 3.3. Live execution must contend with bid–ask spreads, limited liquidity, slippage, and discrete tick sizes. We deploy Gemini 3 with the Brier proper bet on Kalshi for 26 days (April–May 2026) with an initial budget of \$200, operating on a two-hour cadence over markets satisfying an ex-ante eligibility filter (Appendix F.7.2). The agent executed 236 orders across 129 markets, achieving an ROI of **+80.33%** (decomposed under Brier as $\Delta S = +0.7205$, $D = +0.0828$) and a Sharpe ratio of **3.35**. Almost all gains arise from predictive accuracy rather than divergence, consistent with the Conservative persona we identify for Gemini 3 in Appendix F.6. Full implementation, representative trades, and trading logs in Appendix F.7.

5. Conclusion

We resolve the puzzle that AI forecasters can outperform prediction markets in accuracy yet still lose money, by establishing a formal equivalence between predictive accuracy and trading profitability under arbitrary price-impact functions. What is more striking is the magnitude of returns and the practical viability of the framework in real markets—a month-long live deployment achieved $+80.33\%$ ROI with a Sharpe ratio of 3.35.

Several future directions remain open. First, our persona analysis in Appendix F.6 hints at the potential of *data-adaptive scoring-rule selection*: identifying the best rule for each forecaster from observable behavior alone. Second, our guarantees concern *expected* profit; extending them to risk-adjusted criteria would narrow the gap between theory and live deployment.

References

Abernethy, J., Chen, Y., and Vaughan, J. W. Efficient market making via convex optimization, and a connection to online learning. *ACM Transactions on Economics and Computation (TEAC)*, 1(2):1–39, 2013.

Arrow, K. J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J. O., Levmore, S., Litan, R., Milgrom, P., Nelson, F. D., et al. *The Promise of Prediction Markets*. Science and Technology Policy Institute, 2008.

Berg, J., Forsythe, R., Nelson, F., and Rietz, T. Results from a dozen years of election futures markets research. *Handbook of experimental economics results*, 1:742–751, 2008.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Chen, Y. and Pennock, D. M. A utility framework for bounded-loss market makers. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 49–56, 2007.

Chen, Y. and Vaughan, J. W. A new understanding of prediction markets via no-regret learning. In *Proceedings of the 11th ACM conference on Electronic commerce*, pp. 189–198, 2010.

Chen, Y., Dimitrov, S., Sami, R., Reeves, D. M., Pennock, D. M., Hanson, R. D., Fortnow, L., and Gonen, R. Gaming prediction markets: Equilibrium strategies with a market maker. *Algorithmica*, 58(4):930–969, 2010.

Cover, T. M. Universal portfolios. *Mathematical finance*, 1 (1):1–29, 1991.

Della Vedova, J. Who profits from prediction markets? execution, not information. 2026.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Good, I. J. Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114, 1952.

Hanson, R. Combinatorial information market design. In *Proceedings of the 5th ACM Conference on Electronic Commerce (EC)*, pp. 41–47, 2003.

Hanson, R. Logarithmic market scoring rules for modular combinatorial information aggregation. *Journal of Prediction Markets*, 1(1):3–15, 2007.

Hayek, F. A. The use of knowledge in society. *The American Economic Review*, 35(4):519–530, 1945.

Jang, Y., Kim, J., and Zhang, B.-T. The losing winner: An llm agent that predicts the market but loses money. In *Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative AI in Finance*, 2025.

Kalshi. Kalshi rulebook and contract specifications. `https://kalshi.com/regulatory/rulebook`, 2025. Accessed: 2025-09-23.

Kelly, J. L. A new interpretation of information rate. *the bell system technical journal*, 35(4):917–926, 1956.

Lambert, N. S., Pennock, D. M., and Shoham, Y. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pp. 129–138, 2008.

MacLean, L. C., Thorp, E. O., and Ziemba, W. T. *The Kelly capital growth investment criterion: Theory and practice*, volume 3. world scientific, 2011.

McCarthy, J. Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42(9):654–655, 1956a.

McCarthy, J. Measures of the value of information. In *Proceedings of the National Academy of Sciences Symposium on Information Theory*, pp. 654–655, 1956b.

Ng, H., Peng, L., Tao, Y., and Zhou, D. Price discovery and trading in modern prediction markets. *Available at SSRN*, 2026.

Polymarket. Polymarket CLOB documentation. `https://docs.polymarket.com/concepts/prices-orderbook`, 2023. Accessed: 2026.

Rothschild, D. Forecasting elections: Comparing prediction markets, polls, and their biases. *Public Opinion Quarterly*, 73(5):895–916, 2009.

Savage, L. J. *Elicitation of personal probabilities and expectations*. Holt, Rinehart and Winston, 1971.

Thorp, E. O. Portfolio choice and the kelly criterion. In *Stochastic optimization models in finance*, pp. 599–619. Elsevier, 1975.

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

Yang, Q., Mahns, S., Li, S., Gu, A., Wu, J., and Xu, H. Llm-as-a-prophet: Understanding predictive intelligence with prophet arena. *arXiv preprint arXiv:2510.17638*, 2025.

A. Related work

Proper scoring rules. Proper scoring rules originate in the statistical forecasting literature as a mechanism for eliciting truthful probabilistic predictions (Brier, 1950; Good, 1952; McCarthy, 1956b; Savage, 1971). Gneiting & Raftery (2007) formalize strictly proper scoring rules and their relationship to Bregman divergences and convex analysis. Subsequent work extends these ideas to more general prediction settings, including decision-theoretic elicitation (Lambert et al., 2008) and mechanism design (Chen & Vaughan, 2010).

Prediction markets. Prediction markets themselves are a well-studied mechanism for aggregating dispersed beliefs into a consensus probability (Wolfers & Zitzewitz, 2004; Arrow et al., 2008). Hanson (2003; 2007) introduced market scoring rules, which use a proper scoring rule to automate market making and provide subsidized liquidity; this framework underpins modern implementations such as the Logarithmic Market Scoring Rule (LMSR) and its derivatives (Chen & Pennock, 2007; Abernethy et al., 2013). A parallel line of work examines price formation and information aggregation under strategic trading (Chen et al., 2010). Empirical studies of real markets (Berg et al., 2008; Rothschild, 2009) have documented that prices are calibrated and often outperform expert forecasts.

Betting strategies. The problem of translating a probabilistic forecast into a bet size has a long history, beginning with the Kelly criterion (Kelly, 1956; Thorp, 1975; MacLean et al., 2011), which maximizes long-run log-wealth under known margin. More recent work has connected portfolio selection to online learning and proper scoring rules (Cover, 1991; Abernethy et al., 2013), but typically assumes a single scoring-rule objective rather than asking which scoring rule best converts predictions into profit. At the same time, recent studies have highlighted a paradox in which models achieve strong predictive performance yet fail to generate positive returns (Della Vedova, 2026; Jang et al., 2025).

B. Motivating examples

The two examples below show that natural betting heuristics break the accuracy–profit link in both directions; they are referenced from Section 1 and Section 3.

Example B.1 (Accurate forecaster can lose)**.** Consider an event with three outcomes and let $\mathbf{p} = (0.61, 0.39, 0.00)$, $\mathbf{q} = (0.10, 0.89, 0.01)$, and $\mathbf{p}^* = (0.05, 0.10, 0.85)$. Under the quadratic scoring rule, since $\|\mathbf{p}^* - \mathbf{p}\| < \|\mathbf{p}^* - \mathbf{q}\|$, the forecaster outperforms the market. If the forecaster only places a bet on the highest-margin first outcome, their expected profit is $(1, 0, 0) \cdot (\mathbf{p}^* - \mathbf{q}) < 0$.

Example B.2 (Inaccurate forecaster can profit)**.** Consider an event with two outcomes and let $\mathbf{p} = (0.9, 0.1)$, $\mathbf{q} = (0.5, 0.5)$, and $\mathbf{p}^* = (0.6, 0.4)$. Under the quadratic scoring rule, since $\|\mathbf{p}^* - \mathbf{p}\| > \|\mathbf{p}^* - \mathbf{q}\|$, the forecaster underperforms the market. If the forecaster places a bet on the first outcome, which is a highest-margin outcome, their expected profit is $(1, 0) \cdot (\mathbf{p}^* - \mathbf{q}) > 0$.

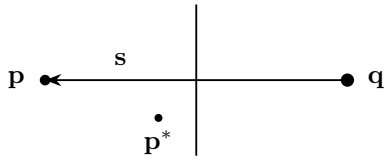


Figure 1. For the quadratic scoring rule, $S(\mathbf{p}; \mathbf{p}^*) > S(\mathbf{q}; \mathbf{p}^*)$ is equivalent to $\mathbf{p}^*$ being closer to $\mathbf{p}$ than $\mathbf{q}$. The proper betting strategy is in the direction of $\mathbf{p} - \mathbf{q}$.

C. Implementations of prediction markets

A prediction market needs an underlying matching mechanism that determines how buy and sell orders interact and at what prices; the choice shapes the liquidity, transaction costs, and incentives forecasters face. Two design families dominate practice—automated market makers and central limit order books—and both are recovered as special cases of the price-impact-function abstraction in Section 2 via specific choices of ρ and L_ρ.

Automated market makers. A classic implementation of a prediction market is an *automated market maker* (AMM) governed by a cost function (Hanson, 2003; 2007; Chen & Pennock, 2007; Abernethy et al., 2013). The AMM maintains a

vector $\mathbf{x} \in \mathbb{R}^K$ recording the net number of shares of each outcome it has sold so far, and commits to a strictly convex, differentiable *cost function* $C : \mathbb{R}^K \to \mathbb{R}$. A forecaster who executes the position vector $\mathbf{s} \in \mathbb{R}^K$ pays the total cost $C(\mathbf{x}+\mathbf{s}) - C(\mathbf{x})$ of moving the share state to $\mathbf{x}+\mathbf{s}$, and receives the payout $\mathbf{s} \cdot \mathbf{1}_y$ for a realized outcome y, with expected profit $\pi^{\text{AMM}}(\mathbf{s}, \mathbf{p}^*) := \mathbf{s} \cdot \mathbf{p}^* - (C(\mathbf{x}+\mathbf{s}) - C(\mathbf{x}))$. The *spot price* is the marginal cost $\mathbf{q} = \nabla C(\mathbf{x})$, and we can assume C is normalized so that $\nabla C(\mathbf{x}) \in \Delta_{[K]}$ for all $\mathbf{x}$. In the price-impact notation of Section 2, an AMM corresponds to $\rho(\mathbf{s}) = \nabla C(\mathbf{x}+\mathbf{s})$ and $L_\rho(\mathbf{s}; \mathbf{q}) = D_C(\mathbf{x}+\mathbf{s}, \mathbf{x})$.

The key structural property of this construction is that the convex conjugate of C on the simplex, $G(\mathbf{p}) = \sup_{\mathbf{x} \in \mathbb{R}^K} \{\mathbf{p} \cdot \mathbf{x} - C(\mathbf{x})\}$, is the convex potential that induces a strictly proper scoring rule $S(\mathbf{p}, y) = G(\mathbf{p}) + \nabla G(\mathbf{p}) \cdot (\mathbf{1}_y - \mathbf{p})$. Under this correspondence the forecaster's expected profit is exactly the score gap, which yields the classical profitability guarantee for informed forecasters:

Proposition C.1 (Forecaster's profitability guarantee in AMMs)**.** *A forecaster who executes the position vector $\mathbf{s}_{\mathbf{q}\to\mathbf{p}}$ that moves the spot price from $\mathbf{q}$ to a target $\mathbf{p}$ earns expected profit*

$$\pi^{AMM}(\mathbf{s}_{\mathbf{q}\to\mathbf{p}}, \mathbf{p}^*) \;=\; S(\mathbf{p}; \mathbf{p}^*) - S(\mathbf{q}; \mathbf{p}^*).$$

In particular, a forecaster who outperforms the market under S earns positive expected profit.

Central limit order books. The largest prediction markets today (e.g. Kalshi, Polymarket) instead adopt a *central limit order book* (CLOB), the matching mechanism standard in equity and futures exchanges: opposing limit orders are matched directly between participants, and liquidity is supplied dynamically by the resting orders rather than by a designated maker. CLOBs are favored over AMMs in practice because the cost-function approach is hard to scale and operate: liquidity must be subsidized up front through a fixed parameter, making thousands of simultaneous markets infeasible; the cost function cannot be updated post-deployment, whereas CLOB makers adjust quotes to news and adverse selection in real time; and derivatives regulations—such as the CFTC's, under which Kalshi operates as a designated contract market—are written around order-book mechanics rather than algorithmic counterparties. Polymarket itself migrated from an AMM to a CLOB after launch (Polymarket, 2023; Ng et al., 2026), citing tighter spreads, native limit-order support, and scalability without per-market subsidy. In a CLOB, ρ is the piecewise-constant function read directly off the order book, and L_ρ captures the depth-induced slippage as the trader walks the book.

Theorem 3.2 is a strict generalization of Proposition C.1. Theorem 3.2 recovers Proposition C.1 as the boundary case in which the AMM's liquidity loss exactly absorbs the Bregman bonus. Specializing to an AMM with cost function C dual to G, the proper bet $\mathbf{s}_G(\mathbf{p}, \mathbf{q}) = \nabla G(\mathbf{p}) - \nabla G(\mathbf{q})$ is precisely the trade $\mathbf{s}_{\mathbf{q}\to\mathbf{p}}$ that moves the AMM's spot price from $\mathbf{q}$ to $\mathbf{p}$, and its liquidity loss

$$L_\rho(\mathbf{s}_G; \mathbf{q}) \;=\; D_C(\mathbf{x}+\mathbf{s}_G,\, \mathbf{x}) \;=\; D_G(\mathbf{q}, \mathbf{p})$$

exactly matches the Bregman-bonus term. The two terms cancel in Theorem 3.2, leaving $\pi(\mathbf{s}_G,\, \mathbf{p}^*) = S(\mathbf{p}; \mathbf{p}^*) - S(\mathbf{q}; \mathbf{p}^*)$, which is exactly Proposition C.1. Theorem 3.2 strictly generalizes Proposition C.1 to handle arbitrary price-impact function ρ.

D. Omitted propositions and proofs

D.1. Proper betting strategies ensure positive profit

Proof of Lemma 3.3. We compute

$$\begin{aligned}
S(\mathbf{p}, y) - S(\mathbf{q}, y) &= (G(\mathbf{p}) - G(\mathbf{q})) + \nabla G(\mathbf{p}) \cdot (\mathbf{1}_y - \mathbf{p}) - \nabla G(\mathbf{q}) \cdot (\mathbf{1}_y - \mathbf{q}) \\
&= (-\nabla G(\mathbf{p}) \cdot (\mathbf{q} - \mathbf{p}) - D_G(\mathbf{q}, \mathbf{p})) + \nabla G(\mathbf{p}) \cdot (\mathbf{1}_y - \mathbf{p}) - \nabla G(\mathbf{q}) \cdot (\mathbf{1}_y - \mathbf{q}) \\
&= (\nabla G(\mathbf{p}) - \nabla G(\mathbf{q})) \cdot (\mathbf{1}_y - \mathbf{q}) - D_G(\mathbf{q}, \mathbf{p}) \\
&= \mathbf{s} \cdot (\mathbf{1}_y - \mathbf{q}) - D_G(\mathbf{q}, \mathbf{p})
\end{aligned}$$

where

$$D_G(\mathbf{q}, \mathbf{p}) = G(\mathbf{q}) - G(\mathbf{p}) - \nabla G(\mathbf{p}) \cdot (\mathbf{q} - \mathbf{p})$$

is the *Bregman divergence* between $\mathbf{p}$ and $\mathbf{q}$. □

Proof of Theorem 3.2. Averaging Lemma 3.3 over $y \sim \mathbf{p}^*$ yields the idealized (frictionless) expected return of the proper bet,

$$\mathbf{s}_G(\mathbf{p}, \mathbf{q}) \cdot (\mathbf{p}^* - \mathbf{q}) \;=\; \underbrace{S(\mathbf{p}; \mathbf{p}^*) - S(\mathbf{q}; \mathbf{p}^*)}_{\text{score gap}} + \underbrace{D_G(\mathbf{q}, \mathbf{p})}_{\text{Bregman divergence}} .$$

Subtracting the liquidity loss $L_\rho(\mathbf{s}^*; \mathbf{q})$ incurred by executing $\mathbf{s}^* = \mathbf{s}_G(\mathbf{p}, \mathbf{q})$ along the price-impact trajectory gives the expected profit

$$\pi(\mathbf{s}^*, \mathbf{p}^*) \;=\; \mathbf{s}^* \cdot (\mathbf{p}^* - \mathbf{q}) - L_\rho(\mathbf{s}^*; \mathbf{q}) \;=\; (S(\mathbf{p}; \mathbf{p}^*) - S(\mathbf{q}; \mathbf{p}^*)) + D_G(\mathbf{q}, \mathbf{p}) - L_\rho(\mathbf{s}^*; \mathbf{q}),$$

which is the decomposition stated in the theorem. Since $D_G(\mathbf{q}, \mathbf{p}) > 0$ for $\mathbf{q} \neq \mathbf{p}$ by strict convexity of G, the expected profit is strictly positive whenever the score gap together with the Bregman bonus exceeds the liquidity loss, namely

$$L_\rho(\mathbf{s}^*; \mathbf{q}) < S(\mathbf{p}; \mathbf{p}^*) - S(\mathbf{q}; \mathbf{p}^*) + D_G(\mathbf{q}, \mathbf{p}).$$

□

D.2. Proper betting strategies uniquely ensure positive profit

In this section we prove in a restricted sense that there do not exist other betting strategies $\mathbf{s}(\mathbf{p}, \mathbf{q})$ that only depend on $\mathbf{p}$ and $\mathbf{q}$ that guarantee positive expected profit. We first give a rather technical definition for when two betting strategies are intrinsically different.

Definition D.1 (Intrinsically different betting strategies)**.** We say that two betting strategies $\mathbf{s}(\mathbf{p}, \mathbf{q})$ and $\mathbf{s}'(\mathbf{p}, \mathbf{q})$ are *intrinsically different* if

$$\lim_{\mathbf{p} \to \mathbf{q}} \frac{\mathbf{s}^\perp}{\|\mathbf{s}^\perp\|} \neq \lim_{\mathbf{p} \to \mathbf{q}} \frac{\mathbf{s}'^\perp}{\|\mathbf{s}'^\perp\|} \tag{1}$$

where $\mathbf{s}^\perp$ denotes the projection of $\mathbf{s}$ onto the subspace orthogonal to the unit bet $\mathbf{1} \in \mathbb{R}^K$. The inequality (1) includes cases when $\lim_{\mathbf{p} \to \mathbf{q}} \frac{\mathbf{s}^\perp}{\|\mathbf{s}^\perp\|}$ does not exist.

We argue that the technicalities present in Definition D.1 are unavoidable. For one, note that starting with the proper scoring rule $\mathbf{s}_G$ and perturbing $\mathbf{s}_G(\mathbf{p}, \mathbf{q})$ infinitesimally for any single pair $(\mathbf{p}, \mathbf{q})$ for which $\mathbf{s}$ ensures strictly positive profit will also yield positive profit, so any definition of *intrinsically different* must be robust to such perturbations. The reason why we scale $\mathbf{s}$ to have unit norm is that the sign of the profit remains invariant under scaling. Finally, the following remark explains the projection onto a subspace:

Remark D.2 (Unit bet). Since the ground truth $\mathbf{p}^*$ and the prediction market price $\mathbf{q}$ are probabilities, we have $\mathbf{1}^\intercal(\mathbf{p}^* - \mathbf{q}) = 0$ hence any betting strategy $\mathbf{s}$ achieves the same profit as $\mathbf{s} + c \cdot \mathbf{1}$ for any $c \in \mathbb{R}$. Without loss of generality it suffices to consider betting strategies restricted to the subspace $\{\mathbf{s} \in \mathbb{R}^K : \mathbf{1}^\intercal \mathbf{s} = 0\}$ orthogonal to the unit bet $\mathbf{1}$. This explains the projection condition in Proposition D.3.

The main result of this section is that any betting strategy that is intrinsically different from the proper one does not guarantee positive profit for accurate forecasts:

Proposition D.3 (Intrinsically different strategy can yield negative profit under accurate prediction)**.** *Fix a scoring rule S with differentiable potential function G, a market price $\mathbf{q} \in \operatorname{int}(\Delta_{[K]})$, and a betting strategy $\mathbf{s}(\mathbf{p}, \mathbf{q})$. If $\mathbf{s}$ is intrinsically different from the proper betting strategy $\mathbf{s}_G$, then there exists a forecaster prediction $\mathbf{p}$ and a ground truth $\mathbf{p}^*$ for which $S(\mathbf{p}; \mathbf{p}^*) > S(\mathbf{q}; \mathbf{p}^*)$ but $\mathbf{s} \cdot (\mathbf{p}^* - \mathbf{q}) < 0$.*

Before proving Proposition D.3, we explain why we need the limit condition in Definition D.1:

Example D.4 (Necessity of the limit)**.** Without the limit condition, Proposition D.3 is false in the sense that for a fixed $\mathbf{p}$ it is not true that $\mathbf{s}_G$ is the unique betting strategy that ensures positive profit. We construct the following counterexample: let S be the quadratic scoring rule, so that $\mathbf{s}_G = 2 \cdot (\mathbf{p} - \mathbf{q})$, $D_G(\mathbf{q}, \mathbf{p}) = \|\mathbf{q} - \mathbf{p}\|^2$, and the condition $S(\mathbf{p}; \mathbf{p}^*) > S(\mathbf{q}; \mathbf{p}^*)$ is equivalent to $\|\mathbf{p} - \mathbf{p}^*\| \leq \|\mathbf{q} - \mathbf{p}^*\|$. Since $\Delta_{[K]}$ is bounded, we have

$$\inf_{\substack{\mathbf{p}^* \in \Delta_{[K]} \\ S(\mathbf{p}; \mathbf{p}^*) \geq S(\mathbf{q}; \mathbf{p}^*)}} \mathbf{s}_G \cdot (\mathbf{p}^* - \mathbf{q}) > 0.$$

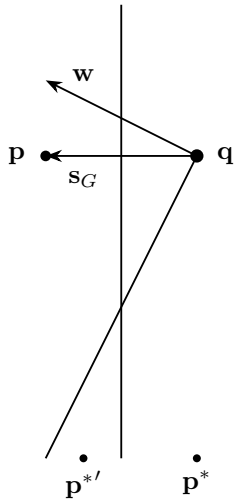


Figure 2. Picture of the proof of Proposition D.3 for the quadratic scoring rule. We need to choose $\mathbf{p}^{*\prime}$ so that $\mathbf{w} \cdot (\mathbf{p}^{*\prime} - \mathbf{q}) < 0$ but $\mathbf{s}_G \cdot (\mathbf{p}^{*\prime} - \mathbf{q}) > 0$.

Hence by perturbing $\mathbf{s}_G$ by an infinitesimal amount and using the fact that $\mathbf{p}^*, \mathbf{q} \in \Delta_{[K]}$ are bounded, we can construct $\mathbf{s}$ with $\frac{\mathbf{s}^\perp}{\|s\|^\perp} \neq \frac{\mathbf{s}_G^\perp}{\|\mathbf{s}_G^\perp\|}$ for which

$$\inf_{\substack{\mathbf{p}^* \in \Delta_{[K]} \\ S(\mathbf{p};\mathbf{p}^*) \geq S(\mathbf{q};\mathbf{p}^*)}} \mathbf{s} \cdot (\mathbf{p}^* - \mathbf{q}) > 0.$$

Proof of Proposition D.3. Since

$$\lim_{\mathbf{p} \to \mathbf{q}} \frac{\mathbf{s}^\perp}{\|\mathbf{s}^\perp\|} \neq \lim_{\mathbf{p} \to \mathbf{q}} \frac{\mathbf{s}_G^\perp}{\|\mathbf{s}_G^\perp\|}$$

and the space of unit vectors is compact, there exists a direction $\mathbf{w}$ with $\|\mathbf{w}\| = 1$ in the subspace $\{\mathbf{s} \in \mathbb{R}^K : \mathbf{1}^\intercal \mathbf{s} = 0\}$ distinct from $\pm \frac{\mathbf{s}_G^\perp}{\|\mathbf{s}_G^\perp\|}$ that is the limit of $\frac{\mathbf{s}^\perp}{\|\mathbf{s}^\perp\|}$ for a subsequence of predictions $\mathbf{p}$ approaching $\mathbf{q}$. Since $\mathbf{q} \in \text{int}(\Delta_K)$ and $\mathbf{p} \to \mathbf{q}$, for all directions in $\mathbf{v}$ in the subspace orthogonal to the unit bet, there exists $\mathbf{p}^*$ such that the vector $\mathbf{p}^* - \mathbf{q}$ has direction $\frac{\mathbf{p}^* - \mathbf{q}}{\|\mathbf{p}^* - \mathbf{q}\|} = \mathbf{v}$ and $\|\mathbf{p}^* - \mathbf{q}\|$ is an arbitrarily large fraction of $\|\mathbf{p} - \mathbf{q}\|$. We apply this to find two ground truths associated to any two opposite directions $\pm\mathbf{v}$ perpendicular to $\mathbf{s}_G$. Since the direction of $\mathbf{w}$ is distinct from that of $\mathbf{s}_G$, one of these ground truths $\mathbf{p}^*$ will satisfy $\mathbf{w} \cdot (\mathbf{p}^* - \mathbf{q}) < 0$.

Note that $\|\mathbf{p}^* - \mathbf{q}\|$ is arbitrarily large compared to $\|\mathbf{p} - \mathbf{q}\|$ as $\mathbf{p} \to \mathbf{q}$ since $\mathbf{s}$ is continuous. For the above subsequence of predictions $\mathbf{p}$ whose betting strategy direction $\frac{\mathbf{s}^\perp}{\|\mathbf{s}^\perp\|}$ approaches $\mathbf{w}$, eventually for all $\mathbf{p}$ in this sequence we will have that $\frac{\mathbf{s}}{\|\mathbf{s}\|} \cdot (\mathbf{p}^* - \mathbf{q})$ is uniformly bounded away from 0, namely less than c for some $c < 0$.

By construction of $\mathbf{p}^*$, we have $\mathbf{s}_G \cdot (\mathbf{p}^* - \mathbf{q}) = 0$. Again utilizing the fact that $\|\mathbf{p}^* - \mathbf{q}\|$ is arbitrarily large compared to $\|\mathbf{p} - \mathbf{q}\|$ and therefore arbitrarily large compared to $D_G(\mathbf{q}, \mathbf{p})$ as well since D_G is continuous, there exists a perturbation $\mathbf{p}^{*\prime}$ of $\mathbf{p}^*$ so that

$$S(\mathbf{p}; \mathbf{p}^{*\prime}) - S(\mathbf{q}; \mathbf{p}^{*\prime}) = \mathbf{s}_G \cdot (\mathbf{p}^{*\prime} - \mathbf{q}) - D_G(\mathbf{q}, \mathbf{p}) > 0$$

but $\mathbf{s} \cdot (\mathbf{p}^{*\prime} - \mathbf{q}) < -\frac{c}{2} < 0$. The result now follows using this perturbed $\mathbf{p}^{*\prime}$ as the ground truth.

□

D.3. Characterization of proper scoring rules for prediction market profitability

First we show that properness of the scoring rule is necessary for the proper betting strategy $\mathbf{s} = \nabla G(\mathbf{p}) - \nabla G(\mathbf{q})$ to always generate nonnegative expected profit.

Proposition D.5 (Proper scoring rules are necessary for proper betting strategy to ensure nonnegative return)**.** *Let*

$$S(\mathbf{p}, y) = G(\mathbf{p}) + \nabla G(\mathbf{p}) \cdot (\mathbf{1}_y - \mathbf{p})$$

be a scoring rule derived from a differentiable function G that is not convex. Then there exist values of $\mathbf{p}$, $\mathbf{q}$, and $\mathbf{p}^*$ such that

$$\mathbb{E}_{y\sim\mathbf{p}^*}\left[S(\mathbf{p}, y)\right] > \mathbb{E}_{y\sim\mathbf{p}^*}\left[S(\mathbf{q}, y)\right]$$

but the betting strategy $\mathbf{s} = \nabla G(\mathbf{p}) - \nabla G(\mathbf{q})$ has negative return $\mathbf{s} \cdot (\mathbf{p}^* - \mathbf{q}) < 0$.

Proof. Since G is not convex, there must exist some $\mathbf{p}$ and $\mathbf{q}$ such that $\mathbf{q} \in \text{int}(\Delta_{[K]})$, $\nabla G(\mathbf{q}) \neq \nabla G(\mathbf{p})$, and

$$D_G(\mathbf{q}, \mathbf{p}) = G(\mathbf{q}) - G(\mathbf{p}) - \nabla G(\mathbf{p}) \cdot (\mathbf{q} - \mathbf{p}) = -\kappa < 0. \tag{2}$$

We claim that we can choose $\mathbf{p}^*$ so that

$$-\frac{\kappa}{2} < \mathbf{s} \cdot (\mathbf{p}^* - \mathbf{q}) < 0. \tag{3}$$

Since $\mathbf{q} \in \text{int}(\Delta_{[K]})$, we can consider a line segment through $\mathbf{q}$ parallel to $\mathbf{s}$, and one choice of $\mathbf{p}^*$ arbitrarily close to $\mathbf{q}$ on one side of the line will satisfy Equation (3). Plugging in Equation (2) and Equation (3) to Lemma 3.3 yields

$$\mathbb{E}_{y\sim\mathbf{p}^*}\left[S(\mathbf{p}, y)\right] - \mathbb{E}_{y\sim\mathbf{p}^*}\left[S(\mathbf{q}, y)\right] > \frac{\kappa}{2} > 0$$

as desired. □

Next we prove Theorem 3.5. The forward direction of Theorem 3.5 is exactly Theorem 3.2, so it suffices to show the contrapositive:

Proposition D.6. *Let*

$$S(\mathbf{p}, y) = G(\mathbf{p}) + \nabla G(\mathbf{p}) \cdot (\mathbf{1}_y - \mathbf{p})$$

be a scoring rule derived from a function G that is not convex. Let $\mathbf{s} = \mathbf{s}(\mathbf{p}, \mathbf{q})$ be a betting strategy that depends only on $\mathbf{p}$ and $\mathbf{q}$ and that is nowhere locally constant in $\mathbf{q}$. Then there exist values for $\mathbf{p}$, $\mathbf{q}$, and $\mathbf{p}^*$ such that $S(\mathbf{p}; \mathbf{p}^*) > S(\mathbf{q}; \mathbf{p}^*)$ but $\mathbf{s} \cdot (\mathbf{p}^* - \mathbf{q}) < 0$.

Proof. Since G is not convex, there must exist some $\mathbf{p}$ and $\mathbf{q}$ such that $\mathbf{q} \in \text{int}(\Delta_{[K]})$ and

$$D_G(\mathbf{q}, \mathbf{p}) = G(\mathbf{q}) - G(\mathbf{p}) - \nabla G(\mathbf{p}) \cdot (\mathbf{q} - \mathbf{p}) = -\kappa < 0. \tag{4}$$

Since $\mathbf{s}(\mathbf{p}, \mathbf{q})$ is not locally constant, we can perturb $\mathbf{q}$ by an infinitesimal amount so that $\mathbf{q} \in \text{int}(\Delta_{[K]})$ and Equation (4) still hold for some κ and also $\mathbf{s} \neq \mathbf{0}$ is a nontrivial betting strategy. Then we can construct $\mathbf{p}^*$ in the same way as the proof of Proposition D.5, noting that we did not use any special property of $\mathbf{s}$ other than the fact that $\mathbf{s} \neq \mathbf{0}$. □

E. Extensions of Theorem 3.2

E.1. Proper betting over multiple events

Theorem 3.2 is a single-event guarantee that requires knowledge of the unobservable ground truth $\mathbf{p}^*$ to verify the score-gap hypothesis. In real prediction markets we only see realized outcomes $y_1, \ldots, y_n$ across a sequence of n events with associated forecaster predictions $\mathbf{p}_1, \ldots, \mathbf{p}_n$ and market prices $\mathbf{q}_1, \ldots, \mathbf{q}_n$. Define the empirical forecaster and market scores

$$\hat{S}_F \;:=\; \frac{1}{n}\sum_{i=1}^{n} S(\mathbf{p}_i, y_i), \qquad \hat{S}_M \;:=\; \frac{1}{n}\sum_{i=1}^{n} S(\mathbf{q}_i, y_i).$$

On any individual event the forecaster's score may well be lower than the market's even when their predictions are on average more accurate, so $\hat{S}_F > \hat{S}_M$ does not directly invoke the per-event guarantee. Averaging Lemma 3.3 across events nevertheless yields an empirical version of Theorem 3.2.

Corollary E.1 (Empirical version of Theorem 3.2)**.** *If $\hat{S}_F > \hat{S}_M$ on a realized sequence of outcomes $y_1, \ldots, y_n$, then the sequential proper bet $\mathbf{s}_i := \mathbf{s}_G(\mathbf{p}_i, \mathbf{q}_i)$ has strictly positive realized return on that same sequence:*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{s}_i \cdot (\mathbf{1}_{y_i} - \mathbf{q}_i) \;=\; (\hat{S}_F - \hat{S}_M) \;+\; \frac{1}{n}\sum_{i=1}^{n} D_G(\mathbf{q}_i, \mathbf{p}_i) \;>\; 0.$$

Proof. Apply Lemma 3.3 at each event i:

$$\mathbf{s}_i \cdot (\mathbf{1}_{y_i} - \mathbf{q}_i) \;=\; [S(\mathbf{p}_i, y_i) - S(\mathbf{q}_i, y_i)] + D_G(\mathbf{q}_i, \mathbf{p}_i).$$

Average over $i \in [n]$ and use the definitions of $\hat{S}_F$, $\hat{S}_M$. The score-gap term is strictly positive by hypothesis and the Bregman sum is non-negative by convexity of G, so the realized return is strictly positive. □

Remark E.2 (Bet sizes should vary across events). A natural intuition is that a forecaster whose accuracy is measured uniformly across events should place uniform bet sizes. Corollary E.1 says otherwise: the correct sizes are proportional to $\nabla G(\mathbf{p}_i) - \nabla G(\mathbf{q}_i)$, with larger bets on events where the forecaster's prediction deviates more from the market.

Remark E.3 (Profit is bounded below by Bregman divergence). The realized return in Corollary E.1 is not only positive but bounded below by the average Bregman divergence $\frac{1}{n}\sum_i D_G(\mathbf{q}_i, \mathbf{p}_i)$. Forecasters whose predictions deviate substantially from the market consensus *and* are more accurate under S must therefore profit substantially. Conversely, positive realized return does not imply accuracy: the observed profit can be attributed entirely to the Bregman bonus, and we experimentally observe a few forecasting agents profiting despite below-market accuracy.

E.2. Proper betting under nonzero bid-ask spread

In real prediction markets, low liquidity and platform fees create a positive gap between the prices at which a contract can be bought and sold. We extend Theorem 3.2 to this setting via the following bid-ask convention. For each outcome $k \in [K]$, let q_k^+ denote the ask price at which a *YES* contract on outcome k (paying 1 if $y = k$, otherwise 0) can be bought, and q_k^- the ask price at which the corresponding *NO* contract (paying 1 if $y \neq k$, otherwise 0) can be bought. Absence of riskless arbitrage on outcome k requires $q_k^+ + q_k^- \geq 1$, with equality recovering the spread-free setting of Section 2; in practice the inequality is strict.

Bid-ask-aware proper bet. Treat each outcome k as an independent binary event $y_k := \mathbf{1}[y = k] \in \{0, 1\}$ and extend the scoring rule to vectors $\mathbf{p} \in [0, 1]^K$ coordinate-wise:

$$S(\mathbf{p}, y) \;:=\; \sum_{k=1}^{K} S(p_k, y_k).$$

The Bregman divergence and proper-bet construction extend coordinate-wise as well. Given a model prediction $\mathbf{p} \in [0, 1]^K$ and prices $(\mathbf{q}^+, \mathbf{q}^-)$, define the bid-ask-aware proper bet $\mathbf{s} \in \mathbb{R}^K$ by

$$s_k \;:=\; \begin{cases} \nabla G(p_k) - \nabla G(q_k^+) & \text{if } p_k > q_k^+ \quad \text{(buy YES on } k \text{ at } q_k^+), \\ \nabla G(p_k) - \nabla G(1 - q_k^-) & \text{if } p_k < 1 - q_k^- \quad \text{(buy NO on } k \text{ at } q_k^-), \\ 0 & \text{if } 1 - q_k^- \leq p_k \leq q_k^+. \end{cases}$$

The middle case is the spread-induced no-bet zone: the model's edge is too small to overcome the spread on outcome k (the interval is non-empty since $q_k^+ + q_k^- \geq 1$). Setting $\tilde{q}_k := q_k^+$, $1 - q_k^-$, or p_k in the three cases respectively, we have $s_k = \nabla G(p_k) - \nabla G(\tilde{q}_k)$ uniformly, and the realized profit is

$$\mathbf{s} \cdot (\mathbf{1}_y - \tilde{\mathbf{q}}) = \sum_{k=1}^{K} s_k(y_k - \tilde{q}_k).$$

Corollary E.4 (Profitability under bid-ask spread)**.** *Let*

$$\tilde{S}_M := S(\tilde{\mathbf{q}}, y) = \sum_{k=1}^{K} S(\tilde{q}_k, y_k)$$

be the bid-ask-aware market score. If

$$\mathbb{E}_{y\sim\mathbf{p}^*}[S(\mathbf{p},y)] > \mathbb{E}_{y\sim\mathbf{p}^*}[\tilde{S}_M],$$

then the bid-ask-aware proper bet has expected profit

$$\mathbb{E}_{y\sim\mathbf{p}^*}\left[\mathbf{s}\cdot(\mathbf{1}_y-\tilde{\mathbf{q}})\right] \;=\; \underbrace{\mathbb{E}_{y\sim\mathbf{p}^*}\left[S(\mathbf{p},y)-\tilde{S}_M\right]}_{\textit{score gap}} \;+\; \underbrace{\sum_{k=1}^{K} D_G(\tilde{q}_k,p_k)}_{\textit{Bregman bonus}} \;>\; 0. \tag{5}$$

Proof. For each coordinate k, $s_k = \nabla G(p_k) - \nabla G(\tilde{q}_k)$ is the binary proper bet with respect to the reference $\tilde{q}_k$. Apply the binary version of Lemma 3.3:

$$s_k(y_k-\tilde{q}_k) \;=\; (\nabla G(p_k)-\nabla G(\tilde{q}_k))\,(y_k-\tilde{q}_k) \;=\; [S(p_k,y_k)-S(\tilde{q}_k,y_k)] + D_G(\tilde{q}_k,p_k).$$

Summing over $k\in[K]$ and taking expectation under $y\sim\mathbf{p}^*$ yields Equation (5). The score-gap term is strictly positive by hypothesis and the Bregman sum is non-negative by convexity of G, so the expected profit is strictly positive. □

Cost of spread. The reference $\tilde{\mathbf{q}}$ depends on which side of the spread each coordinate lands on, so an accuracy edge against $\tilde{\mathbf{q}}$ is harder to attain than one against any spread-free price $\bar{\mathbf{q}}$ with $\bar{q}_k \in [1-q_k^-, q_k^+]$. Specifically, the score gap against $\tilde{\mathbf{q}}$ decomposes as

$$S(\mathbf{p},y) - S(\tilde{\mathbf{q}},y) \;=\; [S(\mathbf{p},y)-S(\bar{\mathbf{q}},y)] \;+\; \underbrace{[S(\bar{\mathbf{q}},y)-S(\tilde{\mathbf{q}},y)]}_{\text{cost of spread}\,\le\,0},$$

where the spread cost vanishes when $q_k^+ + q_k^- = 1$ (zero spread). The bid-ask-aware result thus interpolates between the spread-free guarantee of Theorem 3.2 and a strictly weaker condition that absorbs the per-trade spread cost.

E.3. Proper betting for long time horizon

For events that take days or weeks to resolve, both forecaster and market predictions evolve over time. At each time t, denote the forecaster's prediction and the market price by $\mathbf{p}^t$ and $\mathbf{q}^t$ respectively. We extend Theorem 3.2 to two natural multi-period strategies, each compounding the per-round proper bet but differing in what they hold and what accuracy edge they require.

Corollary E.5 (Fundamental-driven strategy)**.** *The* fundamental *strategy executes the proper bet $\mathbf{s}^t := \nabla G(\mathbf{p}^t) - \nabla G(\mathbf{q}^t)$ at each time t and holds all positions to resolution. If at every round the forecaster outperforms the market under S relative to the ground truth $\mathbf{p}^*$,*

$$S(\mathbf{p}^t;\mathbf{p}^*) > S(\mathbf{q}^t;\mathbf{p}^*),$$

then the cumulative expected profit at resolution decomposes as

$$\sum_{t=1}^{T}\mathbf{s}^t\cdot(\mathbf{p}^*-\mathbf{q}^t) \;=\; \sum_{t=1}^{T}\left[S(\mathbf{p}^t;\mathbf{p}^*)-S(\mathbf{q}^t;\mathbf{p}^*)\right] \;+\; \sum_{t=1}^{T} D_G(\mathbf{q}^t,\mathbf{p}^t) \;>\; 0. \tag{6}$$

Corollary E.6 (Momentum-driven strategy)**.** *The* momentum *strategy maintains the proper position $\mathbf{x}^t := \nabla G(\mathbf{p}^t) - \nabla G(\mathbf{q}^t)$ at each time t, rebalancing after every market update; the marginal trade at time t is $\mathbf{s}^t := \mathbf{x}^t - \mathbf{x}^{t-1}$ with $\mathbf{x}^0 := \mathbf{0}$. If at every round the forecaster outperforms the market under S* evaluated against the next-step market price $\mathbf{q}^{t+1}$*,*

$$S(\mathbf{p}^t;\mathbf{q}^{t+1}) > S(\mathbf{q}^t;\mathbf{q}^{t+1}),$$

then the cumulative mark-to-market return decomposes as

$$\sum_{t=1}^{T}\mathbf{x}^t\cdot(\mathbf{q}^{t+1}-\mathbf{q}^t) \;=\; \sum_{t=1}^{T}\left[S(\mathbf{p}^t;\mathbf{q}^{t+1})-S(\mathbf{q}^t;\mathbf{q}^{t+1})\right] \;+\; \sum_{t=1}^{T} D_G(\mathbf{q}^t,\mathbf{p}^t) \;>\; 0. \tag{7}$$

Equivalently in trade form,

$$\sum_{t=1}^{T}\mathbf{x}^t\cdot(\mathbf{q}^{t+1}-\mathbf{q}^t) = \sum_{t=1}^{T}\mathbf{s}^t\cdot(\mathbf{q}^{T+1}-\mathbf{q}^t),$$

the profit of executing trades $\mathbf{s}^t$ at prices $\mathbf{q}^t$ and settling at $\mathbf{q}^{T+1}$.

Table 4. Calibration and profitability metrics by model and persona group. N is the number of market-level predictions per model. ECE is expected calibration error. Quad. Score represents the model's score under the quadratic scoring rule. $\pm$ values are the bootstraped standard errors over 1,000 resamples. Shaded rows report persona-group averages.

Model	N markets	ECE	Quad. Score	ROI
GPT-5.2 (High)	11,789	0.021 $\pm$0.005	0.916 $\pm$0.006	-7.8% $\pm$3.6 (Log)
GPT-5.2 (Base)	11,789	0.013 $\pm$0.003	0.916 $\pm$0.005	-4.4% $\pm$3.9 (Log)
Grok 4.1 Fast	11,407	0.027 $\pm$0.004	0.913 $\pm$0.005	-9.1% $\pm$4.5 (Log)
Claude Sonnet 4.5	11,785	0.023 $\pm$0.003	0.911 $\pm$0.005	-8.1% $\pm$3.5 (Log)
Kimi K2 Thinking	11,811	0.024 $\pm$0.004	0.912 $\pm$0.005	-8.7% $\pm$3.8 (Log)
DeepSeek V3.2	11,071	0.031 $\pm$0.005	0.899 $\pm$0.006	-12.8% $\pm$3.4 (Log)
Minimax M2	11,688	0.028 $\pm$0.005	0.900 $\pm$0.006	-9.4% $\pm$3.8 (Log)
Brittle (avg)	**11,620**	**0.024**	**0.910**	$\mathbf{-8.6\%}$ **(Log)**
LLaMA 4 Maverick	29,613	0.069 $\pm$0.004	0.871 $\pm$0.005	-4.9% $\pm$2.2 (Log)
Qwen 3 235B	29,923	0.078 $\pm$0.004	0.868 $\pm$0.004	-11.2% $\pm$1.9 (Log)
DeepSeek R1	29,759	0.100 $\pm$0.005	0.848 $\pm$0.005	-5.4% $\pm$1.7 (Log)
Dispersed (avg)	**29,765**	**0.082**	**0.862**	$\mathbf{-7.2\%}$ **(Log)**
Gemini 3	9,721	0.017 $\pm$0.003	0.936 $\pm$0.004	$+14.2\%$ $\pm$8.8 (Brier)
Claude Opus 4.6	3,172	0.011 $\pm$0.004	0.949 $\pm$0.007	$+17.5\%$ $\pm$7.2 (Brier)
Conservative (avg)	**6,446**	**0.014**	**0.942**	$\mathbf{+15.8\%}$ **(Brier)**

The two corollaries differ in what plays the role of the "ground truth" in the proper-bet decomposition: Corollary E.5 uses the unobservable $\mathbf{p}^*$ that resolves the event, so the trader needs an accuracy edge against the truth and is exposed across the entire horizon; Corollary E.6 uses the next-period market price $\mathbf{q}^{t+1}$, so the trader only needs to predict the market's next move and is exposed only one period at a time.

Proof. For Corollary E.5, apply Lemma 3.3 at each round with realized outcome y:

$$\mathbf{s}^t \cdot (\mathbf{1}_y - \mathbf{q}^t) \;=\; \big[S(\mathbf{p}^t, y) - S(\mathbf{q}^t, y)\big] + D_G(\mathbf{q}^t, \mathbf{p}^t).$$

Summing over t and taking expectation under $y \sim \mathbf{p}^*$ gives Equation (6). The score-gap sum is strictly positive by hypothesis and the Bregman sum is non-negative, so the total is positive.

For Corollary E.6, apply Lemma 3.3 (in expectation form) at each round with substitutions $\mathbf{p}^* \to \mathbf{q}^{t+1}$, $\mathbf{p} \to \mathbf{p}^t$, $\mathbf{q} \to \mathbf{q}^t$:

$$\mathbf{x}^t \cdot (\mathbf{q}^{t+1} - \mathbf{q}^t) \;=\; \big[S(\mathbf{p}^t; \mathbf{q}^{t+1}) - S(\mathbf{q}^t; \mathbf{q}^{t+1})\big] + D_G(\mathbf{q}^t, \mathbf{p}^t).$$

Summing gives Equation (7). The trade-form equivalence follows from substituting $\mathbf{x}^t = \sum_{r=1}^{t} \mathbf{s}^r$ and exchanging the order of summation:

$$\sum_{t=1}^{T} \mathbf{x}^t \cdot (\mathbf{q}^{t+1} - \mathbf{q}^t) \;=\; \sum_{r=1}^{T} \mathbf{s}^r \cdot \sum_{t=r}^{T} (\mathbf{q}^{t+1} - \mathbf{q}^t) \;=\; \sum_{r=1}^{T} \mathbf{s}^r \cdot (\mathbf{q}^{T+1} - \mathbf{q}^r),$$

where the inner sum telescopes by

$$\sum_{t=r}^{T} (\mathbf{q}^{t+1} - \mathbf{q}^t) = \mathbf{q}^{T+1} - \mathbf{q}^r.$$

Strict positivity follows similarly to the fundamental-driven strategy. □

F. Additional experiments

F.1. Model details

Table 4 reports the LLMs evaluated across three metrics – expected calibration error (ECE), the score under the quadratic scoring rule, and ROI under each model's best proper betting strategy. GPT-5.2 (Base/High) correspond to different reasoning levels of the same underlying model. The number of markets varies across forecasters because models were

released at different times, resulting in differing amounts of collected data through the Prophet Arena pipeline; to ensure comparability, we standardize evaluations in Section 4.

Interestingly, while the Brittle group has lower ECE and the higher aggregate Brier accuracy as compared to the Dispersed group, it is also more unprofitable. As such, aggregate accuracy and calibration can therefore coexist with systematic capital loss, due to the importance of the Bregman divergence term.

F.2. Implementation details of betting strategies

Setup. We consider a collection of n markets with forecast–price pairs $\{p_i, q_i\}_{i\in[n]}$. A betting strategy is defined by a set of budget allocation weights $\{w_i\}_{i\in[n]}$, which encode both direction and size: if $w_i > 0$, we buy YES on market i with allocation w_i; if $w_i < 0$, we buy NO (equivalently, sell YES) with allocation $-w_i$. Without loss of generality, we restrict $w_i \in [-1, 1]$ for all $i \in [n]$ and normalize total exposure such that $\sum_{i\in[n]} |w_i| = 1$.

Proper Betting Strategies. Proper betting strategies can be constructed as follows:

1. **Brier:** Allocates linearly in the margin: $w_i \propto p_i - q_i$.
2. **Logarithmic:** Scales allocations according to: $w_i \propto \log p_i - \log q_i$.
3. **Spherical:** Allocates based on the difference between the L_2-normalized forecast and market price vectors: $w_i \propto \frac{p_i}{\|p\|} - \frac{q_i}{\|q\|}$.

Baseline Betting Strategies. We also construct a few betting strategies from well-motivated heuristics as baselines:

1. **Max-Margin:** Strategies that bet the margin.
   - *Market-Level:* $w_i \propto \text{sign}(p_i - q_i), \forall i \in [n]$. This strategy places a unit bet on every market, with direction determined by whether $p_i > q_i$.
   - *Grouped:* Let $\{\mathcal{I}_k\}_{k=1}^m$ denote a partition of markets into groups corresponding to the same underlying question. For each group k, define
   $$i^\star(k) = \arg\max_{i\in\mathcal{I}_k} |p_i - q_i|.$$
   This strategy allocates the budget uniformly across groups, placing a single bet per group on the market with the largest margin:
   $$w_{i^\star(k)} = \frac{\text{sign}(p_{i^\star(k)} - q_{i^\star(k)})}{m}, \quad w_i = 0, \quad \forall i \notin \{i^\star(k)\}_{k=1}^m.$$
2. **Inverse-Margin:** $w_i \propto \frac{\text{sign}(p_i - q_i)}{|p_i - q_i|}, \forall i \in [n]$. This heuristic places more weight on markets with smaller margins, under the hypothesis that forecasts are more reliable when deviations from the market are modest.
3. **Kelly-Alike:** $w_i \propto \frac{p_i - q_i}{1 - q_i}, \forall i \in [n]$. This heuristic strategy mimics the ratio in the Kelly criterion.
4. **Kelly Criterion (Kelly, 1956):** Unlike the previous strategies, which allocate a fixed budget B across markets via weights w_i with the goal of maximizing expected profit, Kelly sizing determines the fraction of wealth to wager on each individual market sequentially to maximize the expected log-growth rate. For a binary market with forecast p_i and market price q_i, a unit stake on YES pays $1/q_i$, giving net odds $b_i = (1 - q_i)/q_i$. Maximizing
   $$\mathbb{E}[\log W_{i+1}] = p_i \log(1 + f_i b_i) + (1 - p_i) \log(1 - f_i)$$
   over the wealth fraction f_i yields the closed form $f_i = \frac{|p_i - q_i|}{1 - q_i}$. If f_i exceeds 1, we allow for the use of leverage to purchase additional shares.

Table 5 reports the full comparison of evaluated models across all baseline betting strategies and the proposed Proper (Brier) strategy.

Table 5. ROI (%) of baseline betting strategies and the proposed Proper (Brier) strategy. ΔS is the proper (Brier) score gap: negative values indicate worse performance than the market. Bold marks the best ROI per row.

		ROI (%)					
Model	ΔS	**Proper**	**Max (Mkt)**	**Max (Grp)**	**Inv-Margin**	**Kelly-Alike**	**Kelly**
Claude Opus 4.6	$+0.0016$	**+22.1**	+5.0	-14.0	-3.5	+10.9	-99.9
Gemini 3	$+0.0008$	**+8.1**	+1.3	-0.1	+0.7	+1.8	-42.7
GPT-5.2 (Base)	-0.0347	**+4.5**	+2.4	-14.9	+1.9	-0.1	-99.9
Claude Sonnet 4.5	-0.0426	-3.5	+1.2	-19.8	**+1.4**	+0.7	-99.9
LLaMA 4 Maverick	-0.0450	-13.7	**-4.1**	-15.5	-4.0	-10.0	-99.9
GPT-5.2 (High)	-0.0460	-2.4	-1.6	-13.3	-1.9	**-1.0**	-99.9
Grok 4.1 Fast	-0.0462	-11.6	-7.8	-26.4	-7.0	**-6.3**	-99.9
DeepSeek V3.2	-0.0466	-12.2	-7.3	-30.2	**-5.1**	-5.8	-99.9
Kimi K2 Thinking	-0.0492	-9.5	-3.1	-24.3	**+1.5**	-2.5	-99.9
Minimax M2	-0.0537	**+1.3**	-4.0	-22.8	-6.5	-3.9	-99.9
DeepSeek R1	-0.0636	-20.3	-7.6	-25.4	**+0.7**	-9.6	-99.9
Qwen 3 235B	-0.0838	-13.3	-5.3	-14.8	**+12.2**	-9.3	-99.9

F.3. Empirical decomposition

Table 6 reports the full decomposition across all models and all evaluated proper betting strategies on the standardized 2,418 market dataset. Divergence varies substantially across both models and rules, and often offsets large negative score gaps, meaning models less accurate can still yield positive ROI (e.g., Claude Sonnet 4.5 under Log). Rules such as Log tend to amplify divergence, leading to higher upside but also greater dispersion in outcomes, while Brier and Spherical are comparatively more stable. This reinforces that the mapping from forecasts to returns is highly rule-dependent, and that profitability hinges as much on how aggressively a strategy exploits disagreement as on the underlying predictive accuracy.

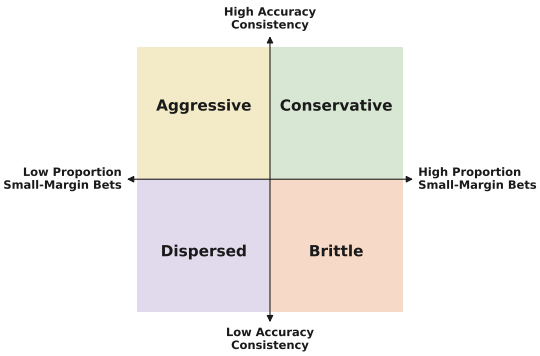
Table 6. ROI (%) decomposition of proper betting strategies. ΔS is the aggregate score difference between the forecast and the market across all bets placed, and D is total Bregman divergence. All quantities are summed across bets, normalized by total cost staked under each rule and multiplied by 100 so $\Delta S + D = \text{ROI}$. Values are rounded to 1 decimal place.

	Brier			**Log**			**Spherical**		
Model	ΔS	D	**ROI**	ΔS	D	**ROI**	ΔS	D	**ROI**
Claude Opus 4.6	$+3.2$	$+17.8$	$+21.0$	$+4.5$	$+16.9$	$+21.4$	-1.3	$+10.1$	$+8.8$
Gemini 3	$+4.0$	$+5.0$	$+9.0$	$+3.9$	$+5.0$	$+8.9$	-2.4	$+2.9$	$+0.4$
GPT-5.2 (Base)	-108.4	$+112.7$	$+4.3$	-5.8	$+15.9$	$+10.1$	-52.7	$+46.7$	-6.0
Claude Sonnet 4.5	-111.2	$+110.2$	-0.9	-10.0	$+17.6$	$+7.5$	-63.4	$+59.9$	-3.6
DeepSeek V3.2	-116.4	$+107.3$	-9.1	-18.0	$+18.7$	$+0.7$	-64.3	$+53.8$	-10.5
GPT-5.2 (High)	-159.2	$+159.5$	$+0.3$	-13.1	$+21.9$	$+8.8$	-85.9	$+81.2$	-4.6
Grok 4.1 Fast	-137.4	$+128.3$	-9.1	-20.3	$+23.8$	$+3.5$	-89.3	$+74.3$	-15.0
LLaMA 4 Maverick	-123.7	$+110.6$	-13.1	-21.0	$+20.2$	-0.8	-53.3	$+38.5$	-14.8
Kimi K2 Thinking	-123.7	$+115.4$	-8.3	-21.5	$+20.4$	-1.2	-71.4	$+65.0$	-6.4
Minimax M2	-138.7	$+137.6$	-1.1	-10.9	$+19.8$	$+8.9$	-87.6	$+70.7$	-17.0
DeepSeek R1	-138.9	$+117.6$	-21.3	-35.5	$+22.9$	-12.6	-74.4	$+55.7$	-18.7
Qwen 3 235B	-148.1	$+138.1$	-9.9	-47.1	$+44.6$	-2.4	-89.4	$+75.5$	-13.8

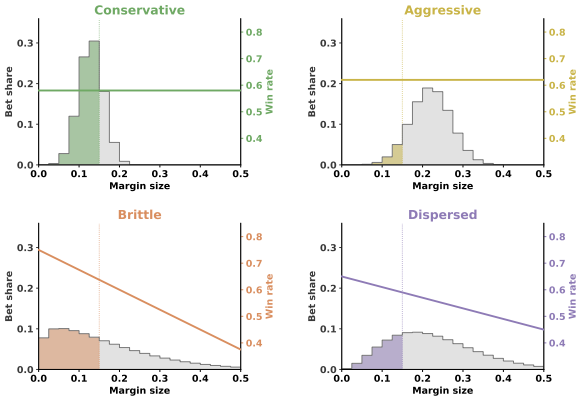
F.4. Synthetic persona generation

We define four personas spanning representative combinations of two dimensions—small-margin share and accuracy-consistency—as shown in Figure 3. **Conservative** models concentrate most forecasts in the small-margin region with flat accuracy across margins. **Aggressive** models place fewer small-margin bets, with accuracy still flat. **Dispersed** models spread bets broadly across margin sizes and show gradual decline in win-rate as margin increases. **Brittle** models have a significant proportion of small-margin forecasts but exhibit a sharply declining win-rate as margin increases.

We evaluate each persona under two regimes: Regime A, where the model outperforms the market and all strategies yield positive ROI, and Regime B, where it underperforms and all strategies incur losses. A symmetric ± 0.05 quadratic score gap



(a) Taxonomy of LLM forecasting personas by proportion of forecasts at small margins and accuracy consistency across margins.



(b) Each panel shows a persona's margin distribution (grey, left axis) and win-rate at different margins (color, right axis).

Figure 3. Schematics for synthetic forecasting personas and their prediction profiles.

Table 7. Simulated persona ROI (%) under proper betting strategies. S-M Share denotes the share of forecasts with small margin: $|p - q| \leq 0.15$. Consistency is the OLS slope of per-bet directional win rate against margin size.

			Regime A (Outperforms)			*Regime B (Underperforms)*		
Persona	**S-M Share**	**Consistency**	**Brier**	**Log**	**Sph.**	**Brier**	**Log**	**Sph.**
Conservative	0.85	0.0	**+48.8**	+46.2	+44.8	-76.8	**-71.5**	-82.7
Aggressive	0.58	0.0	**+37.7**	+37.4	+31.3	-38.8	**-30.1**	-46.9
Dispersed	0.68	−0.2	**+44.3**	+42.5	+38.6	-30.4	**-20.7**	-38.1
Brittle	0.77	−0.6	**+45.4**	+43.8	+41.0	-60.2	**-52.2**	-67.9

(market quadratic score = 0.839) isolates the effect of persona and strategy while holding average accuracy constant. As shown in Table 7, Brier is optimal in Regime A for all personas, while Log is optimal in Regime B. In Regime A, the positive Brier gap implies that expected returns are positive across margin bins, so linear scaling in $|p - q|$ is optimal. In Regime B, losses are dominated by high-margin errors, making Log's sublinear weighting preferable as it attenuates exposure to large, costly mistakes.

For each persona, we generate a synthetic forecast by perturbing the market price by a persona-specific margin:

$$p_i = q_i \pm m_i \quad \text{(clipped to } [\varepsilon,\, 1-\varepsilon]),$$

where $m_i \geq 0$ is the margin magnitude drawn from a persona-specific distribution, and the sign is chosen toward the realized outcome with probability a_i. A single global offset on a_i is calibrated per persona so that all personas attain the same Brier score within each regime.

Synthetic persona construction. Conservative uses $m \sim \text{Beta}(\text{mode} = 0.12, \text{conc} = 60)$ with uniform accuracy $a(m) = a_0$; Aggressive uses $m \sim \text{Beta}(0.20, 25)$ with uniform $a(m) = a_0$; Dispersed uses $m \sim \text{Beta}(0.10, 6)$ with linearly declining accuracy $a(m) = \text{clip}(a_0 - 0.20\, m, 0.01, 0.99)$, where $m = |p - q|$ denotes the margin.

In all cases, a_0 is calibrated via Brent's method on mean Brier so that the quadratic score matches the market within ± 0.05 in Regime A/B. The calibrated values are $a_0 = 0.999/0.073$ for Conservative, $0.877/0.338$ for Aggressive, and $0.971/0.473$ for Dispersed, where the first number corresponds to Regime A (model beats market) and the second to Regime B (model loses to market). Empirically, these yield $\Pr[m \leq 0.15] = 0.85,\ 0.58,\ 0.68$ and OLS win-rate slopes of $+0.00,\ -0.01,\ -0.20$, respectively.

For Brittle, matching the same ΔS (Brier) cannot be achieved with a single linear win-rate schedule: slopes steep enough to capture the empirical tail decline reduce mean accuracy below the Regime A target, while flatter slopes eliminate the pattern. We therefore use a two-piece schedule

$$a(m) \;=\; \text{clip}\big(a_0 - s \cdot \max(0,\, m - \tau),\ 0.01,\ 0.99\big),$$

with $\tau = 0.20$, $s = 1.0$, and $m \sim \text{Beta}(0.12, 20)$, calibrated to $a_0 = 0.999/0.223$ for Regimes A/B. In Figure 3b, we visualize this fitting a linear regression (OLS) for comparability with the other personas.

F.5. Proper betting rule capital allocation surfaces

The three proper scoring rules evaluated – Brier, logarithmic, and spherical – differ in *how much capital* they allocate to each bet (i.e., the number of shares purchased). Figure 4 visualizes these weight functions across the full (q, m) plane, where m is the signed margin.

Under the Brier rule, the weight is simply $w = |p - q|$, so capital scales linearly with divergence and is independent of q, producing vertical contour lines. As a result, Brier is the most aggressive strategy at large divergences, concentrating capital where model overconfidence is most pronounced. In contrast, the logarithmic rule, $w = |\log p - \log q|$, grows sublinearly in divergence and depends on the price level. The inward curvature of its contours at large margins reflects this diminishing sensitivity, so weight increases more slowly as divergence grows. These features make the log rule more conservative. The spherical rule lies between these extremes: its weight is nonlinear in both divergence and price, allocating more than log at moderate divergences—amplifying smaller, more reliable signals—but less than Brier at extreme divergences. It also exhibits mild price dependence, with slightly higher weights near $q = 0.5$, where normalization effects are weakest.

Generally, Brier favors personas whose accuracy holds across margin sizes (e.g., Conservative), but in the regime in which all personas can outperform the market in terms of accuracy, the performance gap is sufficiently large that even less stable personas benefit from linear scaling. For smaller (though still positive) Brier gaps, sharp deterioration in win rate at higher margins can cause linear scaling to over-weight large-margin forecasts, leading to outsized losses. In such cases, Log may be preferable due to its stronger penalization of extreme probabilities.

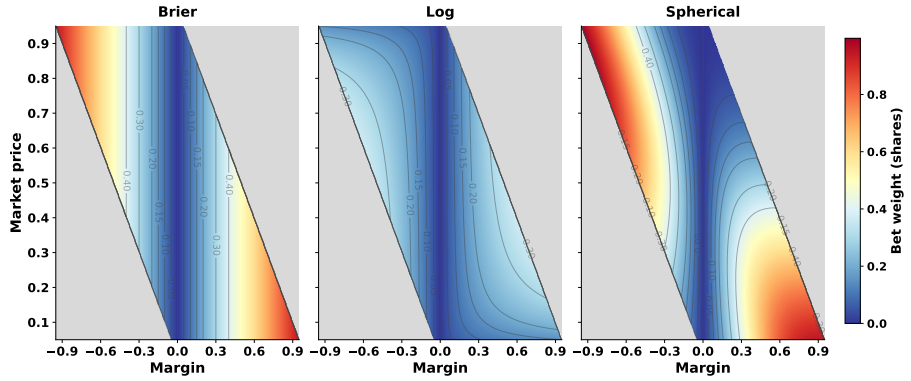


Figure 4. Capital allocation (bet weight in shares) as a function of signed margin and market price (q, y-axis) for the three proper betting strategies. Warmer colors indicate larger bets. Grey regions are infeasible ($p \notin (0, 1)$). Contour lines mark constant weight levels.

F.6. Mapping empirical forecasters to personas

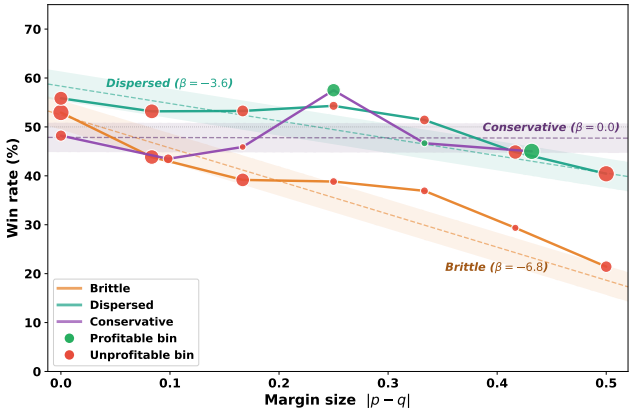
As shown in Table 8, along the margin dimension, models separate into three clear groups: seven place 75%–85% of bets below 0.15 (Brittle), three distribute more mass across margins (Dispersed, $< 70\%$ below 0.15), and two concentrate over 85% at small margins (Conservative). Second, Figure 5b confirms the persona assignments along the accuracy consistency dimension.The models exhibiting the Brittle persona exhibit a steep win-rate slope ($\beta = -6.8$), with win-rate falling from over 50% to below 30% by margin 0.50. Dispersed models decline more gradually ($\beta = -3.6$), while Conservative models are essentially flat ($\beta = 0$).

F.7. Live deployment experiment

We deploy a portfolio-management version of the forecasting framework in Yang et al. (2025). The agent runs on a fixed two-hour cadence; on each cycle it (i) refreshes prices and lifecycle state for every market it currently tracks, (ii) discovers a batch of new markets, (iii) queries the LLM forecaster on each eligible market through a standardized prediction context, and (iv) hands the resulting probabilities to an executor that places limit orders on Kalshi. Summary statistics and cumulative ROI from the live deployment are shown in Figure 6.

Model	S-M Share	ΔS	ROI	
GPT-5.2 (B)	0.82	-0.044	-4.4	(L)
Grok 4.1	0.80	-0.053	-9.1	(L)
Brittle	**0.80**	$\mathbf{-0.054}$	$\mathbf{-8.6}$	**(L)**
LLaMA 4	0.67	-0.115	-4.9	(L)
DeepSeek R1	0.59	-0.140	-5.4	(L)
Dispersed	**0.64**	**-0.120**	**-7.2**	**(L)**
Gemini 3	0.94	+0.002	+14.2	(B)
Claude Opus	0.89	+0.000*	+17.5	(B)
Conservative	**0.91**	**+0.001**	**+15.8**	**(B)**

(a) Persona classification across models. ROI reported is under the best proper scoring rule: Log (L) or Brier (B). * denotes a positive ΔS rounded to $+0.000$.



(b) Binned per-bet win rate vs. margin $|p - q|$. Each dot has area proportional to profit for their persona. Dashed lines are OLS fits with slope β.

Figure 5. Mapping real models to personas.

A Kalshi market is pulled for analysis on a given cycle if it is open and its `close time` lies between 2 and 14 days (we impose a 14-day maximum horizon to facilitate capital turnover and avoid positions being tied up in long-dated markets). To limit exposure to trading fees and LLM-driven variance, a market is skipped on a given cycle if the market price has not moved by at least $10¢$ since the last fill; markets we have never traded on are always re-evaluated.

For each eligible market we issue a single forecast call. The model is given the market's title, resolution rules, and contemporaneous market data and is asked to return a JSON object containing a rationale and the probability that the contract resolves YES. We use the same prompting template and prediction context as Yang et al. (2025). The deployed forecaster is Gemini 3 Pro with high-effort thinking and Google Search grounding enabled.

F.7.1. Prediction context and methodology

Let p denote the model's probability of YES. We do not trade when p lies within the bid–ask band (a consequence of the occasional non-zero bid-ask spread in real-world prediction markets), as this implies no actionable margin.

Whenever a market is traded, we rebalance the position toward the new target τ on each cycle, as is detailed in Appendix E.3.

All orders are placed as limit orders at the prevailing ask, so execution is not guaranteed: on thin Kalshi books, a portion of the requested size often remains unfilled at the end of the cycle. If they do not fill, they expire and are cancelled at the next cycle when updated prices and forecasts produce a new target. The next-cycle delta is then computed based on the positions that actually filled.

Figure 6a presents the full detailed performance and trading statistics for the strategy over the evaluation period. We also provide access to the trading history **here**, which logs performance of the trading agent over the trading period.

F.7.2. Eligibility criteria

Eligibility is determined ex-ante; we exclude three categories of markets and halt trading three hours prior to event resolution, the latter reflecting the reduced informational advantage of LLMs close to resolution time (Yang et al., 2025). First, we exclude the Mentions category, where contracts resolve based on unstructured public statements (e.g., "will X mention Y"), as we find that LLMs perform substantially worse on these questions in tests over publicly available data released by Yang et al. (2025). Second, we exclude events for which a Kalshi market's listed `close_time` is more than one hour after the true resolution time of the event[1], creating a mismatch between the LLM's belief about market closure and the actual outcome realization.

[1]Kalshi's `close_time` is the trading deadline and does not always coincide with when the underlying event resolves. While these align for most events, some have a `close_time` set hours or days after the outcome is publicly known. This discrepancy is observable ex ante via a field in the API deemed the `expected_expiration_time`, which indicates the event's resolution time.

Table 8. Language model margin concentration. Entries report the share of bets (%) in each margin bin. ΔS denotes the aggregate score difference relative to the market. Greyed columns correspond to high-margin bins with sparse data and are excluded from the win-rate analysis. Shares are rounded to the nearest integer; * indicates a value that is positive but rounds to +0.000.

Model	[0,.05)	[.05,.15)	[.15,.25)	[.25,.50)	[.50,.70)	[.70,.90)	[.90,1.0)	ΔS
GPT-5.2 (High)	61	23	7	5	2	1	1	−0.038
GPT-5.2 (Base)	59	23	7	6	2	2	1	−0.044
Grok 4.1 Fast	58	22	7	7	3	2	1	−0.053
Claude Sonnet 4.5	54	25	8	7	3	2	1	−0.055
Kimi K2 Thinking	57	24	6	7	2	2	1	−0.053
DeepSeek V3.2	50	25	8	9	4	3	1	−0.069
Minimax M2	52	25	8	8	3	3	2	−0.064
Brittle (avg)	**56**	**24**	**7**	**7**	**3**	**2**	**1**	**−0.054**
LLaMA 4 Maverick	46	21	9	10	5	4	5	−0.115
Qwen 3 235B	45	21	9	12	5	4	3	−0.106
DeepSeek R1	37	22	10	14	8	6	5	−0.140
Dispersed (avg)	**43**	**21**	**9**	**12**	**6**	**5**	**4**	**−0.120**
Gemini 3	81	13	2	1	0	0	2	+0.002
Claude Opus 4.6	69	20	5	3	1	1	0	+0.000*
Conservative (avg)	**75**	**16**	**4**	**2**	**0**	**0**	**1**	**+0.001**

Starting capital	$200.00
Trading days	26
Forecasts issued	1,605
Trade signals placed (filled)	396 (236)
Markets traded	129
Ending capital	$360.67
ROI	+80.33%
Sharpe ratio ($\sqrt{365}$)	3.35
Shares transacted	2,657
Exchange fees paid	$26.31
Win rate	50.9%

(a) Live trading statistics for the Gemini 3 agent on Kalshi.

(b) ROI over trading days of the forecaster's live Kalshi trades over 26 trading days.

Figure 6. Live deployment results for the Gemini 3 trading agent on Kalshi prediction markets.

In addition to eligibility criteria, we also drop a market from consideration due to an underspecified resolution rule. We have contacted Kalshi about contracts of this kind, and we view such situations as an artifact of the novelty of prediction markets that will phase out as venues mature and resolution-rule conventions standardize; this is not a limitation of our betting strategy but of the underlying contract specification. The event was titled *"What will be the top AI model this month?"* and the published resolution rule reads in full: *"If* `claude-opus-4-6-thinking` *is the top-ranked AI model on Apr 30, 2026 at 10:00 AM ET, then the market resolves to Yes."* The rule does not specify which leaderboard, benchmark, or evaluation suite determines "top-ranked," nor does it name the operator that produces such a ranking. Furthermore, plausible benchmarks — LMArena, Artificial Analysis, the OpenLLM Leaderboard, MMLU, etc. — can even disagree on a given day, so the market's outcome is determined by the market operator's discretionary choice of ranking source rather than by any externally verifiable fact. We therefore treat this market as failing the resolution-clarity precondition for forecasting and exclude it from the agent's [2].

F.7.3. Case studies

We present two representative case studies drawn from live trading runs to illustrate how the proposed strategy operates in practice. These examples highlight both the decision-making process and the resulting performance.

[2] Subsequent monthly versions of this contract now explicitly specify the LM Arena leaderboard as the resolution source.

Case 1: Weekly Executive Order count (`KXEOWEEK-26APR25-0`). This contract asked whether the President of the United States would sign more than 0 Executive Orders between April 19–25, 2026. By the time the agent first evaluated the market on April 21, public news cycles were dominated by a recent executive order (EO) accelerating medical treatments for serious mental illness, and Kalshi traded YES at 0.77. Gemini 3 priced YES essentially in line with the market, but on April 22 issued a `BUY_NO` signal with the rationale:

> "That specific Executive Order (accelerating medical treatments for serious mental illness) was signed on Saturday, April 18, which falls just outside the April 19–25 window. No new Executive Orders have been signed so far between April 19 and April 22, leaving only three days in the market period. The current prediction market probability of 66% is likely inflated by traders conflating the April 18 signing with this week's news cycle, so the true probability of an additional EO being signed in the remaining window is materially lower."

Eight prior forecasts in which model and market had agreed on $p_{\text{yes}} \approx 0.77$, the model's p_{yes} collapsed to 0.31 in a single cycle while Kalshi's `yes_ask` still sat at 0.66, opening a 35-cent edge on the NO side. The strategy converted that edge into 35 NO shares at $\$0.37$ per share.

As the market then drifted in the model's direction over the following five days—first to `no_ask` $= 0.68$ on April 24, then to `no_ask` $= 0.99$ by April 27—the position was gradually reduced as the edge decayed. Rather than holding a fixed exposure, the rebalancer continuously reassessed whether additional upside remained. It sold part of the position as soon as the mispricing meaningfully narrowed, taking 7 shares at $\$0.67$ ($+\$1.87$), and later unwound most of the remaining exposure near certainty at $\$0.98$ ($+\$4.35$), where further gains were negligible.

The remaining 23 shares, accumulated at an average cost of $\$0.43$, were held into resolution: the President signed zero Executive Orders that week, the contract resolved NO, and the residual position paid out at $\$1.00$ per share for an additional $+\$13.01$. The per-market net P&L of $+\$19.23$ on a peak cost basis of $\$15.21$ ($+126\%$) reflects gains accrued both through early liquidation at improving prices and through the final resolution payoff.

Case 2: U.S. Strategic Petroleum Reserve level (`KXSPRLVL-26APR01-T415`). This contract asked whether the U.S. Strategic Petroleum Reserve (SPR) level on April 1, 2026, would be above 415 million barrels. The reserve had been stable at 415.44 million barrels for several weeks, but on March 11 the Department of Energy announced a 172-million-barrel emergency drawdown, which led the market to price in a high probability of the level falling below 415.

By March 28, Gemini 3 was already assigning low probability to YES (around 0.13), and the agent initially accumulated NO positions. However, on March 30 at 14:06 UTC, the agent reversed sharply after identifying a timing mismatch in how the data is measured. It recognized that the EIA Weekly Petroleum Status Report records inventory as of 7:00 a.m. on Fridays, while the early phase of the announced drawdown had not yet been formally recorded in contract awards or physical accounting at that cutoff:

> "The Department of Energy did not award the initial contracts for the first 45.2 million barrels until Friday, March 27. The EIA's Weekly Petroleum Status Report measures inventory strictly as of 7:00 a.m. on Fridays, meaning any physical shipments that commenced later that day will not be captured in the data for the week ending March 27. Consequently, the SPR level—which has held perfectly steady at 415.44 million barrels for several weeks—will almost certainly print above 415 in the upcoming release."

In other words, although drawdown activity had been announced, it had not yet entered the official measurement window used for settlement. This meant that the upcoming report would still reflect the pre-drawdown level.

Acting on this insight, the agent closed its NO position and bought 81 YES shares at $\$0.12$. The following day, the EIA release confirmed the SPR level remained at 415.44 MMbbl, and the contract repriced to $\$0.87$, allowing the agent to sell the full position for a $+\$60.75$ profit on that leg alone (a $+625\%$ return).

After accounting for earlier position adjustments and small late-cycle bets, the net realized P&L was $+\$45.65$ on a $\$10.92$ cost basis. The key driver of performance was not disagreement about the drawdown itself, but the agent's correct identification of the measurement cutoff: it exploited the fact that announced changes had not yet entered the reporting window used for settlement.

F.8. Profitability at extreme margins

While most large margin bets are wrong (Table 9 illustrates a win rate $\ll 50\%$ for each high-margin bin for each model), the ROI can remain positive because the payoff on the rare correct bets scales sharply with margin. In these extreme bins, favorable prices mean that a small number of correct, high-conviction bets can more than offset the many losses, yielding positive returns despite low directional accuracy. This effect is most pronounced in the $[0.70, 0.9)$ bin for Conservative personas, where even a handful of correct predictions drives substantial gains. We highlight two such cases below from Claude Opus 4.6 and Gemini 3.

Table 9. Per-bin profitability at high margins ($|p - q| \geq 0.50$). For each persona and each margin bin, we report the (%) share of total bets, directional win rate, and ROI under the corresponding scoring rule (Log for Brittle / Dispersed, Brier for Conservative).

Persona	Bin	Share (%)	Win %	ROI (%)
Brittle	[.50, .70)	3	16.2	−17.3
	[.70, .90)	2	7.0	−57.4
	[.90, 1.0)	1	15.6	+191.6
Dispersed	[.50, .70)	6	34.3	+0.3
	[.70, .90)	5	14.3	−25.5
	[.90, 1.0)	4	2.7	−57.9
Conservative	[.50, .70)	0	17.6	+25.8
	[.70, .90)	0	10.5	+197.8
	[.90, 1.0)	1	2.9	+76.3

Table 10. Per-market Brier-strategy economics for two extreme-margin case studies. p_{yes} is the model's predicted probability for YES; q_{yes} is the YES-side ask price; y is the realized outcome. Weight is the margin $|p - q_{\text{yes}}|$.

Market (Kalshi ID)	Predicted	Resolved	p_{yes}	q_{yes}	y	Weight	ROI (%)
Case 1: Kirk Cousins's next team – "Stays with Atlanta or Retires" (Claude Opus 4.6)							
Stays with Atlanta or retires	2026-03-11	2026-03-25	0.03	0.78	NO	0.75	+21%
Case 2: Rain in Dallas, January 2026 – "Above 2 inches" (Gemini 3)							
Above 2 inches	2026-01-25	2026-02-01	0.22	0.98	NO	0.76	+1900%

Case 1: Kirk Cousins's next team (`KXNEXTTEAMNFL-26KCOUSINS`). This 32-way market asked which NFL team Kirk Cousins would sign with after his March 11, 2026 release by the Atlanta Falcons. On the "Stays with Atlanta or Retires" contract, Kalshi traded YES at 0.78, while Claude Opus 4.6 assigned $p = 0.03$, an implied disagreement of 0.75. Its rationale (recorded same-day, 2026-03-11):

> "Kirk Cousins is being released March 11, 2026 and wants to play. The Arizona Cardinals have the clearest QB vacancy after releasing Kyler Murray, with strong scheme fit under new HC Mike LaFleur (Shanahan/McVay system). Multiple insiders from NBC Sports, SI, and PFT have identified Arizona as the top landing spot, with Cousins offering a bridge starter option alongside weapons like Marvin Harrison Jr. Pittsburgh is the main alternative if Rodgers retires, while most other QB-needy teams (Vikings targeting Murray, Dolphins signed Willis, Jets got Geno Smith) have already addressed the position. Cousins may take his time signing, but Arizona's need is clear and urgent."

The model's edge here is real-time news synthesis: same-day reports confirming Cousins's intent to keep playing, combined with insider coverage identifying Arizona as the destination, and elimination of competing QB-needy teams that the market price had implicitly assumed were still in play. Cousins did not remain in Atlanta and did not retire. Under the Brier strategy, the position bought 0.75 NO shares at \$0.83, costing \$0.622; the contract resolved NO, paying \$0.75, for a return of +21%.

Case 2: Rain in Dallas, January 2026 (`KXRAINDALM-26JAN`). This 7-way ladder market asked the cumulative January precipitation in Dallas. On the "Above 2 inches" contract, Kalshi priced YES at 0.98 (with NO at 0.05), while Gemini 3 assigned $p = 0.22$. Its rationale (recorded 2026-01-25):

> "NWS data confirms recorded precipitation of 0.48 inches through Jan 23, with a winter storm forecast to add ∼0.5–0.8 inches of liquid equivalent, making the total likely to reach 1.0–1.5 inches. The provided market data was discounted due to internal

> inconsistencies (P(>3) > P(>2)) and contradiction with verified meteorological reports. Probabilities favor exceeding 1 inch but decline sharply for 2+ inches given the dry start to the month and limited remaining forecast precipitation."

The actual January total fell short of two inches. The Brier strategy bought 0.76 NO shares at 0.05, staking 0.038 and paying out 0.76—a return of $+1900\%$. Two factors made the disagreement actionable: (i) verifiable third-party meteorology directly contradicted the market's near-certainty, and (ii) the model identified an internal inconsistency in the order book—specifically, the probability for $P(> 3)$ inches was greater than that of $P(> 2)$ inches, violating monotonicity (due to factors like thin liquidity or delayed updates across related contracts). By detecting this structural mismatch, the model effectively inferred that at least one of the prices must be miscalibrated, and correspondingly discounted the market signal rather than treating it as fully informative.

In both cases, the model's advantage stems from combining verifiable public data with effective information synthesis. The models identify and integrate disparate signals – real-time reporting (Cousins), structured third-party data (NWS) – into coherent forecasts that the market had not yet fully incorporated. The Dallas precipitation case further illustrates an additional capability: detecting and reasoning about internal inconsistencies in market prices themselves.

As such, these examples suggest that LLM forecasters are particularly well-suited to settings with a large, readily available body of public information that is dispersed across sources, or where market prices contain internal inconsistencies. In these environments, LLMs can aggregate, cross-reference, and reconcile information more efficiently than humans, forming a coherent view faster than the market updates.