# When do prophets profit in prediction markets?

**Anri Gu**
University of Chicago
anrigu@uchicago.edu

**Nicole Kagan**
Kalshi Research
nkagan@kalshi.com

**Alec Sun**
University of Chicago
alecsun@uchicago.edu

**Jibang Wu**
New York University, Shanghai
wujibang@nyu.edu

**Haifeng Xu**
University of Chicago
haifengxu@uchicago.edu

## Abstract

Prediction markets aggregate dispersed beliefs into prices that act as probabilistic forecasts of uncertain events. Classical theory establishes how a better-than-market forecast can yield positive trading profit. However, it hinges crucially on the specific automated market maker (AMM) design, and is not applicable to today's popular exchanges, which are based on central limit order books. This paper fills that gap. For any prediction market and 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. Moreover, proper betting is unique, up to rescaling and payoff-irrelevant shifts, among strategies that guarantee profit for every ground truth under which the forecaster outperforms the market. Our proof rests on a decomposition of expected profit that strictly generalizes the classical AMM guarantee and also explains how strategies can profit even without an accuracy edge. Profit arises through two channels: a forecaster more accurate than the market is guaranteed positive expected profit, and a forecaster can also profit while being less accurate than the market when its large disagreements land on the correct side. Empirically, across thousands of forecasts by AI models, proper betting is the only strategy that reliably converts accuracy into profit, and we further identify systematic forecasting personas and show how the optimal proper strategy varies across them. As a feasibility demonstration, we run a month-long live pilot test on Kalshi.

Our result is a robust profitability guarantee, not profit maximization or an equivalence between profit and accuracy; the equivalence we establish is between strict properness of the scoring rule and the existence of such a guarantee. The live trading results are *not* trading suggestions: our limited live experiments only test whether the mechanism survives real-world spreads, fees, and limited liquidity, while professional trading involves many factors (drawdown control, variance, leverage, capital allocation) outside this paper's scope.

## 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"; their prices are well calibrated and often outperform polls and expert panels [Berg et al., 2008, Rothschild, 2009, Wolfers and Zitzewitz, 2004, Arrow et al., 2008]. The market scoring rule literature [Hanson, 2003, 2007, Chen and Pennock, 2007, Abernethy et al., 2013] formalizes the underlying

Authors are ordered alphabetically. The work by Jibang Wu was performed while visiting the University of Chicago.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

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, with expected profit equal to their accuracy edge. 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 their scalability, adaptability, and regulatory fit (see Appendix B). There, a forecaster can pick any bet size but faces whatever price impact the order book delivers, plus bid-ask spreads, finite liquidity, and fees. The link between accuracy and profit is no longer pinned down by the price-formation rule.

Recent empirical work makes this gap visible. Forecasters whose predictions outperform the market on average routinely fail to convert that edge into trading profit [Della Vedova, 2026, Jang et al., 2025]. On the AI leaderboard that turns AI forecasts into prediction-market trades, every agent loses money despite several beating the market under standard proper scoring rules [Yang et al., 2025]. Market frictions are not the whole story: even in an idealized market (no price impact, bid-ask spread, or fees), natural heuristics lose money on reasonably good forecasts (Kelly in Example 1, highest-margin betting in Example 2, and Section 4.1). Stranger still, the converse also fails: simple strategies earn positive expected profit even when the forecaster has no accuracy edge at all (Example 3). Translating accuracy into profit is therefore a problem in its own right, distinct from forecasting accuracy.

**The primary contribution of this paper** is a principled connection between predictive accuracy and robust trading profitability in general prediction markets with an arbitrary price-impact function. First, we exhibit a betting strategy that converts an accuracy edge over the market into profit:

**Theorem** (Informal version of Theorem 1)**.** *If a forecaster outperforms a sufficiently liquid prediction market under any strictly proper scoring rule $S$, then a specific "proper" betting strategy associated with $S$ ensures positive expected profit for the forecaster.*

The expected profit of this proper betting strategy decomposes into three explicit terms—score gap, Bregman divergence, and liquidity loss—which explain how strategies can profit even *without* an accuracy edge (the Bregman divergence alone can suffice). In AMMs, proper betting is precisely the canonical strategy that moves the price to the forecaster's belief, recovering the classic AMM guarantee as the boundary case where the divergence exactly absorbs the liquidity loss. Notably, Theorem 1 is a robust profitability guarantee, not an expected-profit maximization, and not an equivalence between profit and accuracy: the equivalence we establish (Proposition 2) is between strict properness and the existence of such a guarantee. The decomposition thus exposes two channels of profit: (i) a forecaster more accurate than the market is guaranteed positive expected profit (Theorem 1); (ii) a forecaster less accurate than the market can also profit, when its large disagreements land on the correct side of the market price—a channel that carries no guarantee. Our second main result shows that proper betting is essentially the *only* strategy that robustly guarantees profitability on any sufficiently liquid market, i.e., positive profit *whenever* there is a score edge; uniqueness holds up to rescaling and payoff-irrelevant shifts and concerns the whole strategy map $(\mathbf{p}, \mathbf{q}) \mapsto \mathbf{s}(\mathbf{p}, \mathbf{q})$, not a position at one fixed pair:

**Theorem** (Informal version of Theorem 2)**.** *Any robustly profitable betting strategy is "essentially" (Definition 3) the same as proper betting, up to profitability-invariant modifications such as proportional rescaling and constant shifting of bets.*

Together, the two results yield a new characterization of proper scoring rules through the lens of robust profitability: a scoring rule is strictly proper *if and only if* some betting strategy is robustly profitable under it (Proposition 2). We further extend our results to three practically important settings: betting on a sequence of events based on empirical (rather than expected) scores, betting under nonzero bid-ask spreads, and long-horizon trading in which both forecasts and market prices evolve.

Empirically, on thousands of AI-generated forecasts, proper betting shows a clear advantage over standard heuristic baselines, and we identify systematic forecasting personas that favor different proper betting strategies. Finally, to demonstrate practical feasibility, we deployed an AI forecasting agent with real capital on the Kalshi prediction market for a month. Its positive return on investment (ROI) is an existence proof that our strategy can survive real-world spreads, fees, discrete fills, and limited liquidity. Additional related work is discussed in Appendix A.

## 2 Preliminaries

**Notation.** Define $[K] := \{1, \dots, K\}$. Let $\Delta_{[K]} := \{\mathbf{p} \in [0,1]^K : \sum_{k\in[K]} p_k = 1\}$ denote the probability simplex, $\mathbf{1}_y \in \{0,1\}^K$ the one-hot vector with a 1 in the $y$-th entry, $\mathbf{1} \in \mathbb{R}^K$ the all-ones vector, $\nabla G$ the gradient of a function $G$, and $\|\cdot\|$ the Euclidean norm.

**Prediction markets.** Consider a probabilistic event with $K$ *disjoint* outcomes and *ground truth* $\mathbf{p}^* \in \Delta_{[K]}$. A prediction market on this event offers, for each outcome $k \in [K]$, a *contract* that pays 1 if outcome $k$ occurs and 0 otherwise. Let $\mathbf{q} = (q_k)_{k\in[K]} \in [0,1]^K$ denote the price vector, where $q_k$ is the price of the contract on outcome $k$. A forecaster's *betting strategy* is described by a position vector $\mathbf{s} \in \mathbb{R}^K$, where $s_k$ is the position on outcome $k$ ($s_k > 0$ buys, $s_k < 0$ sells). When the outcome $y \sim \mathbf{p}^*$ is realized, the forecaster receives payout $\mathbf{s} \cdot \mathbf{1}_y = s_y$.

We assume $\mathbf{q} \in \Delta_{[K]}$, so that $\mathbf{q}$ is the market's *consensus belief* [Arrow et al., 2008]: a forecaster who believes outcome $k$ occurs with probability $q_k$ has zero expected return from that contract. Any arbitrage-free market has $\sum_k q_k = 1$, since otherwise buying or selling one share of every contract yields a riskless profit $|1 - \sum_k q_k|$; real markets instead keep separate bid and ask prices, and Section 3.3 extends our results to nonzero spreads.

A market has finite liquidity at the quote price $\mathbf{q}$, described by a *price-impact function* $\rho : \mathbb{R}^K \to \Delta_{[K]}$, where $\rho(\mathbf{s})$ is the post-trade spot price upon executing the position $\mathbf{s}$, satisfying the boundary condition $\rho(\mathbf{0}) = \mathbf{q}$ and the monotonicity condition $(\rho(\mathbf{s}) - \rho(\mathbf{s}')) \cdot (\mathbf{s} - \mathbf{s}') \geq 0$ for all $\mathbf{s}, \mathbf{s}' \in \mathbb{R}^K$, i.e., walking the order book in any direction moves the marginal price in that same direction. Executing $\mathbf{s}$ incurs the integrated cost $\int_0^1 \rho(t\mathbf{s}) \cdot \mathbf{s}\, dt$, which decomposes into the spot-price piece $\mathbf{s} \cdot \mathbf{q}$ plus a *liquidity loss* (slippage) capturing the price impact along the trajectory:[1]

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

Monotonicity implies that the liquidity loss is always non-negative; the more liquid a market is, the smaller $L_\rho(\mathbf{s}; \mathbf{q})$ is (often close to 0 for reasonably liquid markets and reasonably sized bets). The forecaster collects $\mathbf{s} \cdot \mathbf{1}_y$ for a realized outcome $y$, so the expected profit is

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

The form of $\rho$ depends on the market design: classic AMMs [Hanson, 2003] set $\rho = \nabla C$ for a convex cost function $C$, whereas the largest prediction markets today run a central limit order book, for which $\rho$ is piecewise constant and read directly off the book [Kalshi, 2025, Polymarket, 2023]. We discuss both designs in Appendix B; our theory applies to any design. We suppress the evolution of $\rho$ over time, as each analysis concerns a single bet against a snapshot of the book.

**Scoring rules.** *Scoring rules* are a standard tool to measure the quality of a probabilistic forecast. Given a prediction $\mathbf{p} \in \Delta_{[K]}$ and a realized outcome $y \in [K]$, a scoring rule $S(\mathbf{p}, y)$ assigns a real-valued reward to $\mathbf{p}$, with expected score $S(\mathbf{p}; \mathbf{p}^*) := \mathbb{E}_{y\sim\mathbf{p}^*}[S(\mathbf{p}, y)]$ under ground truth $\mathbf{p}^*$. A scoring rule is *proper* if it incentivizes a forecaster to report their true belief.

**Definition 1** (Proper scoring rule)**.** *A scoring rule $S$ is* proper *if for every $\mathbf{p}, \mathbf{p}' \in \Delta_{[K]}$, $S(\mathbf{p}; \mathbf{p}) \geq S(\mathbf{p}'; \mathbf{p})$, and* strictly proper *if the inequality is strict for $\mathbf{p}' \neq \mathbf{p}$.*

A proper scoring rule lets us compare any two forecasts under a ground truth $\mathbf{p}^*$. Applying $S$ to the market consensus $\mathbf{q}$ gives the *market score* $S(\mathbf{q}; \mathbf{p}^*)$, and we say the forecaster *outperforms* the market if $S(\mathbf{p}; \mathbf{p}^*) > S(\mathbf{q}; \mathbf{p}^*)$. Importantly, this comparison depends on the ground truth $\mathbf{p}^*$, which is never observed in most market applications; Section 3.3 discusses comparisons based on empirical scores $S(\mathbf{p}, y)$ averaged across markets. Every proper scoring rule corresponds to a convex *potential function* $G$, by the following folklore result.

**Proposition 1** (Characterization of proper scoring rules [McCarthy, 1956a])**.** *A scoring rule $S$ is (strictly) proper if and only if there exists a (strictly) convex function $G : \Delta_K \to \mathbb{R}$ such that $S(\mathbf{p}, y) = G(\mathbf{p}) + \nabla G(\mathbf{p}) \cdot (\mathbf{1}_y - \mathbf{p})$.*

[1] For non-conservative $\rho$ the cost is path-dependent; we use the straight path $t \mapsto t\mathbf{s}$.

Throughout, we take $G$ differentiable on the relative interior of the simplex and, when $\nabla G$ is unbounded at the boundary (e.g., the logarithmic rule), restrict forecasts and prices to the relative interior; in general, Proposition 1 holds with $\nabla G$ replaced by a subgradient [Gneiting and Raftery, 2007]. The *Bregman divergence* of a strictly convex differentiable $G$ 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 if and only if $\mathbf{p} = \mathbf{q}$. Table 1 lists common proper scoring rules with their associated functions (the last column $\mathbf{s}_G$ is introduced in Definition 2).

**Table 1:** Common proper scoring rules with their associated functions.

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

## 3 A theory of proper betting for prediction markets

Our goal is to translate a forecast $\mathbf{p}$ that outperforms the market into a concrete betting strategy that guarantees profit in a general prediction market. There are several natural strategies, such as betting on the highest-margin outcome, scaling inversely with margin, or using the Kelly criterion [Kelly, 1956]. The *highest-margin* strategy bets on the outcome $k$ with the greatest $|p_k - q_k|$, as this outcome intuitively has the most potential for profit: the gap between the perceived value $p_k$ of a share and its price $q_k$ is greatest. Perhaps surprisingly, neither the Kelly criterion nor the highest-margin strategy guarantees positive expected profit, so betting optimization is challenging even for a good forecaster; Section 4 shows that these approaches also perform poorly in practice.

**Example 1** (Accurate forecaster + Kelly can lose)**.** *Consider an event with three outcomes and let* $\mathbf{p} = (0.64, 0.32, 0.04)$*,* $\mathbf{q} = (0.80, 0.10, 0.10)$*, and* $\mathbf{p}^* = (0.10, 0.10, 0.80)$*. Under the quadratic scoring rule, since* $\|\mathbf{p}^* - \mathbf{p}\| < \|\mathbf{p}^* - \mathbf{q}\|$*, the forecaster outperforms the market. The Kelly criterion spends* $p_i$ *to purchase* $p_i/q_i$ *shares of contract* $i$ *for each* $i = 1, 2, 3$*,*[2] *and its expected profit is* $(\mathbf{p} \oslash \mathbf{q}) \cdot (\mathbf{p}^* - \mathbf{q}) < 0$*. The intrinsic reason is that Kelly is derived by assuming the forecast is the ground truth, which in practice can be far off.*

**Example 2** (Accurate forecaster + highest-margin trade 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 bets on the highest-margin first outcome, their expected profit is* $(1, 0, 0) \cdot (\mathbf{p}^* - \mathbf{q}) < 0$*.*

On the other hand, a forecaster can profit even if their prediction underperforms the market:

**Example 3** (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 bets on the first (highest-margin) outcome, their expected profit is* $(1, 0) \cdot (\mathbf{p}^* - \mathbf{q}) > 0$*.*

To resolve this seemingly broken link between accuracy and profitability, we define, for every proper scoring rule $S$, a corresponding *proper* betting strategy.

**Definition 2** (Proper betting strategy)**.** *Under a proper scoring rule* $S$ *with potential function* $G$*, the* proper *betting strategy for a forecast* $\mathbf{p}$ *and a market price* $\mathbf{q}$ *is the position vector* $\mathbf{s}_G(\mathbf{p}, \mathbf{q}) := \nabla G(\mathbf{p}) - \nabla G(\mathbf{q})$*.*

If the $k$-th entry of $\mathbf{s}_G(\mathbf{p}, \mathbf{q})$ is negative, the corresponding execution is to buy a *NOT* $k$ contract at price $1 - q_k$ (assuming no bid-ask spread for now).

### 3.1 Proper betting and its robust profitability

**On robust profitability.** A fundamental challenge in designing betting strategies is that we almost never observe the ground truth $\mathbf{p}^*$. Even when a forecast $\mathbf{p}$ is promised to score better than the market $\mathbf{q}$—in the sense that $S(\mathbf{p}; \mathbf{p}^*) \geq S(\mathbf{q}; \mathbf{p}^*)$ for some proper scoring rule $S$—this can hold for

[2] This is the general many-outcome Kelly solution; the familiar two-outcome formula is a special case (Appendix C.5).

infinitely many $\mathbf{p}^*$'s. A desirable property is thus *robust profitability*: a strategy $\mathbf{s}(\mathbf{p}, \mathbf{q})$ is robustly profitable under $S$ (for a given $\rho$) if it earns positive expected profit for all $\mathbf{p} \neq \mathbf{q}$ and $\mathbf{p}^*$ satisfying $S(\mathbf{p}; \mathbf{p}^*) \geq S(\mathbf{q}; \mathbf{p}^*)$, and is not robustly profitable otherwise. This is a binary notion: it concerns only whether positive profit can be guaranteed, not its magnitude. Our first main result shows that a robustly profitable strategy exists whenever the market's liquidity loss is relatively small; in fact, proper betting is such a strategy.

**Theorem 1** (Robust profitability of proper betting)**.** *For any ground truth $\mathbf{p}^*$, forecast $\mathbf{p}$, market price $\mathbf{q}$, and strictly proper scoring rule $S$ with potential function $G$, the expected profit of the proper betting strategy $\mathbf{s}^* = \mathbf{s}_G(\mathbf{p}, \mathbf{q})$ decomposes as*

$$\pi(\mathbf{s}^*, \mathbf{p}^*) \;=\; \underbrace{S(\mathbf{p}; \mathbf{p}^*) - S(\mathbf{q}; \mathbf{p}^*)}_{\textit{score gap}} \;+\; \underbrace{D_G(\mathbf{q}, \mathbf{p})}_{\textit{Bregman divergence}} \;-\; \underbrace{L_\rho(\mathbf{s}^*; \mathbf{q})}_{\textit{liquidity loss}}. \tag{1}$$

*Therefore, if $D_G(\mathbf{q}, \mathbf{p}) \geq L_\rho(\mathbf{s}^*; \mathbf{q})$—i.e., the divergence offsets the liquidity loss—then proper betting ensures positive profit whenever $S(\mathbf{p}; \mathbf{p}^*) > S(\mathbf{q}; \mathbf{p}^*)$, and is* robustly profitable *if the offset is strict at every pair $(\mathbf{p}, \mathbf{q})$, e.g. in a frictionless market.*

*Proof.* The proof hinges on the following lemma, which may be of independent interest.

**Lemma 1** (Profit decomposition lemma)**.** *For any realized outcome $y$, forecast $\mathbf{p}$, market price $\mathbf{q}$, and scoring rule $S(\mathbf{p}, y)$ with potential function $G$ (not necessarily proper),*

$$\underbrace{\mathbf{s}_G(\mathbf{p}, \mathbf{q}) \cdot (\mathbf{1}_y - \mathbf{q})}_{\textit{Idealized profit from outcome } y} \;=\; \underbrace{S(\mathbf{p}, y) - S(\mathbf{q}, y)}_{\textit{Realized score gap}} + \underbrace{D_G(\mathbf{q}, \mathbf{p})}_{\textit{Bregman divergence}}. \tag{2}$$

The proof is a direct expansion of the definition of $D_G$ (Appendix C.1). Taking the expectation of Eq. (2) over $y \sim \mathbf{p}^*$ gives $\mathbf{s}_G(\mathbf{p}, \mathbf{q}) \cdot (\mathbf{p}^* - \mathbf{q}) = S(\mathbf{p}; \mathbf{p}^*) - S(\mathbf{q}; \mathbf{p}^*) + D_G(\mathbf{q}, \mathbf{p})$, pricing all shares at $\mathbf{q}$; subtracting the slippage $L_\rho(\mathbf{s}^*; \mathbf{q})$ yields $\pi(\mathbf{s}^*, \mathbf{p}^*)$. □

Like Lemma 1, Theorem 1 also holds "pointwise" for each realized outcome $y$ (the special case $\mathbf{p}^* = \mathbf{1}_y$), which lets us extend the guarantee to empirical scores in Section 3.3.

**Remark 1.** *The key insight of Theorem 1 is that $\mathbf{s}_G(\mathbf{p}, \mathbf{q}) = \nabla G(\mathbf{p}) - \nabla G(\mathbf{q})$ is the "proper" strategy to use, and its profit admits a clean characterization. The decomposition formalizes a natural intuition: a profitable forecast is not merely accurate (a large score gap), but sufficiently different from the market while remaining sufficiently accurate. Conversely, positive profit does not imply accuracy: it can be attributed entirely to the Bregman bonus. This explains the paradox in Example 3: when a prediction is far from the market price, a large Bregman divergence can outweigh a negative score gap. The divergence is not free money for disagreeing, however: if the market price is in fact correct ($\mathbf{p}^* = \mathbf{q}$), the score gap equals exactly $-D_G(\mathbf{q}, \mathbf{p})$ and expected profit is $-L_\rho(\mathbf{s}^*; \mathbf{q}) \leq 0$ under every proper rule. Divergence-driven profit is therefore compensation for contrarian risk that is earned only when the forecaster lands on the correct side of the market price—a real but weaker form of skill than outperforming the market under $S$. Note also that the condition $D_G(\mathbf{q}, \mathbf{p}) \geq L_\rho(\mathbf{s}^*; \mathbf{q})$ involves no ground truth, and a small enough rescaling $\alpha \cdot \mathbf{s}^*$ satisfies its scaled form $\alpha D_G(\mathbf{q}, \mathbf{p}) \geq L_\rho(\alpha \cdot \mathbf{s}^*; \mathbf{q})$ in any market whose price impact vanishes for small orders (Remark 4). (Theorem 1 even holds for $\mathbf{q} \notin \Delta_K$; see Appendix D.1.)*

### 3.2 Proper betting is (essentially) the only robustly profitable strategy

Theorem 1 establishes the robust profitability of proper betting in a market with low liquidity loss. Our second main result shows that it is essentially the only robustly profitable strategy, where "essentially" rules out trivial transformations such as rescaling ($\lambda \mathbf{s}_G(\mathbf{p}, \mathbf{q})$, $\lambda > 0$) or constant shifting ($\mathbf{s}_G(\mathbf{p}, \mathbf{q}) + \lambda \mathbf{1}$), which do not change whether a strategy is profitable.

**Definition 3** (Essentially different betting)**.** *Let $\hat{\mathbf{s}} := P\mathbf{s}/\|P\mathbf{s}\|$ be the* normalized direction *of a strategy $\mathbf{s}(\mathbf{p}, \mathbf{q})$, with $P$ the orthogonal projection onto $\mathbf{1}^\perp$; $\hat{\mathbf{s}}$ is invariant under $\mathbf{s} \to \lambda \mathbf{s}$ ($\lambda > 0$) and $\mathbf{s} \to \mathbf{s} + \mu \mathbf{1}$, and undefined if $P\mathbf{s} = \mathbf{0}$. Strategies $\mathbf{s}, \mathbf{s}'$ are* essentially the same *if, for every interior $\mathbf{q} \in \mathrm{int}(\Delta_K)$ and every sequence $\mathbf{p}_t \to \mathbf{q}$ ($\mathbf{p}_t \neq \mathbf{q}$), both directions are eventually defined and $\hat{\mathbf{s}}(\mathbf{p}_t, \mathbf{q}) - \hat{\mathbf{s}}'(\mathbf{p}_t, \mathbf{q}) \to \mathbf{0}$; otherwise they are* essentially different.

The real test is the limit $\mathbf{p} \to \mathbf{q}$: away from it, any robustly profitable strategy can be superficially modified, e.g., $\mathbf{s}_G + \epsilon\,\mathbf{e}_1$ stays profitable whenever $\epsilon$ is below the profit of $\mathbf{s}_G$ (at least $D_G(\mathbf{q}, \mathbf{p})$), even at infinitely many well-separated $(\mathbf{p}, \mathbf{q})$ pairs.

**Theorem 2.** *Any robustly profitable betting strategy is essentially the same as the proper betting* $\mathbf{s}_G(\mathbf{p}, \mathbf{q})$. *Formally, for any strictly proper scoring rule* $S$ *and any betting strategy* $\mathbf{s}(\mathbf{p}, \mathbf{q})$ *that is essentially different from* $\mathbf{s}_G(\mathbf{p}, \mathbf{q})$*, there exist* $\mathbf{p}, \mathbf{p}^*, \mathbf{q}$ *such that* $S(\mathbf{p}; \mathbf{p}^*) > S(\mathbf{q}; \mathbf{p}^*)$ *but the expected profit of* $\mathbf{s}(\mathbf{p}, \mathbf{q})$ *is non-positive, and negative unless* $\mathbf{s}(\mathbf{p}, \mathbf{q}) \in \mathbb{R} \cdot \mathbf{1}$. *This holds even for frictionless markets, i.e.,* $L_\rho = 0$.

Theorem 2 is a statement about the strategy map $(\mathbf{p}, \mathbf{q}) \mapsto \mathbf{s}(\mathbf{p}, \mathbf{q})$, not about a position at one fixed pair: at any fixed $(\mathbf{p}, \mathbf{q})$, every position close enough in direction to $\mathbf{s}_G$ profits under all favorable $\mathbf{p}^*$, so no pointwise uniqueness can hold. The limit $\mathbf{p} \to \mathbf{q}$ enters only through the definition of when two strategies count as distinct; the uniqueness itself is global, i.e., proper betting is unique up to rescaling and payoff-irrelevant shifts among strategies that guarantee profit for all ground truths under which the forecaster outperforms.

*Proof sketch of Theorem 2.* Frictionless markets suffice. By Lemma 1, the ground truths under which $\mathbf{p}$ scores at least as well as $\mathbf{q}$ form a half-space cap $\{\mathbf{p}^* : \mathbf{s}_G \cdot (\mathbf{p}^* - \mathbf{q}) \geq D_G(\mathbf{q}, \mathbf{p})\}$ whose boundary is within $\|\mathbf{p} - \mathbf{q}\|$ of $\mathbf{q}$; as $\mathbf{p} \to \mathbf{q}$ it fills the half-neighborhood of $\mathbf{q}$ where $\mathbf{s}_G \cdot (\mathbf{p}^* - \mathbf{q}) > 0$, so a strategy profitable on all of it must align with $\mathbf{s}_G$ up to an angle $O(\|\mathbf{p} - \mathbf{q}\|)$ (Appendix C.2). □

### 3.3 Implications and real market considerations

**A new characterization of proper scoring rules.** Since $\mathbf{s}_G(\mathbf{p}, \mathbf{q})$ is well-defined for any $G$, is properness necessary for robust profitability? It is, in a strong sense: strict properness of $S$ is necessary for the existence of *any* robustly profitable strategy.

**Proposition 2** (Equivalence between score properness and robust profitability)**.** *A scoring rule* $S(\mathbf{p}, y)$ *is strictly proper* if and only if *there exists a betting strategy* $\mathbf{s}(\mathbf{p}, \mathbf{q})$ *that guarantees positive profit in a frictionless market (*$L_\rho = 0$*) for any* $\mathbf{p}^*$ *and* $\mathbf{p} \neq \mathbf{q}$ *satisfying* $S(\mathbf{p}; \mathbf{p}^*) \geq S(\mathbf{q}; \mathbf{p}^*)$.

Proposition 2 (proved in Appendix C.3) thus characterizes strictly proper scoring rules as precisely those admitting a robustly profitable betting strategy.

**Proper betting strictly generalizes canonical betting in AMMs.** An AMM is governed by a convex cost function $C(\mathbf{x})$, strictly convex (on $\mathbf{1}^\perp$), such that the market price $\mathbf{q} \in \Delta_K$ and the net number of shares of each contract $\mathbf{x} \in \mathbb{R}^K$ satisfy $\nabla C(\mathbf{x}) = \mathbf{q}$ [Hanson, 2007, Chen and Pennock, 2007]. A forecaster with belief $\mathbf{p}$ maximizes their expected profit by purchasing the shares $\mathbf{s}_{\mathbf{q}\to\mathbf{p}}$ that move the price from $\mathbf{q}$ to $\mathbf{p}$, i.e., $\nabla C(\mathbf{x} + \mathbf{s}_{\mathbf{q}\to\mathbf{p}}) = \mathbf{p}$ (Appendix B). The convex conjugate $G(\mathbf{p}) = \sup_{\mathbf{x}} \{\mathbf{p} \cdot \mathbf{x} - C(\mathbf{x})\}$ is precisely the potential of the market scoring rule, and $(\nabla C)^{-1} = \nabla G$. Hence the canonical AMM bet *is* proper betting under the market scoring rule:

$$\mathbf{s}_{\mathbf{q}\to\mathbf{p}} = (\mathbf{x} + \mathbf{s}_{\mathbf{q}\to\mathbf{p}}) - \mathbf{x} = (\nabla C)^{-1}(\mathbf{p}) - (\nabla C)^{-1}(\mathbf{q}) = \nabla G(\mathbf{p}) - \nabla G(\mathbf{q}) = \mathbf{s}_G(\mathbf{p}, \mathbf{q}). \quad (3)$$

A basic fact about AMMs is that moving the price from $\mathbf{q}$ to $\mathbf{p}$ earns $S(\mathbf{p}; \mathbf{p}^*) - S(\mathbf{q}; \mathbf{p}^*)$ [Hanson, 2007], exactly the first term of Theorem 1. This is no coincidence: the divergence and the liquidity loss cancel precisely in any AMM (proved in Appendix C.4).

**Corollary 1** (Instantiation of Theorem 1 in AMMs)**.** *In any AMM with a cost function* $C(\mathbf{x})$ *(strictly convex on* $\mathbf{1}^\perp$*) and market scoring rule* $S$ *induced by its convex conjugate* $G$*,* $D_G(\mathbf{q}, \mathbf{p}) = L_\rho(\mathbf{s}^*; \mathbf{q})$ *for the proper bet* $\mathbf{s}^* = \nabla G(\mathbf{p}) - \nabla G(\mathbf{q})$*. Hence the profit of* $\mathbf{s}^*$ *is* $S(\mathbf{p}; \mathbf{p}^*) - S(\mathbf{q}; \mathbf{p}^*)$.

The foundational result of Hanson [2003, 2007] thus already implies that $\mathbf{s}_{\mathbf{q}\to\mathbf{p}}$ profits under every strict edge in AMMs; Theorem 2 adds that it is essentially the only such strategy.

**Extension to empirical scores.** Since $\mathbf{p}^*$ is unobservable, a forecaster's expected advantage on a single event can never be validated, but an advantage in empirical scores can. Because Theorem 1 holds for each realized outcome, it extends to a sequence of $n$ events with predictions $\mathbf{p}_i$, prices $\mathbf{q}_i$, and realized outcomes $y_i$. Let $\hat{S}_F := \frac{1}{n}\sum_i S(\mathbf{p}_i, y_i)$ and $\hat{S}_M := \frac{1}{n}\sum_i S(\mathbf{q}_i, y_i)$ be the empirical forecast and market scores.

**Corollary 2** (Empirical version of Theorem 1)**.** *The sequential proper bet* $\mathbf{s}_i := \mathbf{s}_G(\mathbf{p}_i, \mathbf{q}_i)$ *on realized outcomes* $y_1, \ldots, y_n$ *yields the realized average profit*

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

Hence sequential proper betting profits if the empirical score gap and the average divergence together offset the average liquidity loss; Section 4 indeed observes forecasters profiting from divergence despite below-market accuracy. Bet sizes should thus *vary* across events, proportionally to $\mathbf{s}_i$, with larger bets where the prediction deviates more from the market. Corollary 2 is a realized (pathwise) statement: on any outcome sequence with $\hat{S}_F > \hat{S}_M$, the realized return exceeds the average Bregman divergence minus the average liquidity loss. The common proportionality constant is a free risk knob: scaling all positions by $c \in (0, 1]$ scales idealized profit linearly and its variance quadratically without changing its sign, and Remark 5 turns the expected guarantee into a high-probability bound on realized return. Choosing the constant optimally depends on liquidity, budget, risk tolerance, and execution constraints, which we leave to future work.

**Extension to nonzero bid-ask spreads and correlated outcomes.** Low liquidity and fees create a spread: for outcome $k$, $q_k^+ + q_k^- \geq 1$, where $q_k^+$ and $q_k^-$ are the prices to buy *YES* $k$ and *NOT* $k$, so a negative entry of $\mathbf{s}_G$ can no longer be executed at $1 - q_k^+$. The key idea is to decompose each outcome $k$ into a binary event with prices $q_k^+$ and $q_k^-$. The resulting proper bet buys *YES* $k$ when $p_k > q_k^+$, buys *NOT* $k$ when $p_k < 1 - q_k^-$, and abstains otherwise; its profit admits the same decomposition up to a cost of spread. Treating each outcome as a separate binary market also covers non-disjoint outcomes (e.g., "Bitcoin at least \$7K", "at least \$8K"; Appendix D.3).

**Extension to sequential betting.** As forecasts $\mathbf{p}^t$ and prices $\mathbf{q}^t$ evolve before resolution, Appendix D.4 extends Theorem 1 to a *fundamental*-driven strategy that executes $\mathbf{s}_G(\mathbf{p}^t, \mathbf{q}^t)$ each round and holds to resolution, and a *momentum*-driven strategy that maintains $\mathbf{s}_G(\mathbf{p}^t, \mathbf{q}^t)$ as a rebalanced position. Both inherit the decomposition summed across rounds; the former needs a per-round edge against the unobservable $\mathbf{p}^*$, the latter an edge against the next price $\mathbf{q}^{t+1}$, which is observable. The rebalancing rules of our live agent (Section 4.3) are empirical choices; we give no formal risk-control guarantee for them.

## 4 Experiments

Testing betting strategies requires data pairing forecasts with contemporaneous market prices and realized outcomes, which is hard to collect from human forecasters at scale; AI forecasting benchmarks built on real-money prediction markets, most notably Prophet Arena [Yang et al., 2025], provide a scalable alternative. Our evaluation has two stages: an offline study on thousands of archived markets, where we simulate each market under zero price impact, and a live pilot test that exposes our strategy to the full set of real-market frictions.

### 4.1 How good are proper betting strategies?

We use forecasts made by 18 frontier models, each with predictions on at least 600 resolved Kalshi and Polymarket events (up to approximately 3,900 for the most heavily covered models), for a total of 94,453 forecasts[3] collected in collaboration with Sooth Labs, spanning domains such as sports, politics, economics, and crypto. Letters appended to model names denote distinct prompt variants of the same underlying model. Each market includes a forecast, the contemporaneous bid and ask prices, and the realized outcome; for each model and strategy, we simulate the resulting trades and compute the realized return after resolution. *Provenance.* This dataset was added after peer review; it was collected by Sooth Labs' production forecasting pipeline, independently of the Prophet Arena data used in the submitted version (whose tables, with event-clustered intervals, are in Appendix F.1), and is described in Appendix F.5. Two cuts of these data exist. Tables 2 and 3 (and their full versions in Appendices E.2 and E.3) use the per-lane *conditioned* cut released during review, with 29,713

[3]Many forecasts may be made per event at different times. Full betting data is released here. At the time of writing the site serves the conditioned cut used in this section; the unconditioned full set of Appendix F.5 is not yet posted there.

**Table 2:** Proper Betting (Brier) vs. each baseline heuristic for selected top forecasting agents; each column reports the difference in ROI (%) between Proper Betting and the heuristic. Intervals are paired, event-clustered 95% bootstrap CIs; bold marks differences whose CI excludes zero. ROI is realized profit over total cost staked, $\sum_i |w_i|\, q_i$, with $w_i$ in shares of the traded contract and $q_i$ its price (Appendix E.2). Kelly compounds one bankroll; Kelly-Alike is its fresh-budget analogue. Kelly entries reflect uncapped leverage: per-outcome Kelly fractions are summed within an event without a joint budget constraint, so the stake on an event can exceed the bankroll (on the submitted Prophet Arena data, up to $34\times$ on one event; Appendices E.2 and F.1).

| Model | $\Delta S$ | − Max (Mkt) | − Max (Grp) | − Inv-Margin | − Kelly-Alike | − Kelly |
|---|---|---|---|---|---|---|
| Gemini 3.1 Pro (B) | **+0.0076** [+0.0014, +0.0136] | **+30.3** [+20.7, +39.9] | **+30.1** [+15.5, +42.9] | **+39.2** [+27.7, +50.6] | **+24.3** [+16.1, +32.8] | **+138.9** [+127.3, +150.4] |
| Claude Fable 5 | **+0.0039** [+0.0008, +0.0073] | **+26.8** [+15.3, +40.0] | **+48.3** [+34.2, +63.1] | **+33.6** [+20.0, +48.9] | **+16.8** [+7.3, +27.4] | +129.0 [$-1.3 \times 10^9$, +141.8] |
| GPT-5.6 Sol | **+0.0033** [+0.0003, +0.0067] | **+22.4** [+12.2, +35.4] | **+39.9** [+24.3, +55.7] | **+28.5** [+15.7, +43.7] | **+15.7** [+6.0, +27.1] | -1220.4 [$-2.2 \times 10^8$, +122.5] |
| GPT-5.5 (B) | -0.0059 [-0.0129, +0.0009] | +13.0 [-1.2, +29.3] | +20.3 [-0.7, +41.1] | **+18.6** [+0.1, +38.9] | +8.6 [-2.7, +21.4] | **+116.8** [+99.8, +136.0] |
| GPT-5.5 (C) | -0.0056 [-0.0101, -0.0012] | +8.8 [-1.5, +18.7] | +3.6 [-28.4, +27.3] | +7.0 [-5.5, +19.4] | +2.0 [-6.3, +10.0] | **+107.3** [+94.9, +119.2] |

events across the 18 models; Appendix F.5 repeats the analysis on the *unconditioned* full set (30,889 events), in which no event or forecast is removed on the basis of its score or return. The two differ in a small share of events, but the difference matters: under the full set no model has a significantly positive score gap, while proper betting remains significantly profitable for several models through the divergence term alone. The tables below should be read together with Appendix F.5. The per-lane selection rule behind the conditioned cut was not recorded when that cut was released, so it cannot be restated here; Appendix F.5, which uses every forecast, is therefore the reference analysis, and the conditioned-cut tables are kept because they are what the reviewers saw.

A betting strategy determines a set of weights—how to allocate a fixed budget, and in which direction—over a collection of markets with forecasts and prices. We implement the three proper betting strategies of Table 1 and compare them with baselines derived from intuitive heuristics: (i) **Max-Margin**, which concentrates capital on the largest disagreements with the market, either across all markets or via the single largest-margin bet per event group; (ii) **Inverse-Margin**, which down-weights large disagreements under the assumption that extreme deviations are less reliable; (iii) **Kelly-Alike**, which scales exposure by Kelly-style edge-to-odds ratios; and (iv) the **Kelly Criterion** [Kelly, 1956], which maximizes asymptotic log-wealth growth; Kelly compounds a single bankroll across bets, whereas all other strategies, including Kelly-Alike—its de-compounded analogue—are evaluated on a fresh unit budget, so Kelly-Alike is the like-for-like comparison.[4] Appendix E.2 details the implementation of all strategies.

Table 2 reveals why proper betting warrants study: heuristic allocations fail to yield consistently positive returns, and for stronger models, Brier-weighted allocation is the only strategy that reliably produces positive ROI. Inverse-Margin limits downside for weaker models by concentrating capital on near-market bets, but by design also limits the upside. Accuracy alone is thus insufficient to create profit: how probabilities become position sizes is a key determinant of returns (all models in Appendix E.2).

### 4.2 How to choose proper betting strategies for different forecaster personas?

**Score gap vs. divergence.** Table 3 illustrates the decomposition of Lemma 1, with the score gap and divergence summed across bets and normalized by total spend (ROI units). Log and Sph denote the *per-contract* log and spherical rules that our executable, buys-only implementation uses on separate YES/NO books (Appendix E.2); their score gaps are measured under the induced per-contract scores and coincide with the standard (netted) log and spherical rules only at zero spread. Profitability depends not only on the score gap ($\Delta S$) but also on the divergence ($D$), so models with worse accuracy can achieve higher ROI if they generate sufficiently large divergence. For example, under the Log rule, Gemini 3.1 Pro (B) achieves higher ROI than GPT-5.6 Sol despite a worse score gap. Moreover, a model's score under a given rule does not by itself determine which rule yields the best returns, since divergence is rule-dependent: for GPT-5.5 (B), Brier yields a worse score gap than Log ($-15.4$ vs. $-9.5$) but a higher ROI ($+16.8\%$ vs. $+11.7\%$), driven by a substantially larger divergence (32.2 vs. 21.2). Appendix E.3 reports the decomposition for all models. GPT-5.5 (B) is the clearest

[4] Kelly performs particularly poorly because it is highly sensitive to miscalibrated probabilities, which makes it unstable under model error, and because its per-outcome fractions are summed without a joint budget cap, so its stakes can exceed the bankroll (Appendix E.2).

**Table 3: ROI decomposition** $\mathrm{ROI} = \Delta S + D$ for each model and proper scoring rule. Bold: significantly positive score gap (95% CI above zero); bold italic: significantly positive ROI. All terms are summed over bets and divided by total cost staked $\sum_i |w_i| q_i$; Log/Sph are the per-contract rules of Appendix E.2. $\Delta S$, $D$, and ROI are computed separately, so the identity is checked rather than imposed.

| Model | Brier $\Delta S$ | Brier $D$ | Brier ROI | Log $\Delta S$ | Log $D$ | Log ROI | Sph $\Delta S$ | Sph $D$ | Sph ROI | Events |
|---|---|---|---|---|---|---|---|---|---|---|
| Gemini 3.1 Pro (B) | **+11.7** [+2.4, +21.1] | +27.2 [+22.7, +32.1] | ***+38.9*** *[+27.4, +50.9]* | **+7.5** [+0.8, +14.0] | +20.9 [+18.2, +23.8] | ***+28.4*** *[+20.4, +36.2]* | **+11.3** [+3.5, +19.5] | +21.9 [+18.3, +25.8] | ***+33.2*** *[+23.9, +42.9]* | 1113 |
| Claude Fable 5 | **+12.3** [+3.2, +22.0] | +16.7 [+11.8, +22.2] | ***+29.0*** *[+16.7, +42.7]* | **+8.6** [+1.3, +15.6] | +12.9 [+9.8, +16.5] | ***+21.5*** *[+12.2, +30.6]* | **+10.2** [+1.4, +18.6] | +11.9 [+8.5, +15.9] | ***+22.1*** *[+11.6, +33.0]* | 1140 |
| GPT-5.6 Sol | **+10.9** [+1.0, +21.7] | +15.9 [+12.1, +21.9] | ***+26.7*** *[+14.4, +41.4]* | **+10.7** [+3.0, +18.5] | +12.6 [+10.1, +16.1] | ***+23.3*** *[+14.5, +32.4]* | **+12.1** [+3.2, +21.4] | +11.4 [+8.5, +15.7] | ***+23.5*** *[+13.0, +35.8]* | 866 |
| GPT-5.5 (B) | -15.4 [-29.6, -0.4] | +32.2 [+25.8, +39.5] | ***+16.8*** *[+2.3, +33.3]* | -9.5 [-20.7, +0.5] | +21.2 [+17.9, +24.9] | ***+11.7*** *[+0.3, +22.2]* | -11.5 [-23.0, -0.7] | +24.4 [+19.4, +30.1] | ***+12.9*** *[+0.9, +25.0]* | 842 |
| GPT-5.5 (C) | -15.6 [-26.3, -5.6] | +22.9 [+18.6, +27.3] | +7.3 [-3.3, +17.3] | -13.1 [-21.4, -5.5] | +18.4 [+15.2, +21.6] | +5.2 [-3.8, +13.9] | -14.6 [-23.4, -6.2] | +16.5 [+13.4, +20.1] | +1.9 [-6.6, +10.7] | 979 |

divergence-driven case: its ROI is significantly positive under all three rules although its score gap is negative (significantly so under Brier and Sph). As Theorem 1 permits, such profit is compensation for contrarian risk on the correct side of the market (Section 3), not evidence of superior accuracy.

**Forecaster personas.** As Table 3 shows, scoring rules yield materially different returns. To identify when each rule is preferable, we construct four synthetic forecasting "personas" that achieve the same Brier score on the same real markets (27,516 markets across 3,511 questions from the Prophet Arena pipeline, August 1, 2025 to April 15, 2026) but deviate from market prices differently. Personas are defined by (i) the share of small-margin bets ($|p - q| \leq 0.15$) and (ii) the accuracy drop-off across margins: the slope of the directional win rate (the fraction of bets whose chosen side matches the outcome) against the margin $|p - q| \in [0, 0.5]$; larger margins are too sparse for reliable analysis (see Appendix E.8). As illustrated in Figure 1 (Appendix E.4), **Conservative** personas concentrate most forecasts at small margins with flat accuracy across margins; **Aggressive** personas place fewer small-margin bets, also with flat accuracy; **Dispersed** personas spread bets broadly across margins with a gradual decline in win rate; and **Brittle** personas have many small-margin forecasts but a sharply declining win rate as the margin increases (generation details in Appendix E.4).

We evaluate each persona in two regimes with a symmetric $\pm 0.05$ quadratic score gap relative to the market (market score $0.839$), holding average accuracy constant: in Regime A the persona outperforms the market and all strategies profit; in Regime B it underperforms and all strategies lose. As shown in Table 7 (Appendix E.4), Brier is optimal in Regime A for all personas, while Log is optimal in Regime B. In Regime A, the positive Brier gap implies positive expected returns across margin bins, so linear scaling in $|p - q|$ is optimal. In Regime B, losses are dominated by high-margin errors, making the per-contract Log rule's sublinear weighting preferable as it attenuates exposure to large, costly mistakes; the netted standard log bet, $\mathrm{logit}(p) - \mathrm{logit}(q)$, instead grows superlinearly, so this conclusion is specific to the per-contract rule. This effect is most pronounced for Conservative personas, which must incur a high error rate on small-margin bets to sustain a negative Brier gap. Weighting visualizations are provided in Appendix E.5.

**Mapping real models to personas.** We evaluate each real model on the two defining dimensions over its full event history, since these patterns only emerge reliably at scale. Figure 3a reports each model's share of small-margin forecasts, and Figure 3b its win rate as a function of margin (both in Appendix E.6). Models separate cleanly into three persona groups: Brittle models concentrate most bets at small margins with steeply declining win rates, Dispersed models spread forecasts more broadly with gradually declining win rates, and Conservative models place nearly all bets at small margins with roughly flat win rates. Personas thus summarize how model behavior translates into returns and can guide the choice of betting strategy (all models are classified in Appendix E.6).

### 4.3 A live pilot test under active portfolio management

We further evaluate proper betting in a live portfolio setting using the momentum-driven strategy of Section 3.3 (details in Appendix D.4). Our goal is *not* to establish statistically reliable trading alpha or optimal risk-adjusted performance, which is impossible given our resource and legal constraints, but to *test whether proper betting can survive live execution*, which—unlike the backtests—must contend with bid-ask spreads, limited liquidity, slippage, and discrete tick sizes.

We deploy Gemini 3 with the Brier-derived betting strategy on Kalshi for 45 calendar days (24 March–7 May 2026), trading on 26 of them, with an initial budget of $200. Every two hours, the agent evaluates all open markets that pass an ex-ante eligibility filter excluding contracts with

ambiguous resolution criteria and settings known to impair LLM forecasting (Appendix E.7.2). Figure 4 (Appendix E.7) shows the cumulative ROI; the full trading history is documented here. The agent executed 236 orders across 129 markets, ending the window with a net profit of \$160.67 on \$200 (ROI $+80.33\%$). This headline figure carries wide uncertainty: $+\$56.32$ ($+28.2\%$) was realized on settled or closed markets and the remaining 65% was the marked value of 31 positions still open at the end of the window; the three most profitable markets account for 54% of gross profit and removing them flips the sign of the net P&L, while a symmetric trim of the three best and three worst markets leaves $+24.2\%$. The annualized Sharpe ratio over the 44 daily returns is $4.22$ (calendar-day convention, $\sqrt{365}$, used throughout), but its 95% block-bootstrap interval $[-0.89, +7.69]$ includes zero; the maximum drawdown was 15.7% (\$31.47). The win rate of 50.9% counts all unresolved positions as losses; among markets resolved by the end of the window it is 67.4%. (The export's own summary tallies 56 winning and 35 losing settled markets, 61.5%, with 32 still open.) All orders were limit orders at the ask, so slippage beyond the quote was zero by construction; thin depth surfaced as unfilled size (236 of 396 requested trades filled completely), bounding position size rather than per-trade edge; fees were \$26.31 against \$160.67 of net profit. The 160 orders that did not fill completely may be adversely selected; this is an unquantified potential bias. We do not attribute the live return to forecasting accuracy: the forecaster's average Brier score on the traded markets was slightly worse than the market's, and the profit is best described as a concentrated, partly unrealized outcome of large divergence positions that landed (Appendix F.4). The run is an implementation proof that the mechanism survives real frictions, not statistical evidence of risk-adjusted performance. Implementation details, representative trades, and trading logs are in Appendix E.7.

## 5 Conclusion

We investigate the puzzle that forecasters can outperform prediction markets in accuracy yet still lose money, by establishing a robust profitability guarantee that converts an accuracy edge into expected profit, together with an equivalence between strict properness of a scoring rule and the existence of such a guarantee. Key to our result is a "proper" betting strategy for arbitrary prediction markets, which strictly generalizes the canonical betting strategy of AMMs, is essentially the only robustly profitable strategy, and is deeply connected to proper scoring rules. Through this lens, it is unsurprising that proper betting is the only strategy reliably converting accuracy into ROI across thousands of archived AI forecasts. While our live experiments are limited by resource and legal constraints, the month-long pilot on Kalshi (profitable, but with the profit concentrated in a few markets and a Sharpe interval that includes zero) is an implementation proof that the mechanism survives real frictions, not statistical evidence of risk-adjusted performance.

**Limitations.** Our results have limitations on both the theoretical and the empirical side. On the theoretical side, the guarantees concern *expected* profit and require the divergence to offset the liquidity loss (which a small enough rescaling achieves, Remark 4); they assume a differentiable potential on the relative interior (Section 2); a single monotone price-impact map does not capture every order-book feature (separate ladders, discontinuities, path dependence, asynchronous updates) nor how markets adapt once profitable signals are traded repeatedly; and the choice of the proportionality constant and of risk-optimal proper strategies remains open (Remark 5). On the empirical side, all experiments use binary contracts ($K = 2$), with multi-outcome events entering only as families of binary books, so simultaneous execution against a joint $K$-dimensional book is untested. Offline ROIs assume zero price impact beyond the quoted ask and are upper bounds on executable returns (Appendix F.3); the Log and Sph columns evaluate per-contract rules (Appendix E.2); and the persona analysis reports no event-clustered intervals. Personas are synthetic and were tested live in only one persona–rule cell (Conservative, Brier); repeated live runs, a live persona $\times$ rule grid, and live replays of the baselines were infeasible; and the live run is a single, short, tail-driven deployment on one venue, Kalshi (Section 4.3). Open directions include data-adaptive scoring-rule selection, market design at scale, and risk-adjusted position sizing under limited liquidity.

## References


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

Kenneth J. Arrow, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O. Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D. Nelson, et al. *The Promise of Prediction Markets*. Science and Technology Policy Institute, 2008.

Joyce Berg, Robert Forsythe, Forrest Nelson, and Thomas Rietz. Results from a dozen years of election futures markets research. *Handbook of experimental economics results*, 1:742–751, 2008.

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

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

Yiling Chen and Jennifer Wortman Vaughan. A new understanding of prediction markets via no-regret learning. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 189–198, 2010.

Yiling Chen, Stanko Dimitrov, Rahul Sami, Daniel M Reeves, David M Pennock, Robin D Hanson, Lance Fortnow, and Rica Gonen. Gaming prediction markets: Equilibrium strategies with a market maker. *Algorithmica*, 58(4):930–969, 2010.

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

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

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

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

Robin Hanson. Combinatorial information market design. *Information Systems Frontiers*, 5(1): 107–119, 2003.

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

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

Youwon Jang, Joochan Kim, and Byoung-Tak Zhang. 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.

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

Nicolas S Lambert, David M Pennock, and Yoav Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, 2008.

Yingkai Li, Jason D Hartline, Liren Shan, and Yifan Wu. Optimization of scoring rules. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 988–989, 2022.

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

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

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

Hunter Ng, Lin Peng, Yubo Tao, and Dexin Zhou. 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.

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

Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971. ISSN 01621459, 1537274X.

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

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

Qingchuan Yang, Simon Mahns, Sida Li, Anri Gu, Jibang Wu, and Haifeng Xu. Llm-as-a-prophet: Understanding predictive intelligence with prophet arena. 2025. URL https://arxiv.org/abs/2510.17638.

## A Additional 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 and Raftery [2007] formalize strictly proper scoring rules and their relationship to Bregman divergences and convex analysis. Our machinery—proper scoring rules as convex potentials and score gaps as Bregman divergences—is this classical one; our contribution is to carry it to an arbitrary price-impact function and to characterize properness through robust profitability (Theorem 2 and Proposition 2). Subsequent work extends these ideas to more general prediction settings, including decision-theoretic elicitation [Lambert et al., 2008], mechanism design [Chen and Vaughan, 2010], and optimal information acquisition [Li et al., 2022].

**Prediction markets.** Prediction markets themselves are a well-studied mechanism for aggregating dispersed beliefs into a consensus probability [Wolfers and 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 several implementation proposals such as the Logarithmic Market Scoring Rule (LMSR) and its derivatives [Chen and 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 prediction 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 probabilities. Our work is also about betting-strategy design, but differs from Kelly in a key respect: our strategy is oblivious to the ground-truth probabilities and comes with robustness guarantees, whereas the Kelly criterion assumes perfect knowledge of the (usually unobservable) ground-truth probabilities and can perform poorly otherwise (see Example 1).

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 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 $\rho$ and $L_\rho$.

**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 and 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 convex, differentiable *cost function* $C : \mathbb{R}^K \to \mathbb{R}$ (strictly convex on $\mathbf{1}^\perp$: since $\nabla C \in \Delta_{[K]}$, $C(\mathbf{x}+c\mathbf{1}) = C(\mathbf{x})+c$, so every such $C$ is affine along $\mathbf{1}$). 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 cost function $C$ is chosen so that it always satisfies $\nabla C(\mathbf{x}) \in \Delta_{[K]}$ for all $\mathbf{x}$. The marginal cost $\mathbf{q} = \nabla C(\mathbf{x})$ is the *spot price*. 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 $S(\mathbf{p}, y) = G(\mathbf{p}) + \nabla G(\mathbf{p}) \cdot (\mathbf{1}_y - \mathbf{p})$.

**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, $\rho$ is the piecewise-constant function read directly off the order book, and $L_\rho$ captures the depth-induced slippage as the trader walks the book.

# C Omitted proofs

## C.1 Proof of Lemma 1

*Proof of Lemma 1.* By the definition of $D_G(\mathbf{q}, \mathbf{p})$ (used here for any differentiable $G$, where it may be negative),

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

## C.2 Proof of Theorem 2

Throughout, $P$ is the orthogonal projection onto $\mathbf{1}^\perp$; for $\mathbf{v} \in \mathbf{1}^\perp$ one has $\mathbf{s} \cdot \mathbf{v} = (P\mathbf{s}) \cdot \mathbf{v}$, and $\mathbf{p}^* - \mathbf{q} \in \mathbf{1}^\perp$ for $\mathbf{p}^*, \mathbf{q} \in \Delta_K$, so only $P\mathbf{s}$ affects profits. It suffices to treat frictionless markets ($L_\rho = 0$): liquidity loss only lowers profit, so a strategy that is robustly profitable in some market is robustly profitable without friction.

**The favorable set is a half-space cap.** Taking expectations in Lemma 1, $S(\mathbf{p}; \mathbf{p}^*) - S(\mathbf{q}; \mathbf{p}^*) = \mathbf{s}_G(\mathbf{p}, \mathbf{q}) \cdot (\mathbf{p}^* - \mathbf{q}) - D_G(\mathbf{q}, \mathbf{p})$, so the ground truths under which $\mathbf{p}$ performs at least as well as $\mathbf{q}$ are

$$F(\mathbf{p}, \mathbf{q}) = \{\mathbf{p}^* \in \Delta_K : \ \mathbf{g} \cdot (\mathbf{p}^* - \mathbf{q}) \geq D_G(\mathbf{q}, \mathbf{p})\}, \qquad \mathbf{g} := P\mathbf{s}_G(\mathbf{p}, \mathbf{q}).$$

Here $\mathbf{g} \neq \mathbf{0}$ for $\mathbf{p} \neq \mathbf{q}$, because the three-point identity $D_G(\mathbf{q}, \mathbf{p}) + D_G(\mathbf{p}, \mathbf{q}) = \mathbf{s}_G(\mathbf{p}, \mathbf{q}) \cdot (\mathbf{p} - \mathbf{q}) = \mathbf{g} \cdot (\mathbf{p} - \mathbf{q})$ has a positive left-hand side for strictly convex $G$. The same identity and the Cauchy–Schwarz inequality give $D_G(\mathbf{q}, \mathbf{p}) \leq \mathbf{g} \cdot (\mathbf{p} - \mathbf{q}) \leq \|\mathbf{g}\| \|\mathbf{p} - \mathbf{q}\|$, so the bounding hyperplane of $F(\mathbf{p}, \mathbf{q})$ lies at distance

$$d(\mathbf{p}, \mathbf{q}) := D_G(\mathbf{q}, \mathbf{p}) / \|\mathbf{g}\| \leq \|\mathbf{p} - \mathbf{q}\|$$

from $\mathbf{q}$.

**Robust profitability pins down the direction.** Fix an interior $\mathbf{q}$ and let $r := \min_k q_k > 0$; then $\mathbf{q} + \mathbf{v} \in \Delta_K$ for every $\mathbf{v} \in \mathbf{1}^\perp$ with $\|\mathbf{v}\| \leq r$. Let $\mathbf{s}$ be robustly profitable, fix $\mathbf{p} \neq \mathbf{q}$ with $\|\mathbf{p} - \mathbf{q}\| < r$, and write $\mathbf{w} := P\mathbf{s}(\mathbf{p}, \mathbf{q})$, $\hat{\mathbf{g}} := \mathbf{g}/\|\mathbf{g}\|$, $\eta := d(\mathbf{p}, \mathbf{q})/r \leq \|\mathbf{p} - \mathbf{q}\|/r < 1$. For every unit vector $\mathbf{u} \in \mathbf{1}^\perp$ with $\hat{\mathbf{g}} \cdot \mathbf{u} \geq \eta$ the point $\mathbf{p}^* = \mathbf{q} + r\mathbf{u}$ lies in $F(\mathbf{p}, \mathbf{q})$, so robust profitability requires

$$\mathbf{w} \cdot \mathbf{u} > 0 \qquad \text{for every unit } \mathbf{u} \in \mathbf{1}^\perp \text{ with } \hat{\mathbf{g}} \cdot \mathbf{u} \geq \eta. \tag{4}$$

Taking $\mathbf{u} = \hat{\mathbf{g}}$ shows $\mathbf{w} \neq \mathbf{0}$ and $\hat{\mathbf{w}} \cdot \hat{\mathbf{g}} > 0$ for $\hat{\mathbf{w}} := \mathbf{w}/\|\mathbf{w}\|$. Decompose $\hat{\mathbf{w}} = (\hat{\mathbf{w}} \cdot \hat{\mathbf{g}})\hat{\mathbf{g}} + \mathbf{b}$ with $\mathbf{b} \in \mathbf{1}^\perp$, $\mathbf{b} \perp \hat{\mathbf{g}}$. If $\mathbf{b} \neq \mathbf{0}$, the unit vector $\mathbf{u} = \eta\hat{\mathbf{g}} - \sqrt{1 - \eta^2}\, \mathbf{b}/\|\mathbf{b}\|$ satisfies $\hat{\mathbf{g}} \cdot \mathbf{u} = \eta$, and Eq. (4) gives $\eta(\hat{\mathbf{w}} \cdot \hat{\mathbf{g}}) - \sqrt{1 - \eta^2}\, \|\mathbf{b}\| > 0$, i.e.

$$\|\mathbf{b}\| < \frac{\eta}{\sqrt{1 - \eta^2}}, \qquad \text{hence} \qquad \|\hat{\mathbf{w}} - \hat{\mathbf{g}}\|^2 = 2 - 2\sqrt{1 - \|\mathbf{b}\|^2} \leq 2\|\mathbf{b}\|^2 < \frac{2\eta^2}{1 - \eta^2}. \tag{5}$$

(For $K = 2$, $\mathbf{1}^\perp$ is a line, so $\mathbf{b} = \mathbf{0}$ and $\hat{\mathbf{w}} = \hat{\mathbf{g}}$ exactly.) Since $\eta \le \|\mathbf{p} - \mathbf{q}\|/r \to 0$ as $\mathbf{p} \to \mathbf{q}$, the normalized directions of $\mathbf{s}$ and $\mathbf{s}_G$ coincide in the limit along every sequence $\mathbf{p}_t \to \mathbf{q}$, for every interior $\mathbf{q}$: $\mathbf{s}$ is essentially the same as $\mathbf{s}_G$. This proves the first sentence of the theorem, with the explicit rate Eq. (5).

**The contrapositive with strict inequalities.** Let $\mathbf{s}$ be essentially different from $\mathbf{s}_G$. By Definition 3 there are an interior $\mathbf{q}$ and a sequence $\mathbf{p}_t \to \mathbf{q}$ along which either $P\mathbf{s}(\mathbf{p}_t, \mathbf{q}) = \mathbf{0}$ infinitely often or the two directions are defined but do not converge to each other; after passing to a subsequence, either $P\mathbf{s}(\mathbf{p}_t, \mathbf{q}) = \mathbf{0}$ for all $t$, or there is $\epsilon > 0$ with $\|\hat{\mathbf{w}}_t - \hat{\mathbf{g}}_t\| \ge \epsilon$ for all $t$, where $\hat{\mathbf{w}}_t, \hat{\mathbf{g}}_t$ are the directions of $P\mathbf{s}(\mathbf{p}_t, \mathbf{q})$ and $P\mathbf{s}_G(\mathbf{p}_t, \mathbf{q})$. In the first case $\mathbf{s}(\mathbf{p}_t, \mathbf{q})$ earns 0 under every ground truth, in particular under $\mathbf{p}^* = \mathbf{p}_t \in F(\mathbf{p}_t, \mathbf{q})$, where $S(\mathbf{p}_t; \mathbf{p}_t) > S(\mathbf{q}; \mathbf{p}_t)$ by strict properness, which is the non-positive case of the theorem; assume now the second case. Take $t$ so large that $\eta_t := d(\mathbf{p}_t, \mathbf{q})/r < 1/2$, and decompose $\hat{\mathbf{w}}_t = (\hat{\mathbf{w}}_t \cdot \hat{\mathbf{g}}_t)\hat{\mathbf{g}}_t + \mathbf{b}_t$ as above. If $\hat{\mathbf{w}}_t \cdot \hat{\mathbf{g}}_t < 0$, choose $\mathbf{p}^* = \mathbf{q} + r\hat{\mathbf{g}}_t$: the score gap is $\|\mathbf{g}_t\|\, r\,(1 - \eta_t) > 0$ and the expected profit is $r\|\mathbf{w}_t\|\,(\hat{\mathbf{w}}_t \cdot \hat{\mathbf{g}}_t) < 0$. Otherwise $\|\hat{\mathbf{w}}_t - \hat{\mathbf{g}}_t\| \ge \epsilon$ forces $\hat{\mathbf{w}}_t \cdot \hat{\mathbf{g}}_t \le 1 - \epsilon^2/2$ and hence $\|\mathbf{b}_t\| \ge \epsilon' := \epsilon\sqrt{1 - \epsilon^2/4} > 0$; choose $\mathbf{p}^* = \mathbf{q} + r\mathbf{u}$ with the unit vector $\mathbf{u} = 2\eta_t\hat{\mathbf{g}}_t - \sqrt{1 - 4\eta_t^2}\,\mathbf{b}_t/\|\mathbf{b}_t\|$: the score gap is $\|\mathbf{g}_t\|\, r\,\eta_t > 0$, while the expected profit is $r\|\mathbf{w}_t\|\big(2\eta_t(\hat{\mathbf{w}}_t \cdot \hat{\mathbf{g}}_t) - \sqrt{1 - 4\eta_t^2}\,\|\mathbf{b}_t\|\big) \le r\|\mathbf{w}_t\|\big(2\eta_t - \sqrt{1 - 4\eta_t^2}\,\epsilon'\big) < 0$ for all large $t$. In every case $(\mathbf{p}_t, \mathbf{p}^*, \mathbf{q})$ witnesses the theorem. □

**Remark 2** (On the equivalence)**.** *Since $G$ lives on the simplex, $\nabla G$ and hence $\mathbf{s}_G$ are determined only up to a multiple of $\mathbf{1}$ (Table 1 fixes it by extending $G$ off the simplex); this is the constant shift of Definition 3, payoff-irrelevant in a frictionless market, and the reason the proof works with $P\mathbf{s}$. "Essentially the same" is an asymptotic equivalence of directions, broader than exact rescaling or shifting: a perturbation of $\mathbf{s}_G$ whose angle to $\mathbf{s}_G$ vanishes as $\mathbf{p} \to \mathbf{q}$ is essentially the same as $\mathbf{s}_G$ whether or not it is robustly profitable. Theorem 2 is therefore a uniqueness statement, not a characterization.*

### C.3 Proof of Proposition 2

The "only if" direction is Theorem 1 with $L_\rho = 0$: for $\mathbf{p} \ne \mathbf{q}$ and $S(\mathbf{p}; \mathbf{p}^*) \ge S(\mathbf{q}; \mathbf{p}^*)$ the expected profit of $\mathbf{s}_G(\mathbf{p}, \mathbf{q})$ is at least $D_G(\mathbf{q}, \mathbf{p}) > 0$. For the "if" direction suppose that $S$ is not strictly proper. Then there are $\mathbf{p} \ne \mathbf{q}$ in $\Delta_K$ with $S(\mathbf{p}; \mathbf{q}) \ge S(\mathbf{q}; \mathbf{q})$. Take the ground truth $\mathbf{p}^* := \mathbf{q}$. The triple satisfies $S(\mathbf{p}; \mathbf{p}^*) \ge S(\mathbf{q}; \mathbf{p}^*)$, yet every position $\mathbf{s}$ has expected profit $\mathbf{s} \cdot (\mathbf{p}^* - \mathbf{q}) = 0$ in a frictionless market. Hence no betting strategy guarantees positive profit, which is the contrapositive of the claim. The argument uses no structure of $S$ beyond its definition. □

**Remark 3** (The weak inequality is essential)**.** *If positive profit were required only when $S(\mathbf{p}; \mathbf{p}^*) > S(\mathbf{q}; \mathbf{p}^*)$, the "if" direction would fail: the constant rule $S \equiv 0$ is proper but not strictly proper, and it never makes the strict requirement bite, so every strategy would qualify. A strategy that is robust in the sense of Proposition 2 must in particular survive the ground truth $\mathbf{p}^* = \mathbf{q}$, a market that is exactly right, and strict properness is precisely the statement that no $\mathbf{p} \ne \mathbf{q}$ then scores as well as $\mathbf{q}$.*

### C.4 Proof of Corollary 1

We only need to show $D_G(\mathbf{q}, \mathbf{p}) = L_\rho(\mathbf{s}^*; \mathbf{q})$ in any AMM governed by a cost function $C(\mathbf{x})$ (strictly convex on $\mathbf{1}^\perp$), whose convex conjugate is $G(\mathbf{p})$. The rest of the statement then follows from Theorem 1.

Let $\mathbf{s}^* = \nabla G(\mathbf{p}) - \nabla G(\mathbf{q})$ denote the proper bet. We compute

$$\begin{aligned} L_\rho(\mathbf{s}^*; \mathbf{q}) &= C(\mathbf{s}^* + \mathbf{x}) - C(\mathbf{x}) - \nabla C(\mathbf{x}) \cdot \mathbf{s}^* \\ &= D_C(\mathbf{s}^* + \mathbf{x}, \mathbf{x}) \\ &= D_G(\nabla C(\mathbf{x}), \nabla C(\mathbf{s}^* + \mathbf{x})) \;=\; D_G(\mathbf{q}, \mathbf{p}), \end{aligned}$$

where the third equality is the conjugate-duality identity $D_C(\mathbf{x}', \mathbf{x}) = D_G(\nabla C(\mathbf{x}), \nabla C(\mathbf{x}'))$ (note the order of the arguments) and the last one uses $\nabla C(\mathbf{x} + \mathbf{s}^*) = \mathbf{p}$. To verify the identity, write $\mathbf{q} = \nabla C(\mathbf{x})$ and $\mathbf{p} = \nabla C(\mathbf{x}')$, so that $\mathbf{x} = \nabla G(\mathbf{q})$, $\mathbf{x}' = \nabla G(\mathbf{p})$, $C(\mathbf{x}) = \mathbf{q} \cdot \mathbf{x} - G(\mathbf{q})$ and $C(\mathbf{x}') = \mathbf{p} \cdot \mathbf{x}' - G(\mathbf{p})$; then $D_C(\mathbf{x}', \mathbf{x}) = \mathbf{p} \cdot \mathbf{x}' - G(\mathbf{p}) - \mathbf{q} \cdot \mathbf{x} + G(\mathbf{q}) - \mathbf{q} \cdot (\mathbf{x}' - \mathbf{x}) = G(\mathbf{q}) - G(\mathbf{p}) - \nabla G(\mathbf{p}) \cdot (\mathbf{q} - \mathbf{p}) = D_G(\mathbf{q}, \mathbf{p})$. (For LMSR this is $\mathrm{KL}(\mathbf{q} \,\|\, \mathbf{p})$, not $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})$.)

### C.5 The many-outcome Kelly criterion in Example 1

In Example 1, the Kelly criterion spends $p_i$ to purchase $p_i/q_i$ shares of each contract $i$. This is the general solution of the many-outcome Kelly criterion, and the familiar two-outcome formula of spending $\frac{p_1-q_1}{1-q_1}$ to bet on outcome 1 (shorting it if negative) is a special case. To see this, observe that, given $\mathbf{q} \in \Delta_K$, betting $\frac{p_1-q_1}{1-q_1} = p_1 - q_1\frac{1-p_1}{1-q_1}$ on outcome 1 and 0 on outcome 2, and additionally buying $\frac{1-p_1}{1-q_1}$ shares of both outcome 1 at price $q_1$ and outcome 2 at price $1-q_1$ (a riskless position), together yield total bet amounts $p_1$ and $p_2$ on outcomes 1 and 2, respectively.

## D Extensions of Theorem 1

### D.1 Beyond the probability simplex

The decomposition of Theorem 1 (Lemma 1, which needs no convexity) holds even when $\mathbf{q} \notin \Delta_K$. As an application, suppose $Q := \sum_{k=1}^K q_k < 1$, so that there is a trivial arbitrage: buying $\mathbf{s} = \mathbf{1}$ costs $Q\,(< 1)$ and returns 1 deterministically (assuming little liquidity loss). The decomposition captures this situation as a special case. Choosing $G(\mathbf{q}) = (\sum_{k=1}^K q_k)^2$, the proper bet is $\mathbf{s}_G = 2[\sum_{k=1}^K (p_k - q_k)] \cdot \mathbf{1}$, which yields the deterministic profit $2(1-Q)^2$ for any $\mathbf{p}^*$. One can further verify that both the Bregman divergence and the score gap equal $(1-Q)^2$, summing to the $2(1-Q)^2$ profit.

### D.2 Feasible scale and realized-return concentration

**Remark 4** (Feasible scale)**.** *Let $\mathbf{s}^* = \mathbf{s}_G(\mathbf{p}, \mathbf{q})$ with $\mathbf{p} \neq \mathbf{q}$ and $G$ strictly convex, and consider the rescaled bet $\alpha \cdot \mathbf{s}^*$, $\alpha > 0$. By Lemma 1, its expected profit is $\pi(\alpha \cdot \mathbf{s}^*, \mathbf{p}^*) = \alpha\,[S(\mathbf{p}; \mathbf{p}^*) - S(\mathbf{q}; \mathbf{p}^*) + D_G(\mathbf{q}, \mathbf{p})] - L_\rho(\alpha \cdot \mathbf{s}^*; \mathbf{q})$, and $L_\rho(\alpha \cdot \mathbf{s}^*; \mathbf{q}) = \alpha \int_0^1 (\rho(t\alpha \cdot \mathbf{s}^*) - \mathbf{q}) \cdot \mathbf{s}^*\, dt$. If $\rho$ is continuous at $\mathbf{0}$—or, as in an order book with positive depth at the quoted price, $\rho(\mathbf{s}) = \mathbf{q}$ for all sufficiently small $\mathbf{s}$—then $L_\rho(\alpha \cdot \mathbf{s}^*; \mathbf{q})/\alpha \to 0$ as $\alpha \downarrow 0$. Since $D_G(\mathbf{q}, \mathbf{p}) > 0$, there is $\bar{\alpha}(\mathbf{p}, \mathbf{q}) > 0$, which does not depend on $\mathbf{p}^*$, such that $L_\rho(\alpha \cdot \mathbf{s}^*; \mathbf{q}) \leq \alpha D_G(\mathbf{q}, \mathbf{p})$ for all $\alpha \in (0, \bar{\alpha}]$; every such $\alpha \cdot \mathbf{s}^*$ is robustly profitable in the sense of Theorem 1. The price of feasibility is a proportionally smaller absolute profit.*

**Remark 5** (High-probability bound on realized return)**.** *Consider $n$ events with forecasts $\mathbf{p}_i$, prices $\mathbf{q}_i$, positions $\mathbf{s}_i = c\,\mathbf{s}_G(\mathbf{p}_i, \mathbf{q}_i)$ fixed before the outcomes, and independent outcomes $y_i \sim \mathbf{p}_i^*$. The realized profit $X_i = \mathbf{s}_i \cdot (\mathbf{1}_{y_i} - \mathbf{q}_i) - L_\rho(\mathbf{s}_i; \mathbf{q}_i)$ has mean $\pi_i = c\,[S(\mathbf{p}_i; \mathbf{p}_i^*) - S(\mathbf{q}_i; \mathbf{p}_i^*) + D_G(\mathbf{q}_i, \mathbf{p}_i)] - L_\rho(\mathbf{s}_i; \mathbf{q}_i)$ (Lemma 1) and lies in an interval of length $r_i = \max_k s_{i,k} - \min_k s_{i,k}$. By Hoeffding's maximal inequality (with events indexed in resolution order), for every $t > 0$,*

$$\Pr\left[\min_{m \leq n} \textstyle\sum_{i \leq m} (X_i - \pi_i) \leq -t\right] \;\leq\; \exp\Big(-2t^2 / \textstyle\sum_{i \leq n} r_i^2\Big).$$

*Hence, with probability at least $1-\delta$, the realized total profit is at least $\sum_i \pi_i - \sqrt{\frac{1}{2}\sum_i r_i^2 \ln(1/\delta)}$; and if every event satisfies the conditions of Theorem 1 (so $\pi_i > 0$), the cumulative realized profit never falls below $-t$ with probability at least $1 - \exp\big(-2t^2/\sum_i r_i^2\big)$, a bound on the probability of a given loss from the initial capital at fixed exposure. Since $r_i$ scales with $c$, lowering the proportionality constant tightens the bound while preserving the sign of the expected guarantee. Positions are sized by disagreement rather than as fractions of a compounding bankroll, so losses are bounded by the capital staked. A theory of variance- or drawdown-optimal proper betting is beyond the scope of this paper.*

### D.3 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 1 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 a separate 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:

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

where $S(p_k, y_k)$ is the score for the binary event *YES/NO* $k$, in reduced form with 1-D parameter $p_k$. Let $G : [0, 1] \to \mathbb{R}$ be its associated convex potential function; for this one-dimensional $G$, $\nabla G$ below denotes the ordinary derivative $G'$ (written $g'$ in Appendix E.2).

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

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

To see this is a valid betting strategy under a nonzero bid-ask spread, we observe that $\nabla G(p_k) - \nabla G(q_k^+) \geq 0$ when $p_k > q_k^+$ due to convexity of $G$ and $\nabla G(p_k) - \nabla G(1 - q_k^-) < 0$ when $p_k < 1 - q_k^-$ for similar reasons. Hence, $s_k$ never needs to buy the other side (which may have an inconsistent price due to the nonzero bid-ask spread). 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 always have $s_k = \nabla G(p_k) - \nabla G(\tilde{q}_k)$, 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 3** (Profitability under bid-ask spread (omitting liquidity loss))**.** *Let $\mathbf{s}, \tilde{\mathbf{q}}$ be as defined above and $S(\tilde{\mathbf{q}}, y) = \sum_{k=1}^{K} S(\tilde{q}_k, y_k)$ be the bid-ask-aware market score. Then the expected profit of the bid-ask-aware proper bet admits the following decomposition*

$$\mathbb{E}_{y \sim \mathbf{p}^*} \left[ \mathbf{s} \cdot (\mathbf{1}_y - \tilde{\mathbf{q}}) \right] \;=\; \underbrace{\mathbb{E}_{y \sim \mathbf{p}^*} \left[ S(\mathbf{p}, y) - S(\tilde{\mathbf{q}}, y) \right]}_{\textit{score gap}} \;+\; \underbrace{\sum_{k=1}^{K} D_G(\tilde{q}_k, p_k)}_{\textit{Bregman bonus}} \tag{6}$$

*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 1:

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

Summing over $k \in [K]$ and taking expectation under $y \sim \mathbf{p}^*$ yields Eq. (6). □

**Remark 6** (Correlated outcomes)**.** *Corollary 3 holds even for outcomes $\{1, \cdots, K\}$ that need not be disjoint but can be arbitrarily correlated. In this case, we simply treat each outcome $k$ as a separate (though possibly correlated) binary event.*

**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

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

where the spread cost vanishes at zero spread ($q_k^+ + q_k^- = 1$). Its sign is not fixed outcome by outcome: on a YES leg ($\tilde{q}_k = q_k^+ \geq \bar{q}_k$) and $y_k = 0$, $S(\cdot, 0)$ is non-increasing, so $S(\bar{q}_k, 0) \geq S(q_k^+, 0)$. In expectation, $x \mapsto S(x; p_k^*)$ is unimodal with mode $p_k^*$ (its derivative is $G''(x)(p_k^* - x)$), so on a traded coordinate the expected spread cost is non-positive whenever the ground truth lies beyond the traded ask ($p_k^* \geq q_k^+$ on a YES leg, $p_k^* \leq 1 - q_k^-$ on a NO leg); on an untraded coordinate ($\tilde{q}_k = p_k$) it equals minus the forgone edge $S(p_k; p_k^*) - S(\bar{q}_k; p_k^*)$. Where these signs hold, the bid-ask-aware condition is stronger than the spread-free one by the spread cost, and the two coincide at zero spread.

### D.4 Proper betting for long time horizons

For events that take days or weeks to resolve, both forecaster and market predictions evolve over time. A forecaster can make a different prediction $\mathbf{p}^t$ at each time $t = 1, \cdots, T$ and trade on $\mathbf{p}^t$ as well as the market price $\mathbf{q}^t$ at that time. We extend Theorem 1 to two natural multi-period strategies, each repeating the per-round proper bet but differing in what they hold and what accuracy edge they require. In both corollaries below, we omit liquidity loss for simplicity.

**Corollary 4** (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. The cumulative expected profit at resolution then decomposes as*

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

**Corollary 5** (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* $\mathbf{x}^0 := \mathbf{0}$*. The cumulative mark-to-market return then decomposes as*

$$\sum_{t=1}^{T} \mathbf{x}^t \cdot (\mathbf{q}^{t+1} - \mathbf{q}^t) \;=\; \sum_{t=1}^{T} \left[S(\mathbf{p}^t; \mathbf{q}^{t+1}) - S(\mathbf{q}^t; \mathbf{q}^{t+1})\right] \;+\; \sum_{t=1}^{T} D_G(\mathbf{q}^t, \mathbf{p}^t) \tag{8}$$

*Moreover, the accumulated profit of executing trades* $\mathbf{s}^t$ *at prices* $\mathbf{q}^t$ *and settling at* $\mathbf{q}^{T+1}$ *can be equivalently expressed in the following 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 two corollaries differ in what plays the role of the "ground truth" in the proper-bet decomposition: Corollary 4 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 5 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 4, apply Lemma 1 at each round with realized outcome $y$:

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

Summing over $t$ and taking expectation under $y \sim \mathbf{p}^*$ gives Eq. (7). Since the Bregman sum is non-negative, the total is positive whenever the score-gap sum is.

For Corollary 5, apply Lemma 1 (in expectation form) at each round with substitutions $\mathbf{p}^* \to \mathbf{q}^{t+1}$, $\mathbf{p} \to \mathbf{p}^t$, $\mathbf{q} \to \mathbf{q}^t$:

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

Summing gives Eq. (8). 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

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

Positivity under a positive score-gap sum follows as for the fundamental strategy. □

**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 bootstrapped 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)** |

# E Additional experiments

## E.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, although the Brittle group has a lower ECE and a higher aggregate Brier accuracy than the Dispersed group, it is also less profitable. Aggregate accuracy and calibration can therefore coexist with systematic capital loss, because of the Bregman divergence term.

## E.2 Implementation details of betting strategies

**Setup.** We consider a collection of $n$ markets with forecast–price pairs $\{p_i, q_i\}_{i\in[n]}$. *Units.* A betting strategy is defined by positions $\{w_i\}_{i\in[n]}$ measured in *shares*, which encode both direction and size: if $w_i > 0$ we buy $w_i$ YES shares of market $i$ at the YES ask $q_i^+$; if $w_i < 0$ we buy $|w_i|$ NO shares at the NO ask $q_i^-$. Writing $q_i$ for the price of the contract actually bought, positions are scaled so that the total cost staked equals the unit budget, $\sum_{i\in[n]} |w_i|\, q_i = 1$, and ROI is realized profit divided by this total cost. Lemma 1 applies to $w_i$ directly because it is stated in shares; all table captions use this convention.

*Two-contract (per-contract) construction.* Each binary market is treated as two one-sided markets in the sense of Appendix D.3: a YES market with price $q_i^+$ and prediction $p_i$, and a NO market with price $q_i^-$ and prediction $1 - p_i$. In 99.96% of the markets $q_i^+ + q_i^- > 1$, so a sell signal on a contract we do not hold cannot execute; the implementation executes buy signals only. The two buy signals are mutually exclusive ($p_i > q_i^+$ forces $1 - p_i < q_i^-$), so at most one leg is held per market, and no trade is placed when $p_i \in [1 - q_i^-,\, q_i^+]$. A NO bet is sized by the NO contract's own coordinate, not by the YES coordinate with a sign rule. When $q_i^+ + q_i^- = 1$, executing the signed signals of both legs gives exactly the netted binary bet (e.g., $\mathrm{logit}(p_i) - \mathrm{logit}(q_i)$ for Log); otherwise the buys-only rule differs from it.

**Proper Betting Strategies.** On the traded contract, with prediction $\pi$ and price $c$ (i.e., $(\pi, c) = (p_i, q_i^+)$ for a YES bet and $(1-p_i, q_i^-)$ for a NO bet), the per-contract proper bet buys $|w_i| \propto g'(\pi) - g'(c)$ shares for a convex potential $g$ on $[0,1]$:

1. **Brier:** $g(x) = x^2$, so $|w_i| \propto \pi - c$ (linear in the margin).
2. **Logarithmic:** $g(x) = x\log x - x$, so $|w_i| \propto \log\pi - \log c$.
3. **Spherical:** $g(x) = \frac{1}{2}\|(x, 1-x)\| + \frac{1}{2\sqrt{2}}\operatorname{arsinh}(2x-1)$, so $|w_i| \propto \frac{\pi}{\|(\pi,1-\pi)\|} - \frac{c}{\|(c,1-c)\|}$.

*Induced scores.* Each $g$ induces the per-contract score $S_g(x,y) = g(x) + g'(x)(y-x)$ for a contract paying $y \in \{0,1\}$: $S_g(x,y) = 2xy - x^2$ (Brier), $S_g(x,y) = y\log x - x$ (Log), and $S_g(x,y) = g(x) + (y-x)\,x/\|(x,1-x)\|$ (Spherical), and Lemma 1 holds contract by contract: $(g'(\pi) - g'(c))(y-c) = S_g(\pi,y) - S_g(c,y) + D_g(c,\pi)$. The executed buys-only strategy is therefore scored by $T(p,y) = \mathbf{1}[p > q^+]\,[S_g(p,y) - S_g(q^+,y)] + \mathbf{1}[1-p > q^-]\,[S_g(1-p,1-y) - S_g(q^-,1-y)]$, which is proper (though not strictly proper, since every report inside the spread band scores zero); the Log and Sph score gaps $\Delta S$ in Tables 3, 6 and 7 are gaps under these per-contract scores. For Log, scoring *both* legs would give $y\log p + (1-y)\log(1-p) - 1$, the standard log score up to a constant; the buys-only rule scores only the traded leg, so the Log and Sph columns are not comparisons of the standard (netted) log and spherical rules. For Brier the per-contract and netted bets coincide up to a factor of 2, which the budget normalization removes.

**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:* $w_i \propto \operatorname{sign}(p_i - q_i), \forall i \in [n]$. 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:
   $$w_{i^\star(k)} = \frac{\operatorname{sign}(p_{i^\star(k)} - q_{i^\star(k)})}{m}, \quad w_i = 0, \quad \forall i \notin \{i^\star(k)\}_{k=1}^m.$$
2. **Inverse-Margin:** $w_i \propto \frac{\operatorname{sign}(p_i - q_i)}{|p_i - q_i|}, \forall i \in [n]$. 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 $f_i = \frac{|p_i - q_i|}{1-q_i}$. Each contract's fraction is at most 1 (for a NO bet the side-specific fraction is $(q_i - p_i)/q_i$; see *Side rule* below); leverage enters only when the fractions of the contracts of one event are summed (see *Leverage* below). *Side rule.* The bet is placed on YES when $p_i > q_i$ and on NO when $p_i < q_i$. In signed form, $f_i^* = (p_i - q_i)/(1-q_i)$, where a negative value is a short-YES notional of $|f_i^*|$ (i.e., $|f_i^*|/q_i$ shares); this is the same trade as spending the side-specific fraction $(q_i - p_i)/q_i$ of wealth on NO shares at $1-q_i$. The YES fraction is $(p_i - q_i)/(1-q_i)$. Both side-specific fractions are at most 1, so a value $|f_i^*| > 1$ is a notional, not leverage. With separate books, the NO fraction uses the NO ask: $(1 - p_i - q_i^-)/(1 - q_i^-)$. *Leverage.* Each contract's fraction is at most 1, but the implementation

**Table 5:** Pairwise comparison of the Proper Betting (Brier) strategy against each of the five baselines on the expanded dataset (conditioned cut, Section 4.1), with each column after $\Delta S$ reporting the difference in ROI (%; realized profit over total cost staked $\sum_i |w_i| q_i$) between Proper Betting and the corresponding baseline. Confidence intervals are paired, event-clustered 95% bootstrap intervals; bold indicates differences where the CI excludes zero. $\Delta S$ is in Brier units, unlike Tables 3 and 6. Kelly compounds one bankroll; Kelly-Alike is its fresh-budget analogue. Kelly entries reflect uncapped leverage: per-outcome Kelly fractions are summed within an event without a joint budget constraint, so the stake on an event can exceed the bankroll (on the submitted Prophet Arena data, up to $34\times$ on one event; Appendices E.2 and F.1).

| Model | $\Delta S$ | − Max (Mkt) | − Max (Grp) | − Inv-Margin | − Kelly-Alike | − Kelly |
|---|---|---|---|---|---|---|
| Gemini 3.1 Pro (B) | **+0.0076** [+0.0014, +0.0136] | **+30.3** [+20.7, +39.9] | **+30.1** [+15.5, +42.9] | **+39.2** [+27.7, +50.6] | **+24.3** [+16.1, +32.8] | **+138.9** [+127.3, +150.4] |
| Claude Fable 5 | **+0.0039** [+0.0008, +0.0073] | **+26.8** [+15.3, +40.0] | **+48.3** [+34.2, +63.1] | **+33.6** [+20.0, +48.9] | **+16.8** [+7.3, +27.4] | +129.0 [$-1.3 \times 10^9$, +141.8] |
| Gemini 3.1 Pro (A) | **+0.0039** [+0.0016, +0.0063] | **+25.9** [+20.4, +31.5] | **+36.6** [+28.3, +44.5] | **+29.9** [+23.5, +36.5] | **+16.3** [+11.7, +21.0] | **+127.9** [+121.5, +134.6] |
| GPT-5.6 Sol | **+0.0033** [+0.0003, +0.0067] | **+22.4** [+12.2, +35.4] | **+39.9** [+24.3, +55.7] | **+28.5** [+15.7, +43.7] | **+15.7** [+6.0, +27.1] | -1220.4 [$-2.2 \times 10^8$, +122.5] |
| Claude Sonnet 5 | **+0.0028** [+0.0002, +0.0057] | **+20.1** [+10.3, +32.1] | **+58.8** [+45.0, +73.2] | **+25.7** [+13.4, +40.1] | **+8.3** [+1.3, +16.9] | +125.9 [-5766.9, +133.1] |
| GPT-5.5 (B) | -0.0059 [-0.0129, +0.0009] | +13.0 [-1.2, +29.3] | +20.3 [-0.7, +41.1] | **+18.6** [+0.1, +38.9] | +8.6 [-2.7, +21.4] | **+116.8** [+99.8, +136.0] |
| GPT-5.5 (C) | -0.0056 [-0.0101, -0.0012] | +8.8 [-1.5, +18.7] | +3.6 [-28.4, +27.3] | +7.0 [-5.5, +19.4] | +2.0 [-6.3, +10.0] | **+107.3** [+94.9, +119.2] |
| Gemini 3.5 Flash | -0.0188 [-0.0299, -0.0081] | -2.4 [-11.9, +6.8] | +1.8 [-19.5, +19.3] | +1.3 [-10.8, +13.0] | +0.1 [-7.5, +7.6] | **+105.9** [+94.2, +117.7] |
| Qwen 3.5+ (B) | -0.0342 [-0.0415, -0.0272] | **+8.3** [+0.6, +16.0] | +9.6 [-7.7, +24.2] | +7.7 [-1.9, +17.4] | -0.1 [-5.2, +5.2] | **+105.6** [+96.4, +114.6] |
| Kimi K3 | -0.0030 [-0.0065, +0.0004] | +1.9 [-10.7, +15.5] | **+40.3** [+24.2, +56.4] | +4.5 [-12.6, +22.3] | -1.9 [-10.9, +7.9] | **+104.8** [+88.8, +121.6] |
| GPT-5.4 | -0.0203 [-0.0258, -0.0159] | +2.3 [-2.9, +7.6] | **+12.5** [+1.9, +22.5] | +5.1 [-2.5, +13.1] | -2.5 [-5.6, +0.7] | **+99.8** [+92.5, +107.4] |
| GPT-5.5 (D) | -0.0065 [-0.0081, -0.0050] | -0.3 [-3.8, +3.1] | **+15.7** [+9.4, +21.7] | +1.3 [-3.5, +6.3] | -3.1 [-6.2, -0.0] | **+99.8** [+95.2, +104.2] |
| Muse Spark 1.1 | -0.0133 [-0.0203, -0.0066] | -4.2 [-18.0, +11.4] | +16.5 [-3.2, +35.5] | -1.9 [-19.2, +16.1] | -2.3 [-11.1, +7.9] | **+98.9** [+82.6, +116.8] |
| GLM 5.2 | -0.0194 [-0.0248, -0.0139] | -2.3 [-10.0, +5.6] | **+22.7** [+9.2, +34.9] | -4.3 [-14.6, +6.3] | -3.0 [-7.1, +1.4] | **+98.6** [+88.8, +108.4] |
| Claude Sonnet 4.6 (A) | -0.0292 [-0.0353, -0.0233] | -1.2 [-7.9, +5.5] | +7.6 [-5.8, +19.7] | -1.1 [-10.9, +8.6] | -2.8 [-6.5, +1.2] | **+94.9** [+86.9, +103.0] |
| Grok 4.5 | -0.0115 [-0.0164, -0.0067] | -4.6 [-15.5, +8.3] | **+24.0** [+7.8, +40.0] | -4.4 [-18.6, +11.6] | -5.1 [-11.9, +2.8] | **+94.5** [+80.9, +110.0] |
| Claude Sonnet 4.6 (B) | -0.0236 [-0.0346, -0.0128] | -0.9 [-13.8, +13.2] | -16.0 [-63.5, +19.5] | -10.5 [-37.6, +12.6] | -2.4 [-9.2, +6.1] | **+94.2** [+78.5, +110.9] |
| Gemma 4 31B | -0.0339 [-0.0410, -0.0268] | -3.3 [-9.3, +2.8] | +10.1 [-0.6, +19.9] | -5.8 [-13.7, +2.4] | **-4.5** [-8.5, -0.4] | **+93.2** [+85.4, +100.8] |

sums the per-outcome fractions of the contracts of one event without a joint budget constraint, so the total stake on an event can exceed the bankroll (on the submitted Prophet Arena data, up to $34\times$ on a single event). The Kelly columns of Tables 2 and 5 therefore reflect uncapped leverage; the capped variant for the submitted data is reported in Appendix F.1. *Compounding.* Kelly is applied sequentially to one compounding bankroll (compounding is intrinsic to the log-growth criterion), whereas every other strategy uses a fresh unit budget. Kelly-Alike is the de-compounded analogue under identical fresh-budget accounting and is the like-for-like comparison; the Kelly column shows what the criterion itself does under model miscalibration.

Table 5 reports the full comparison of evaluated models across all baseline betting strategies and the proposed Proper (Brier) strategy. The Kelly criterion exhibits particularly high variance in ROI because it compounds the bankroll while effectively treating the forecaster's estimated disagreement with the market as the true edge. When this edge is poorly estimated, Kelly can substantially oversize individual positions. A single sufficiently large loss can therefore deplete the bankroll, causing compounded returns to collapse to $-100\%$ for many models.

### E.3 Empirical decomposition

Table 6 reports the full decomposition across all models and all evaluated proper betting strategies. Divergence varies substantially across both models and rules, and often offsets large negative score gaps, meaning less accurate models can still yield positive ROI. 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.

### E.4 Synthetic persona generation

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.

**Table 6: ROI decomposition on the expanded dataset (conditioned cut, Section 4.1)**, where $\text{ROI} = \Delta S + D$ for each model and proper scoring rule. Bold denotes significantly positive score gaps (95% CI above zero), and bold italic denotes significantly positive ROI. Terms are divided by total cost staked; Log/Sph are the per-contract rules of Appendix E.2.

| Model | Brier $\Delta S$ | Brier $D$ | Brier ROI | Log $\Delta S$ | Log $D$ | Log ROI | Sph $\Delta S$ | Sph $D$ | Sph ROI |
|---|---|---|---|---|---|---|---|---|---|
| Gemini 3.1 Pro (B) | **+11.7** [+2.4, +21.1] | +27.2 [+22.7, +32.1] | ***+38.9*** ***[+27.4, +50.9]*** | **+7.5** [+0.8, +14.0] | +20.9 [+18.2, +23.8] | ***+28.4*** ***[+20.4, +36.2]*** | **+11.3** [+3.5, +19.5] | +21.9 [+18.3, +25.8] | ***+33.2*** ***[+23.9, +42.9]*** |
| Claude Fable 5 | **+12.3** [+3.2, +22.0] | +16.7 [+11.8, +22.2] | ***+29.0*** ***[+16.7, +42.7]*** | **+8.6** [+1.3, +15.6] | +12.9 [+9.8, +16.5] | ***+21.5*** ***[+12.2, +30.6]*** | **+10.2** [+1.4, +18.6] | +11.9 [+8.5, +15.9] | ***+22.1*** ***[+11.6, +33.0]*** |
| Gemini 3.1 Pro (A) | **+8.7** [+3.3, +13.7] | +19.2 [+17.0, +21.7] | ***+27.9*** ***[+21.1, +34.3]*** | **+4.8** [+0.8, +8.5] | +16.4 [+14.7, +18.2] | ***+21.2*** ***[+16.4, +25.8]*** | **+6.5** [+2.1, +10.8] | +13.6 [+12.1, +15.4] | ***+20.1*** ***[+14.9, +25.7]*** |
| GPT-5.6 Sol | **+10.9** [+1.0, +21.7] | +15.9 [+12.1, +21.9] | ***+26.7*** ***[+14.4, +41.4]*** | **+10.7** [+3.0, +18.5] | +12.6 [+10.1, +16.1] | ***+23.3*** ***[+14.5, +32.4]*** | **+12.1** [+3.2, +21.4] | +11.4 [+8.5, +15.7] | ***+23.5*** ***[+13.0, +35.8]*** |
| Claude Sonnet 5 | **+10.2** [+0.2, +19.9] | +15.8 [+11.8, +20.9] | ***+25.9*** ***[+13.4, +38.9]*** | **+9.1** [+1.9, +16.4] | +12.4 [+9.9, +15.6] | ***+21.5*** ***[+13.0, +30.7]*** | **+12.3** [+2.9, +21.7] | +12.4 [+9.5, +16.4] | ***+24.7*** ***[+13.9, +36.4]*** |
| GPT-5.5 (B) | -15.4 [-29.6, -0.4] | +32.2 [+25.8, +39.5] | ***+16.8*** ***[+2.3, +33.3]*** | -9.5 [-20.7, +0.5] | +21.2 [+17.9, +24.9] | ***+11.7*** ***[+0.3, +22.2]*** | -11.5 [-23.0, -0.7] | +24.4 [+19.4, +30.1] | ***+12.9*** ***[+0.9, +25.0]*** |
| GPT-5.5 (C) | -15.6 [-26.3, -5.6] | +22.9 [+18.6, +27.3] | +7.3 [-3.3, +17.3] | -13.1 [-21.4, -5.5] | +18.4 [+15.2, +21.6] | +5.2 [-3.8, +13.9] | -14.6 [-23.4, -6.2] | +16.5 [+13.4, +20.1] | +1.9 [-6.6, +10.7] |
| Gemini 3.5 Flash | -26.3 [-39.9, -14.1] | +32.2 [+26.1, +39.2] | +5.9 [-4.4, +15.8] | -17.4 [-26.8, -8.2] | +25.5 [+21.8, +29.8] | +8.1 [-0.5, +16.4] | -22.5 [-35.4, -10.3] | +27.4 [+21.8, +33.5] | +4.9 [-5.0, +14.8] |
| Qwen 3.5+ (B) | -58.4 [-68.9, -48.4] | +64.0 [+57.7, +70.8] | +5.6 [-2.1, +13.3] | -33.1 [-39.6, -27.0] | +38.9 [+35.6, +41.8] | +5.8 [-0.6, +11.7] | -48.6 [-57.5, -40.5] | +48.9 [+43.5, +54.4] | +0.3 [-6.4, +6.5] |
| Kimi K3 | -15.0 [-29.1, -0.7] | +19.9 [+15.2, +25.1] | +4.8 [-8.2, +18.4] | -7.4 [-18.9, +3.7] | +13.7 [+11.7, +16.0] | +6.4 [-4.6, +17.4] | -10.4 [-22.4, +1.5] | +14.2 [+10.4, +18.7] | +3.8 [-7.3, +15.3] |
| GPT-5.4 | -65.0 [-82.5, -51.3] | +64.8 [+55.1, +79.3] | -0.2 [-7.8, +7.3] | -32.0 [-39.2, -24.8] | +33.2 [+30.9, +35.7] | +1.2 [-5.4, +8.0] | -50.8 [-63.9, -39.1] | +50.4 [+42.3, +60.6] | -0.4 [-7.0, +6.8] |
| GPT-5.5 (D) | -23.8 [-28.9, -18.4] | +23.6 [+21.4, +26.1] | -0.2 [-4.6, +4.4] | -15.1 [-19.4, -10.7] | +16.6 [+15.6, +17.7] | +1.5 [-2.5, +5.7] | -18.1 [-22.9, -13.4] | +17.4 [+15.6, +19.5] | -0.7 [-4.8, +3.3] |
| Muse Spark 1.1 | -35.3 [-50.4, -19.5] | +34.2 [+27.7, +41.1] | -1.1 [-15.2, +13.9] | -23.5 [-35.1, -11.4] | +24.4 [+20.7, +28.1] | +0.9 [-10.2, +12.3] | -31.1 [-44.6, -17.4] | +28.5 [+23.2, +34.6] | -2.6 [-14.7, +10.2] |
| GLM 5.2 | -50.3 [-61.5, -39.1] | +48.8 [+42.4, +56.0] | -1.4 [-9.2, +6.9] | -27.2 [-34.5, -19.8] | +27.9 [+25.3, +30.4] | +0.7 [-6.1, +7.4] | -38.8 [-47.7, -29.9] | +37.2 [+32.1, +42.6] | -1.6 [-8.7, +5.7] |
| Claude Sonnet 4.6 (A) | -50.4 [-58.5, -42.5] | +45.3 [+41.0, +49.9] | -5.1 [-11.5, +1.4] | -35.9 [-41.6, -29.8] | +31.6 [+29.4, +33.7] | -4.3 [-9.7, +1.4] | -42.1 [-49.5, -34.9] | +35.8 [+32.2, +39.6] | -6.3 [-12.0, +0.0] |
| Grok 4.5 | -37.5 [-50.3, -24.8] | +32.0 [+25.1, +39.4] | -5.5 [-17.4, +7.4] | -25.0 [-33.6, -16.0] | +20.4 [+17.4, +23.6] | -4.6 [-13.6, +4.8] | -31.1 [-42.3, -19.7] | +25.5 [+20.3, +31.2] | -5.6 [-15.8, +5.4] |
| Claude Sonnet 4.6 (B) | -38.7 [-52.8, -24.4] | +32.8 [+26.8, +39.6] | -5.8 [-19.0, +8.5] | -34.3 [-45.6, -22.7] | +28.0 [+24.6, +31.3] | -6.3 [-16.8, +4.2] | -35.1 [-47.4, -23.1] | +25.8 [+21.1, +31.2] | -9.3 [-20.9, +2.3] |
| Gemma 4 31B | -52.4 [-60.9, -43.4] | +45.6 [+40.7, +50.7] | -6.8 [-13.2, -0.1] | -33.6 [-40.3, -27.0] | +29.8 [+27.4, +32.2] | -3.8 [-9.7, +1.8] | -41.6 [-49.0, -34.3] | +35.8 [+31.8, +40.1] | -5.9 [-11.5, -0.2] |

**Table 7:** Simulated persona ROI (%) under proper betting strategies (Log and Sph. are the per-contract rules). 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 |

**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 $\pm 0.05$ 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 $\Delta S$ (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

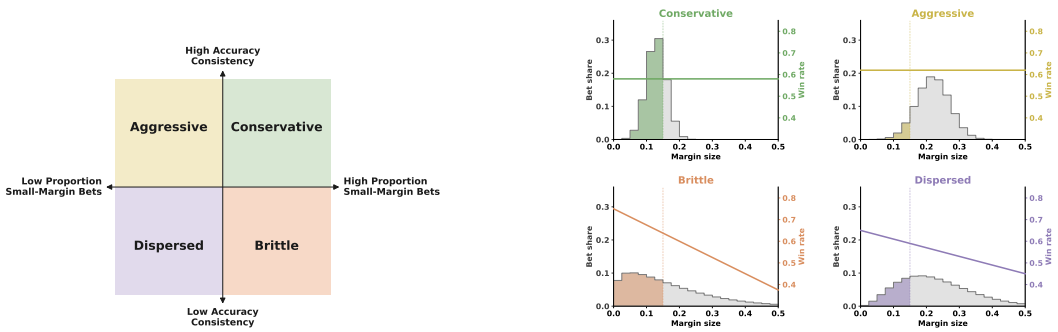
$$a(m) \;=\; \text{clip}\big(a_0 - s \cdot \max(0,\, m - \tau),\ 0.01,\ 0.99\big),$$

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 1b, we visualize this by fitting a linear regression (OLS) for comparability with the other personas.

### E.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 2 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.




**(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 1:** Schematics for synthetic forecasting personas and their prediction profiles.

In contrast, the per-contract logarithmic rule, $w = |\log p - \log q|$ on the traded leg, 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. This property belongs to the one-leg, per-contract rule only: the netted standard log bet $\mathrm{logit}(p) - \mathrm{logit}(q)$ grows superlinearly and is unbounded as $p \to 1$. 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.

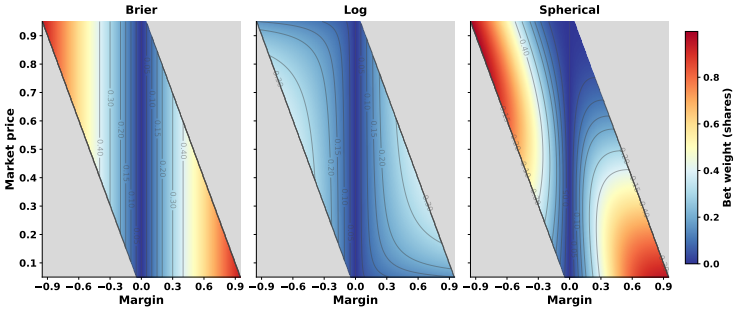


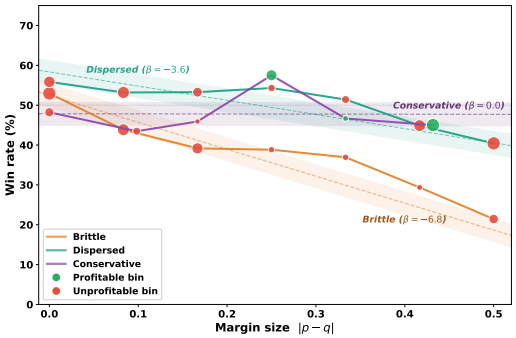

**Figure 2:** Per-contract capital allocation (bet weight in shares on one leg) as a function of signed margin and market price ($q$, $y$-axis) for the three proper betting strategies. Warmer colors indicate larger bets. Grey regions are infeasible ($p \notin (0, 1)$). Contour lines mark constant weight levels. The negative-margin half of each panel is consistent in shape with the NO-leg weight $g'(1-p) - g'(q^-)$ (for Log, larger at high market prices, as the NO leg is). The Brier and Spherical surfaces reproduce the formulas of Appendix E.2 in shares; the Log surface's color scale could not be re-derived from its formula and should be read qualitatively (sublinear growth, price dependence).

### E.6 Mapping empirical forecasters to personas

First, 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,

| Model | S-M Share | $\Delta S$ | 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 $\Delta S$ rounded to $+0.000$.




**(b)** Binned per-bet win rate vs. margin $|p-q|$. Each dot has area proportional to profit for its persona. Dashed lines are OLS fits where $\beta$ denotes the slope (in percentage points of win rate per margin bin). Overlapping points are jittered horizontally for clarity.

**Figure 3:** Mapping real models to personas.

Figure 3b 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 ($\beta = 0$). Here $\beta$ is the OLS slope in percentage points of win rate per margin bin (bins of width $1/12$ in Figure 3b); per unit margin this is about $-0.8$ for Brittle and $-0.4$ for Dispersed, on the scale of the Consistency column of Table 7.

**Table 8:** Language model margin concentration. Entries report the share of bets (%) in each margin bin. $\Delta S$ 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) | $\Delta S$ |
|---|---|---|---|---|---|---|---|---|
| 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** | $\mathbf{-0.054}$ |
| 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** | $\mathbf{-0.120}$ |
| 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** | $\mathbf{+0.001}$ |

### E.7 Details about the live pilot test

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.

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.

### E.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 nonzero bid-ask spread present in almost all real-world prediction markets), as this implies no actionable margin.

Whenever a market is traded, we rebalance the position toward the new target $\tau$ on each cycle, as is detailed in Appendix D.4. The rebalancing schedule, the 10¢ re-evaluation threshold, and the horizon limits are empirical choices without a formal risk-control justification.

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.

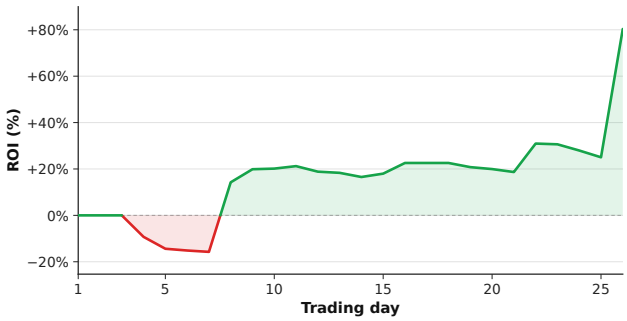



**Figure 4:** ROI of the live Gemini 3 agent on Kalshi over 26 trading days (statistics in Table 9).

We report the Sharpe ratio only for the live test, whose daily returns form a continuous time series; offline, events resolve independently and most models lose money, so a Sharpe ratio would be uninformative. Table 9 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.

### E.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[5], creating a mismatch between the LLM's belief about market closure and the actual outcome realization.

[5]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 named the `expected_expiration_time`, which indicates the event's resolution time.

**Table 9:** Live trading statistics for the Gemini 3 agent on Kalshi. Robustness rows and the Sharpe ratio are from the frozen trade export analyzed in Appendix F.4; the Sharpe ratio uses the calendar-day ($\sqrt{365}$) convention throughout. The frozen export records 226 executed trades on 123 markets, while the order counts in the table (396 placed, 236 filled, 129 markets) come from the agent's order log. The two artifacts were frozen at different times and are not reconciled; every robustness statistic uses the export.

| | |
|---|---|
| Starting capital | \$200.00 |
| Trading window | 24 Mar–7 May 2026 |
| Calendar days (days with trades) | 45 (26) |
| Forecasts issued | 1,605 |
| Trade signals placed (filled) | 396 (236) |
| Markets traded | 129 |
| Ending capital | \$360.67 |
| ROI | $+80.33\%$ |
| 3-trimmed ROI | $+24.2\%$ |
| Realized ROI (settled/closed markets) | $+28.2\%$ |
| Unrealized share of profit | $65\%$ |
| Gross turnover | $5.5\times$ capital |
| Sharpe ratio ($\sqrt{365}$, 44 daily returns) | 4.22 |
| 95% CI of Sharpe (block bootstrap) | $[-0.89, +7.69]$ |
| Shares transacted | 2,657 |
| Exchange fees paid | \$26.31 |
| Win rate (unresolved counted as losses) | $50.9\%$ |
| Win rate (resolved markets only) | $67.4\%$ |
| Maximum drawdown (smoothed NAV; 20.68% pre-smoothing) | \$31.47 (15.7%) |
| Exchange fees vs. net profit | \$26.31 / \$160.67 |

In addition to these eligibility criteria, we also drop a market from consideration due to an under-specified 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 Open LLM 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 eligible set[6].

### E.7.3 Case studies

We present two representative case studies drawn from live trading runs to illustrate how the proposed strategy operates in practice. Both cases come from the live run: the frozen trade export (Appendix F.4) lists the SPR trades of Case 2 on 28–31 March 2026 and the Executive Order trades of Case 1 on 22–24 April 2026, both inside the 24 March–7 May window. These examples highlight both the decision-making process and the resulting performance.

**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

[6] Subsequent monthly versions of this contract now explicitly specify the LMArena leaderboard as the resolution source.

> 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."

After eight prior forecasts in which model and market had agreed on $p_{\text{yes}} \approx 0.77$, the model's $p_{\text{yes}}$ 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, selling 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.

### E.8 Profitability at extreme margins

While most large-margin bets are wrong (Table 10 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 $[0.70, 0.90)$ 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. Both are drawn from the offline (backtested) dataset, not from the live deployment of Section 4.3.

**Table 10:** 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 11:** Per-market Brier-strategy economics for two extreme-margin case studies. $p_{\text{yes}}$ is the model's predicted probability for YES; $q_{\text{yes}}$ is the YES-side ask price; $y$ is the realized outcome. Weight is the margin $|p - q_{\text{yes}}|$.

| Market (Kalshi ID) | Predicted | Resolved | $p_{\text{yes}}$ | $q_{\text{yes}}$ | $y$ | 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 about 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 market's probability for $> 3$ inches was greater than that for $> 2$ 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.

# F Additional analyses added after review

This section reports analyses added after peer review. They address four questions: how much statistical uncertainty surrounds the offline comparison (Appendix F.1); whether proper betting remains competitive against stronger heuristic baselines that reduce exposure when the forecaster is uncertain (Appendix F.2); whether the offline results survive execution frictions and are not driven by a few tail trades (Appendix F.3); and how robust the live-deployment result is (Appendix F.4). Appendix F.5 repeats the offline comparison on a second, larger data source.

## F.1 Statistical uncertainty for the offline comparison

**Resampling unit.** The ROIs in the submitted version's Tables 2 and 3 (Prophet Arena data; reproduced with intervals in Tables 12 and 13) and in their expanded counterparts are ratios of sums, $\mathrm{ROI} = \sum_i \mathrm{profit}_i / \sum_i \mathrm{cost}_i$, not means of per-bet returns, and contracts within one event are mutually exclusive and therefore dependent. We estimate uncertainty with an event-clustered bootstrap: we resample events with replacement ($B = 6{,}000$), carry every contract of a drawn event together, and recompute each statistic on the resample. Standard errors are the standard deviation of the resampling distribution, and 90% intervals are point $\pm\ 1.645\,\mathrm{SE}$ unless stated otherwise. Resampling individual contracts instead of events understates the standard error by 24% at the median across models (up to $1.8\times$), because it treats the roughly $2{,}300$ contracts as independent when the effective sample size is the number of events. Clustering more coarsely than the event (by market series; 53% of the events in the subset are NBA game markets) does not widen the intervals further, so the event is the appropriate unit.

**Construction of the standardized subset.** The standardized subset used for the Prophet Arena comparison of the submitted version is the intersection of events forecast by all compared models (205 events), from which 11 events were removed by a greedy procedure until Gemini 3 and Claude Opus 4.6 both attain a relative Brier score of at least $+0.001$ against the market, leaving 194 events and $2{,}345$ contracts (of which $2{,}344$ carry a nonzero bid-ask spread; median spread $2¢$). On the full 205-event intersection no model attains a positive relative Brier score.

**What the intervals show.** Across the strategy comparison in Table 12, only one cell is separated from zero at the 90% level: proper (Brier) betting on Claude Opus 4.6, with interval $[+8.8, +35.4]$. The relative Brier score of the two strongest models straddles zero, while that of the weaker models

**Table 12:** Submitted Table 2 (Prophet Arena standardized subset) with event-clustered 90% confidence intervals (mean [lower, upper]). ROI in %; $\Delta S$ in Brier units. The Kelly column reports the no-leverage capped Kelly rule ($\sum_i f_i \leq 1$ per event) with a bounded percentile interval from without-replacement event subsamples, because the leveraged variant used in the original column has no finite standard error (see text). Bold marks the best ROI per row.

| Model | $\Delta S$ | Proper | Max (Mkt) | Max (Grp) | Inv-Margin | Kelly-Alike | Kelly (capped) |
|---|---|---|---|---|---|---|---|
| Claude Opus 4.6 | +0.0016 [−0.0037, +0.0069] | **+22.1** [+8.8, +35.4] | +5.0 [−2.3, +12.3] | −14.0 [−32.6, +4.6] | −3.5 [−9.4, +2.4] | +10.9 [+4.6, +17.2] | −99.7 [−99.9, −90.6] |
| Gemini 3 | +0.0008 [−0.0021, +0.0037] | **+8.1** [−2.2, +18.4] | +1.3 [−6.3, +8.9] | −0.1 [−24.5, +24.3] | +0.7 [−8.5, +9.9] | +1.8 [−3.8, +7.4] | +4.7 [−56.7, +170.4] |
| GPT-5.2 (Base) | −0.0347 [−0.0551, −0.0143] | **+4.5** [−13.1, +22.1] | +2.4 [−5.3, +10.1] | −14.9 [−31.6, +1.8] | +1.9 [−3.8, +7.6] | −0.1 [−7.3, +7.1] | −100.0 [−100.0, −100.0] |
| LLaMA 4 Maverick | −0.0450 [−0.0645, −0.0255] | −13.7 [−31.2, +3.8] | −4.1 [−13.5, +5.3] | −15.5 [−36.6, +5.6] | −4.0 [−13.9, +5.9] | −10.0 [−18.3, −1.7] | −100.0 [−100.0, −100.0] |
| Grok 4.1 Fast | −0.0462 [−0.0650, −0.0274] | −11.6 [−31.6, +8.4] | −7.8 [−14.9, −0.7] | −26.4 [−41.0, −11.8] | −7.0 [−14.8, +0.8] | −6.3 [−15.1, +2.5] | −100.0 [−100.0, −100.0] |

**Table 13:** Submitted Table 3 (Prophet Arena standardized subset) with event-clustered 90% confidence intervals. All quantities in ROI units (%), with $\Delta S + D = \text{ROI}$ holding in every resample.

| | Brier | | | Log | | | Spherical | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | $\Delta S$ | $D$ | ROI | $\Delta S$ | $D$ | ROI | $\Delta S$ | $D$ | ROI |
| Claude Opus 4.6 | +3.2 [−10.5, +16.9] | +17.8 [+12.9, +22.7] | +21.0 [+7.7, +34.3] | +4.5 [−5.2, +14.2] | +16.9 [+9.0, +24.8] | +21.4 [+10.0, +32.8] | −1.3 [−12.8, +10.2] | +10.1 [+6.6, +13.6] | +8.8 [−2.1, +19.7] |
| Gemini 3 | +4.0 [−6.0, +14.0] | +5.0 [+3.9, +6.1] | +9.0 [−1.3, +19.3] | +3.9 [−5.8, +13.6] | +5.0 [+3.6, +6.4] | +8.9 [−1.1, +18.9] | −2.4 [−11.6, +6.8] | +2.9 [+2.0, +3.8] | +0.4 [−8.3, +9.1] |
| GPT-5.2 (Base) | −108.4 [−196.3, −20.5] | +112.7 [+29.4, +196.0] | +4.3 [−13.3, +21.9] | −5.8 [−22.0, +10.4] | +15.9 [+12.5, +19.3] | +10.1 [−6.0, +26.2] | −52.7 [−91.8, −13.6] | +46.7 [+8.1, +85.3] | −6.0 [−21.3, +9.3] |
| LLaMA 4 Maverick | −123.7 [−188.9, −58.5] | +110.6 [+49.1, +172.1] | −13.1 [−30.6, +4.4] | −21.0 [−40.3, −1.7] | +20.2 [+15.0, +25.4] | −0.8 [−18.6, +17.0] | −53.3 [−82.5, −24.1] | +38.5 [+10.5, +66.5] | −14.8 [−23.2, −6.4] |
| Grok 4.1 Fast | −137.4 [−222.1, −52.7] | +128.3 [+50.6, +206.0] | −9.1 [−29.1, +10.9] | −20.3 [−37.9, −2.7] | +23.8 [+15.8, +31.8] | +3.5 [−17.4, +24.4] | −89.3 [−137.5, −41.1] | +74.3 [+29.5, +119.1] | −15.0 [−27.2, −2.8] |

is significantly negative. Because per-event returns are heavy-tailed and only 194 events are available, the fine-grained ranking of models in the middle of the table is not statistically supported. Relative claims are better identified than absolute ones: evaluating the difference in ROI between two strategies on the same resampled events cancels the common event draw. Under this paired test, proper betting on Claude Opus 4.6 beats Max (Mkt), Inv-Margin, and Max (Grp) by $+17$, $+26$, and $+36$ ROI points respectively, with 95% intervals excluding zero, and its proper-betting ROI significantly exceeds that of the six weakest models. In Table 13, the Brier score-gap and divergence terms of the weaker models are large and offsetting, so the individual terms carry uncertainty of $\pm 50$ ROI points or more even though their sum is tight ($\pm 10$). The Log decomposition is the most precisely estimated. Tables 14 and 15 give the corresponding standard errors for all twelve models.[7]

**The Kelly column.** The Kelly baseline in the submitted Table 2 sums independent per-outcome Kelly fractions $f_i = (p_i - q_i)/(1 - q_i)$ within an event without a joint budget constraint and lets the resulting stake exceed the bankroll. On this data the summed fractions reach $34\times$ the bankroll on a single event for Claude Opus 4.6, so the $-99.9\%$ entries are a consequence of leverage rather than of the Kelly criterion itself. Capping deployment at the bankroll ($\sum_i f_i \leq 1$ per event) gives the values reported in Table 12: Gemini 3 becomes marginally profitable ($+4.7\%$), Claude Opus 4.6 still loses $99.7\%$, and every other model is ruined. Kelly compounds a rolling bankroll, so its return is a terminal-wealth statistic that is not comparable cell-for-cell with the per-bet ROI of the other columns, and resampling events with replacement duplicates the few decisive events and inflates the tails multiplicatively; we therefore report a percentile interval from without-replacement subsamples, floored at $-100\%$. Half-Kelly is identical to Kelly-Alike on the per-bet basis, since halving every stake cancels in a ratio of sums; on the compounded basis it does not rescue the strategy (most models are still ruined). We keep the leveraged column in the main tables and state this caveat in their captions.

### F.2 Stronger exposure-reducing baselines

The heuristic baselines of Section 4.1 do not explicitly reduce exposure when the forecaster is uncertain. We therefore compare proper (Brier) betting against a family of simple rules that do, all evaluated on the same standardized subset with the same within-spread mask and the same event-clustered bootstrap. Writing $e_i = |p_i - q_i|$ for the margin on the positive-edge side and $H(p) = -p \ln p - (1-p)\ln(1-p)$ for the binary entropy of the forecast, the rules are:

- **Cap-Edge**($\kappa$): $w_i = \min(e_i, \kappa)$, proper betting with a ceiling on any single stake, so that extreme disagreements with the market are not fully trusted.

[7] The strategy tables and the decomposition tables are produced by two scripts whose greedy drop steps select slightly different events, which is why proper betting on Claude Opus 4.6 appears as $+22.1\%$ in Tables 12 and 14 and as $+21.0\%$ in Tables 13 and 15, and Inv-Margin as $-3.5\%$ here and $-4.6\%$ in Table 16. The differences are within one standard error.

**Table 14:** ROI (%) $\pm$ event-clustered standard error for all twelve evaluated models on the standardized subset. $\Delta S$ is the relative Brier score. The leveraged Kelly baseline has no finite standard error and is omitted (see Table 12 for the capped variant).

| Model | $\Delta S$ | Proper | Max (Mkt) | Max (Grp) | Inv-Margin | Kelly-Alike |
|---|---|---|---|---|---|---|
| Claude Opus 4.6 | $+0.0016$ | $+22.1 \pm 8.1$ | $+5.0 \pm 4.4$ | $-14.0 \pm 11.3$ | $-3.5 \pm 3.6$ | $+10.9 \pm 3.8$ |
| Gemini 3 | $+0.0008$ | $+8.1 \pm 6.2$ | $+1.3 \pm 4.6$ | $-0.1 \pm 14.8$ | $+0.7 \pm 5.6$ | $+1.8 \pm 3.4$ |
| GPT-5.2 (Base) | $-0.0347$ | $+4.5 \pm 10.7$ | $+2.4 \pm 4.7$ | $-14.9 \pm 10.1$ | $+1.9 \pm 3.5$ | $-0.1 \pm 4.4$ |
| Claude Sonnet 4.5 | $-0.0426$ | $-3.5 \pm 8.0$ | $+1.2 \pm 3.9$ | $-19.8 \pm 8.1$ | $+1.4 \pm 3.1$ | $+0.7 \pm 2.9$ |
| LLaMA 4 Maverick | $-0.0450$ | $-13.7 \pm 10.6$ | $-4.1 \pm 5.7$ | $-15.5 \pm 12.8$ | $-4.0 \pm 6.0$ | $-10.0 \pm 5.0$ |
| GPT-5.2 (High) | $-0.0460$ | $-2.4 \pm 9.7$ | $-1.6 \pm 4.1$ | $-13.3 \pm 10.6$ | $-1.9 \pm 3.7$ | $-1.0 \pm 3.4$ |
| Grok 4.1 Fast | $-0.0462$ | $-11.6 \pm 12.2$ | $-7.8 \pm 4.3$ | $-26.4 \pm 8.9$ | $-7.0 \pm 4.7$ | $-6.3 \pm 5.4$ |
| DeepSeek V3.2 | $-0.0466$ | $-12.2 \pm 9.4$ | $-7.3 \pm 4.9$ | $-30.2 \pm 7.0$ | $-5.1 \pm 4.2$ | $-5.8 \pm 4.3$ |
| Kimi K2 Thinking | $-0.0492$ | $-9.5 \pm 9.6$ | $-3.1 \pm 4.1$ | $-24.3 \pm 8.0$ | $+1.5 \pm 3.6$ | $-2.5 \pm 3.2$ |
| Minimax M2 | $-0.0537$ | $+1.3 \pm 12.6$ | $-4.0 \pm 5.4$ | $-22.8 \pm 8.4$ | $-6.5 \pm 6.2$ | $-3.9 \pm 9.1$ |
| DeepSeek R1 | $-0.0636$ | $-20.3 \pm 8.6$ | $-7.6 \pm 4.7$ | $-25.4 \pm 9.4$ | $+0.7 \pm 3.6$ | $-9.6 \pm 3.4$ |
| Qwen 3 235B | $-0.0838$ | $-13.3 \pm 13.4$ | $-5.3 \pm 5.7$ | $-14.8 \pm 11.7$ | $+12.2 \pm 7.2$ | $-9.3 \pm 6.2$ |

**Table 15:** Decomposition of ROI (%) into score gap $\Delta S$ and Bregman divergence $D$, $\pm$ event-clustered standard error, for all twelve models on the standardized subset.

| | Brier | | | Log | | | Spherical | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | $\Delta S$ | $D$ | ROI | $\Delta S$ | $D$ | ROI | $\Delta S$ | $D$ | ROI |
| Claude Opus 4.6 | $+3.2 \pm 8.3$ | $+17.8 \pm 3.0$ | $+21.0 \pm 8.1$ | $+4.5 \pm 5.9$ | $+16.9 \pm 4.8$ | $+21.4 \pm 6.9$ | $-1.3 \pm 7.0$ | $+10.1 \pm 2.1$ | $+8.8 \pm 6.6$ |
| Gemini 3 | $+4.0 \pm 6.1$ | $+5.0 \pm 0.7$ | $+9.0 \pm 6.2$ | $+3.9 \pm 5.9$ | $+5.0 \pm 0.8$ | $+8.9 \pm 6.1$ | $-2.4 \pm 5.6$ | $+2.9 \pm 0.6$ | $+0.4 \pm 5.3$ |
| GPT-5.2 (Base) | $-108.4 \pm 53.4$ | $+112.7 \pm 50.7$ | $+4.3 \pm 10.7$ | $-5.8 \pm 9.9$ | $+15.9 \pm 2.1$ | $+10.1 \pm 9.8$ | $-52.7 \pm 23.7$ | $+46.7 \pm 23.5$ | $-6.0 \pm 9.3$ |
| Claude Sonnet 4.5 | $-111.2 \pm 39.9$ | $+110.2 \pm 36.9$ | $-0.9 \pm 8.0$ | $-10.0 \pm 7.7$ | $+17.6 \pm 2.1$ | $+7.5 \pm 7.2$ | $-63.4 \pm 23.8$ | $+59.9 \pm 22.0$ | $-3.6 \pm 6.0$ |
| LLaMA 4 Maverick | $-123.7 \pm 39.6$ | $+110.6 \pm 37.4$ | $-13.1 \pm 10.6$ | $-21.0 \pm 11.7$ | $+20.2 \pm 3.2$ | $-0.8 \pm 10.8$ | $-53.3 \pm 17.7$ | $+38.5 \pm 17.0$ | $-14.8 \pm 5.1$ |
| GPT-5.2 (High) | $-159.2 \pm 66.2$ | $+159.5 \pm 61.0$ | $+0.3 \pm 9.7$ | $-13.1 \pm 12.7$ | $+21.9 \pm 9.1$ | $+8.8 \pm 7.6$ | $-85.9 \pm 37.1$ | $+81.2 \pm 34.5$ | $-4.6 \pm 7.7$ |
| Grok 4.1 Fast | $-137.4 \pm 51.5$ | $+128.3 \pm 47.2$ | $-9.1 \pm 12.2$ | $-20.3 \pm 10.7$ | $+23.8 \pm 4.9$ | $+3.5 \pm 12.7$ | $-89.3 \pm 29.3$ | $+74.3 \pm 27.2$ | $-15.0 \pm 7.4$ |
| DeepSeek V3.2 | $-116.4 \pm 38.5$ | $+107.3 \pm 35.3$ | $-9.1 \pm 9.4$ | $-18.0 \pm 9.6$ | $+18.7 \pm 2.9$ | $+0.7 \pm 8.4$ | $-64.3 \pm 21.1$ | $+53.8 \pm 20.0$ | $-10.5 \pm 6.5$ |
| Kimi K2 Thinking | $-123.7 \pm 45.7$ | $+115.4 \pm 42.3$ | $-8.3 \pm 9.6$ | $-21.5 \pm 8.1$ | $+20.4 \pm 3.2$ | $-1.2 \pm 7.8$ | $-71.4 \pm 25.5$ | $+65.0 \pm 24.1$ | $-6.4 \pm 6.8$ |
| Minimax M2 | $-138.7 \pm 55.8$ | $+137.6 \pm 49.9$ | $-1.1 \pm 12.6$ | $-10.9 \pm 9.4$ | $+19.8 \pm 3.0$ | $+8.9 \pm 10.3$ | $-87.6 \pm 28.1$ | $+70.7 \pm 19.9$ | $-17.0 \pm 15.8$ |
| DeepSeek R1 | $-138.9 \pm 37.1$ | $+117.6 \pm 35.4$ | $-21.3 \pm 8.6$ | $-35.5 \pm 8.8$ | $+22.9 \pm 3.6$ | $-12.6 \pm 8.2$ | $-74.4 \pm 21.8$ | $+55.7 \pm 21.4$ | $-18.7 \pm 6.2$ |
| Qwen 3 235B | $-148.1 \pm 45.1$ | $+138.1 \pm 41.3$ | $-9.9 \pm 13.4$ | $-47.1 \pm 23.3$ | $+44.6 \pm 15.9$ | $-2.4 \pm 11.7$ | $-89.4 \pm 29.1$ | $+75.5 \pm 27.5$ | $-13.8 \pm 9.4$ |

- **Conf-Prop** and **Conf$^2$-Prop**: $w_i = e_i\,|2p_i - 1|$ and $w_i = e_i\,(2p_i - 1)^2$, proper betting scaled down toward coin-flip forecasts.
- **Conf-Gate**$(\tau)$: proper betting only when $|2p_i - 1| \geq \tau$, otherwise abstain.
- **Divide-Entropy**: $w_i = e_i / H(p_i)$, proper betting scaled up by forecast confidence.
- **Certainty**: $w_i = |2p_i - 1|$, sizing by confidence alone.
- **Half-Kelly**, with deployment capped at the bankroll.

**Findings.** The two families behave in opposite ways. Rules that reduce exposure on near-coin-flip forecasts hurt, and they hurt monotonically in their aggressiveness: the mean ROI over the twelve models falls from $-3.2$ for proper betting to $-7.6$ for Conf-Prop and $-8.8$ for Conf$^2$-Prop, and Conf-Gate degrades from $-3.9$ at $\tau = 0.05$ to $-6.7$ at $\tau = 0.2$ and $-10.4$ at $\tau = 0.4$. Rules that increase exposure with confidence hurt more: Divide-Entropy concentrates capital on the forecasts with the most extreme probabilities, which is exactly where the models are least calibrated, giving a mean of $-19.3$ and ruining GPT-5.2 (Base); sizing by certainty alone discards the margin signal and gives $-4.9$. Proper betting is never significantly worse than any of these rules in more than one of twelve models under the paired test, and on Claude Opus 4.6 it is significantly better than the certainty-weighted rules. Rules that cap large disagreements behave differently. Cap-Edge is a statistical tie with proper betting in all twelve models at any $\kappa$ between $0.03$ and $0.25$, and at the cross-validated $\kappa^* \approx 0.27$ it improves the pooled mean by $2.7$ points (Table 17); Inv-Margin has the best pooled mean of all rules ($-0.9$). Both gains come from damage control on weak forecasters, whose large disagreements with the market are mostly errors, and both rules give up return on the profitable models (Claude Opus 4.6: $+21.0$ under proper betting versus $-4.6$ under Inv-Margin

**Table 16:** ROI (%) $\pm$ event-clustered standard error of proper betting and three exposure-adjusting rules on the standardized subset. Inv-Margin is the baseline of Section 4.1; Divide-Entropy and Certainty increase exposure with forecast confidence.

| **Model** | **Proper** | **Inv-Margin** | **Divide-Entropy** | **Certainty** $\lvert 2p-1 \rvert$ |
|---|---|---|---|---|
| Claude Opus 4.6 | $+21.0 \pm 8.1$ | $-4.6 \pm 3.6$ | $+37.7 \pm 26.2$ | $+4.3 \pm 3.7$ |
| Gemini 3 | $+9.0 \pm 6.3$ | $+1.6 \pm 5.6$ | $+1.0 \pm 1.7$ | $+0.4 \pm 3.9$ |
| GPT-5.2 (High) | $+0.3 \pm 9.8$ | $-2.9 \pm 3.6$ | $-45.8 \pm 21.1$ | $-3.3 \pm 3.1$ |
| GPT-5.2 (Base) | $+4.3 \pm 10.7$ | $+2.0 \pm 3.4$ | $-100.0 \pm 48.7$ | $-4.3 \pm 3.6$ |
| Kimi K2 Thinking | $-8.3 \pm 9.6$ | $+1.4 \pm 3.6$ | $+9.8 \pm 16.9$ | $-3.4 \pm 2.9$ |
| Grok 4.1 Fast | $-9.1 \pm 12.2$ | $-3.3 \pm 4.7$ | $-24.2 \pm 29.3$ | $-7.6 \pm 3.5$ |
| Claude Sonnet 4.5 | $-0.9 \pm 7.8$ | $+1.8 \pm 3.1$ | $-8.1 \pm 41.6$ | $+0.3 \pm 2.3$ |
| DeepSeek V3.2 | $-9.1 \pm 9.4$ | $-7.2 \pm 4.2$ | $+14.2 \pm 6.3$ | $-8.4 \pm 3.5$ |
| Qwen 3 235B | $-9.9 \pm 13.4$ | $+12.2 \pm 7.2$ | $-22.2 \pm 23.0$ | $-10.6 \pm 4.2$ |
| LLaMA 4 Maverick | $-13.1 \pm 10.6$ | $-3.2 \pm 6.0$ | $-13.0 \pm 11.3$ | $-9.1 \pm 4.2$ |
| Minimax M2 | $-1.1 \pm 12.5$ | $-7.2 \pm 6.1$ | $-41.1 \pm 34.5$ | $-7.1 \pm 5.4$ |
| DeepSeek R1 | $-21.3 \pm 8.7$ | $-0.9 \pm 3.6$ | $-40.0 \pm 14.7$ | $-10.6 \pm 2.9$ |
| **Mean** | $-3.2$ | $-0.9$ | $-19.3$ | $-4.9$ |

**Table 17:** Proper betting versus Cap-Edge at the in-sample optimal cap $\kappa^* = 0.27$ (ROI, %). Leave-one-model-out cross-validation selects $\kappa \in [0.25, 0.29]$ in every fold; the held-out mean ROI of $-0.58$ matches the in-sample $-0.47$. The pooled gain of $+2.7$ points over proper betting (95% event-clustered CI $[+0.4, +5.1]$) comes entirely from the weaker forecasters; on the one model that is reliably profitable the cap costs return.

| **Model** | **Proper** | **Cap-Edge**$(0.27)$ | $\Delta$ |
|---|---|---|---|
| Claude Opus 4.6 | $+21.0$ | $+18.9$ | $-2.0$ |
| Gemini 3 | $+9.0$ | $+9.0$ | $0.0$ |
| GPT-5.2 (High) | $+0.3$ | $+3.8$ | $+3.5$ |
| GPT-5.2 (Base) | $+4.3$ | $+7.1$ | $+2.8$ |
| Claude Sonnet 4.5 | $-0.9$ | $+3.0$ | $+3.9$ |
| Kimi K2 Thinking | $-8.3$ | $-6.1$ | $+2.2$ |
| DeepSeek V3.2 | $-9.1$ | $-5.5$ | $+3.6$ |
| Qwen 3 235B | $-9.9$ | $-4.4$ | $+5.5$ |
| LLaMA 4 Maverick | $-13.1$ | $-5.6$ | $+7.5$ |
| DeepSeek R1 | $-21.3$ | $-16.0$ | $+5.3$ |

and $+18.9$ under Cap-Edge). The summary is that the profit lives in the margin $|p - q|$, not in how confident the forecast sounds: with a well-calibrated forecaster, uncapped proper betting is the best rule tested, and the value of an exposure cap grows with how untrustworthy the forecaster is. Half-Kelly capped at the bankroll ruins most models (mean $-89$) and is not competitive.

### F.3 Execution frictions and tail concentration

The archived data record top-of-book quotes (best bid and ask, top-level size) and cumulative volume at the time of each forecast, but not the full depth of the order book, so a literal multi-level book walk cannot be reconstructed. The markets on which bets are placed are liquid (median spread $1¢$, median volume $2{,}170$ contracts), so we stress the results with a parametric execution model: every bet already pays the ask, and we add a fixed slippage of $3¢$ or $5¢$ to the paid price, applied before the betting decision so that bets with a margin smaller than the slippage drop out.

Two conclusions follow. The positive returns of the well-calibrated forecasters are not an artifact of frictionless execution: Claude Opus 4.6 moves from $+21.0\%$ to $+20.5\%$ at $+3¢$ and Gemini 3

**Table 18:** Mean ROI (%) across the twelve models under added slippage, and the ROI of proper betting for the two profitable models. Slippage does not reorder the strategies; proper betting, which takes the largest positions, degrades fastest on the losing models and is nearly unchanged on the profitable ones.

| | Mean over 12 models | | | | | Proper | |
|---|---|---|---|---|---|---|---|
| **Slippage** | **Proper** | **Max (Mkt)** | **Max (Grp)** | **Inv-Margin** | **Kelly-Alike** | **Claude Opus 4.6** | **Gemini 3** |
| $0¢$ | $-3.2$ | $-1.2$ | $-17.8$ | $+0.5$ | $-1.1$ | $+21.0$ | $+9.0$ |
| $+3¢$ | $-17.8$ | $-3.6$ | $-25.1$ | $-0.7$ | $-1.9$ | $+20.5$ | $+9.0$ |
| $+5¢$ | $-27.0$ | $-4.4$ | $-30.9$ | $-0.7$ | $-2.3$ | — | — |

**Table 19:** Tail concentration of proper-betting returns (ROI, %) after removing the $k$ most profitable trades out of roughly 600–800 per model. The ten largest winning trades account for 30–40% of gross winning P&L.

| **Model** | **All trades** | **Drop top 5** | **Drop top 10** |
|---|---|---|---|
| Claude Opus 4.6 | $+21.0$ | $+12.9$ | $+8.4$ |
| Gemini 3 | $+9.0$ | $+2.6$ | $-3.4$ |
| GPT-5.2 (Base) | $+4.3$ | $-6.4$ | $-12.9$ |

is unchanged, and Claude Opus 4.6 remains profitable after its ten best trades are removed. The pooled picture is less favorable to proper betting: on the mean over all twelve models it ranks fourth of five at every slippage level, because the edge-damping baselines barely deploy capital and so have little to lose, whereas proper betting takes edge-sized positions on models that are worse than the market. The returns of Gemini 3 and GPT-5.2 (Base) are also tail-dependent. The defensible statement is the per-model one: for a forecaster that beats the market, proper betting converts the edge into money and does so robustly to slippage; for a forecaster that does not, no allocation rule helps and the low-exposure heuristics simply lose less.

### F.4 Live deployment: uncertainty, concentration, and attribution

The figures below are computed from the frozen trade export (`paper-trades.json`: 123 markets, 226 trades, net $+\$160.67$ on \$200) and the smoothed daily NAV series (45 marks, 44 daily returns). Throughout the paper the Sharpe ratio is annualized with the calendar-day convention ($\sqrt{365}$), and Section 4.3 and Table 9 quote the statistics of this section. The deployment ran from 24 March to 7 May 2026 (45 calendar days, hence 45 NAV marks and 44 daily returns), with trades placed on 26 of those days; these are the window and day counts reported in Table 9, and the 24–30 March drawdown falls within this window.

**Uncertainty of the Sharpe ratio.** The deployment produced 44 daily returns. The daily Sharpe ratio is 0.221, which annualizes to 4.22 ($\times\sqrt{365}$) or 3.51 ($\times\sqrt{252}$). Returns are strongly right-skewed (skewness $+3.25$, excess kurtosis $+11.8$): a few resolution-day jumps carry the mean. Table 20 reports the standard error under three estimators. The annualization-invariant $t$-ratio is between 1.45 and 2.18 depending on the non-normality correction; the parametric interval that excludes zero relies on the same positive skew produced by the tail trades whose influence is at issue, and both nonparametric bootstrap intervals include zero at the 95% level. At the 90% level the block bootstrap gives $[+0.46, +7.14]$ and the IID bootstrap $[-0.03, +7.16]$ for the $\sqrt{365}$-annualized ratio. Bootstrapping the terminal return over the 44 daily returns gives a median of $+75\%$ with 95% interval $[-12\%, +343\%]$ (IID) or $[-8\%, +302\%]$ (block), and $\Pr[\text{ROI} > 0] \approx 0.94$–$0.95$. Because the NAV series was smoothed after the summary statistics were frozen, these standard errors are, if anything, too small. The honest summary is that the realized Sharpe ratio is directionally strong but not statistically distinguishable from zero after 45 days; the many-market offline study, not the live run, carries the significance claim.

**Per-market profit and concentration.** Of the $+\$160.67$ net profit, $+\$56.32$ ($+28.2\%$ of capital) was realized on 92 settled or closed markets and $+\$104.35$ (65% of the headline) was the mark-to-

**Table 20:** Standard error and 95% interval of the annualized ($\sqrt{365}$) Sharpe ratio of the live deployment under three estimators ($n = 44$ daily returns, point estimate 4.22).

| Estimator | SE | 95% CI | $t$ |
|---|---|---|---|
| Lo (2002), IID normal | 2.92 | $[-1.49, +9.94]$ | 1.45 |
| Mertens (2002), skew/kurtosis adj. | 1.94 | $[+0.43, +8.02]$ | 2.18 |
| Bootstrap, IID ($B = 10^5$) | 2.28 | $[-1.45, +7.68]$ | — |
| Bootstrap, block ($L = 5$) | 2.11 | $[-0.89, +7.69]$ | — |

market value of 31 positions still open at the end of the window. Gross winning P&L was $+\$330.97$ against gross losses of $-\$170.30$ and $\$22.35$ in fees; 68 markets won and 55 lost, with a median per-market P&L of $+\$0.15$, a maximum of $+\$62.23$ and a minimum of $-\$25.47$. The three most profitable markets account for 54% of gross profit (Herfindahl index of winners 0.11, Gini coefficient of |P&L| 0.75). Removing the single best market leaves $+\$98.44$; removing the best three flips the sign to $-\$17.67$, and on realized P&L alone removing the best market already flips it to $-\$2.95$. A symmetric trim of the three best and three worst markets leaves $+\$48.47$ ($+24.2\%$; $+11.4\%$ on realized P&L), so the book is right-skewed rather than driven by one side. The maximum drawdown was $-15.7\%$ ($-\$31.47$, 24–30 March) on the smoothed NAV and $-20.68\%$ on the pre-smoothing series. Gross traded notional was $\$1{,}101$, or $5.5\times$ capital over 26 trading days (about $21\%$ of capital per day), with peak deployment of $\$196.14$ ($98\%$ of the budget).

**Same-event exposure.** The 123 traded markets belong to 83 distinct events and 62 series; 40 markets (33%) were additional legs on an event already held. The two largest winners are two mutually exclusive legs of the same Survivor-elimination event, which together with a third leg contributed $+\$111.25$, or 69% of the headline profit, with two of the three legs unresolved at the freeze. The effective number of independent bets is therefore well below 123, and the concentration result above is best read as "about two events determine the sign."

**Attribution of the live return.** The score-gap/divergence split reported for the live run in the submitted version ($\Delta S = +72.05\%$, $D = +8.28\%$) is computed per market from the entry forecast and entry price, scaled so that the two terms sum to that market's realized (or, for open positions, marked) P&L, and then summed. It is therefore a P&L-weighted attribution rather than an unweighted measure of forecasting skill, and about two thirds of it is measured against marks rather than outcomes. Re-running the same procedure on the current frozen export gives $\Delta S = +70.33\%$, $D = +10.01\%$: the total is fixed by construction and only the split moves. The unweighted measures point the other way. The forecaster's average Brier score over the traded markets (0.0683) is worse than the market's (0.0606), and a bet-to-bet decomposition that marks the proper position $2(p_t - q_t)$ against the next observed price over the full 4,097-update path gives a negative skill term ($\Delta S = -91.5$, $D = +94.7$ in score units, robust to using mid instead of ask prices). The live profit is real but is best described as a concentrated, partly unrealized outcome of large divergence positions that landed, not as evidence that the forecaster out-predicted the market on average. The main text (Section 4.3) no longer reports this split and describes the deployment as an implementation proof.

### F.5 Expanded evaluation on a second data source

**Data.** Two cuts of the Sooth data exist. This section analyzes the *unconditioned* full set, in which no event or forecast is removed on the basis of its score or return (30,889 events entering the betting analysis across the 18 lanes); the main text (Section 4.1) reports the per-lane *conditioned* cut released during review (29,713 events). The per-lane rule behind that cut was not recorded (Section 4.1). The two cuts share most of their events, but under the full set no lane has a significantly positive score gap, while proper betting remains significantly profitable for four lanes through the divergence term (see Findings below). To test whether the offline findings generalize beyond the Prophet Arena models and time window, we repeat the comparison on forecasts collected through the production forecasting pipeline of Sooth Labs from 17 March 2026 until the export of 1 August 2026. The data comprise 18 model lanes, that is, frontier models run under a fixed harness configuration ("Variant" labels distinguish configurations of the same base model), on Kalshi and Polymarket markets. Each row pairs a forecast with the order-book asks observed at forecast time and the realized outcome, and all rows are used: no event or forecast is removed on the basis of its score or return. Sooth-internal

ensemble and orchestration lanes are excluded. For 15 lanes the export holds one forecast per model and market; for Gemini 3.1 Pro (Variant A), GPT-5.4 and GPT-5.5 (Variant D) it holds every forecast the lane issued (about three per market on average), and each forecast is treated as a separate bet.[8] Rows for Sooth's generated questions (13,005) carry no market quotes and are excluded from every betting analysis. Table 21 summarizes the data. The betting rules, within-spread mask, cost accounting and event-clustered bootstrap are identical to those of Appendix F.1 ($B = 2{,}000$ event resamples; 90% percentile intervals).

**Table 21:** Composition of the expanded dataset by model lane. "Bets" is the number of contracts on which proper (Brier) betting takes a position after the within-spread filter. Rows exceed markets for the three lanes recorded with every forecast.

| Model lane | Events | Markets | Rows | Kalshi / Polymarket | Bets (Brier) |
|---|---|---|---|---|---|
| Gemini 3.1 Pro (Variant B) | 1,651 | 2,940 | 2,940 | 1,869 / 1,071 | 1,937 |
| GPT-5.5 (Variant B) | 1,259 | 2,190 | 2,190 | 1,397 / 793 | 1,414 |
| Gemini 3.1 Pro (Variant A) | 8,160 | 12,327 | 37,306 | 15,866 / 16,493 | 13,937 |
| GPT-5.5 (Variant D) | 6,804 | 10,706 | 30,309 | 14,900 / 11,455 | 12,487 |
| Claude Sonnet 5 | 1,544 | 2,928 | 2,928 | 1,275 / 1,653 | 2,016 |
| GPT-5.5 (Variant C) | 1,693 | 3,172 | 3,172 | 1,822 / 1,350 | 1,744 |
| Claude Fable 5 | 1,851 | 3,371 | 3,371 | 1,604 / 1,767 | 2,080 |
| Gemini 3.5 Flash | 886 | 1,589 | 1,589 | 649 / 940 | 1,054 |
| Kimi K3 | 1,112 | 2,072 | 2,072 | 868 / 1,204 | 1,360 |
| GPT-5.4 | 6,598 | 9,854 | 30,112 | 12,878 / 13,130 | 13,225 |
| GPT-5.6 Sol | 1,400 | 2,616 | 2,616 | 1,102 / 1,514 | 1,634 |
| Muse Spark 1.1 | 1,112 | 2,072 | 2,072 | 866 / 1,206 | 1,347 |
| GLM 5.2 | 3,025 | 5,712 | 5,712 | 3,023 / 2,689 | 4,090 |
| Qwen 3.5+ (Variant B) | 3,719 | 6,811 | 6,811 | 4,327 / 2,484 | 3,853 |
| Claude Sonnet 4.6 (Variant A) | 3,755 | 6,918 | 6,918 | 4,451 / 2,467 | 4,668 |
| Grok 4.5 | 1,546 | 2,932 | 2,932 | 1,277 / 1,655 | 2,095 |
| Claude Sonnet 4.6 (Variant B) | 1,099 | 1,866 | 1,866 | 1,073 / 793 | 1,223 |
| Gemma 4 31B | 2,904 | 5,414 | 5,414 | 3,452 / 1,962 | 3,489 |
| **Total** | | | 150,330 | 72,699 / 64,626 | 73,653 |

**Findings.** No lane beats the market on accuracy: the per-forecast relative Brier score is negative for all 18 lanes, with 90% intervals excluding zero in every case (Table 23). Proper betting nevertheless has a positive point estimate for 11 of the 18 lanes, and its interval excludes zero for four: Gemini 3.1 Pro (Variant B) at $+21.4\%$ $[+11.8, +31.5]$, GPT-5.5 (Variant B) at $+16.8\%$ $[+1.9, +32.6]$, Gemini 3.1 Pro (Variant A) at $+10.0\%$ $[+6.3, +13.8]$, and GPT-5.5 (Variant D) at $+8.2\%$ $[+3.7, +12.8]$. Table 22 shows where this profit comes from: in every lane the score-gap term is negative and the divergence term positive, and the sign of the ROI is decided by which of the two dominates. This is the second regime of the decomposition in Section 3, profit without an accuracy edge, and it appears here at scale; the first regime, profit from accuracy, does not appear in this data. Consistent with Appendix F.2, proper betting is the best rule for the profitable lanes and the paired intervals in Table 24 exclude zero against every heuristic for Gemini 3.1 Pro (Variant B) and against most heuristics for the other three; for the unprofitable lanes proper betting is not distinguishable from, or is worse than, the low-exposure heuristics, and its only consistent significant win is against the high-variance Max (Grp) rule. Sorting lanes by relative Brier score does not recover the sorting by ROI, which is the point of the decomposition.

**Availability.** The per-bet data behind the main-text tables (the conditioned cut) are released at `https://prophets-dataset.onrender.com`; the unconditioned full set behind Tables 21 to 24 is not yet posted there. The data were collected in partnership with Sooth Labs.

[8] Restricting these three lanes to a single forecast per market (the forecast with the median probability) moves their Brier proper-betting ROI from $+10.0$ to $+5.7$ $[+1.3, +10.0]$ for Gemini 3.1 Pro (Variant A), from $+2.8$ to $+0.1$ $[-6.5, +7.0]$ for GPT-5.4, and from $+8.2$ to $+8.7$ $[+3.3, +14.0]$ for GPT-5.5 (Variant D); the signs, the significance of the two profitable lanes, and every conclusion below are unchanged.

**Table 22:** Decomposition of proper-betting ROI (%) into score gap $\Delta S$ and divergence $D$ on the expanded dataset, with event-clustered 90% intervals, one block per scoring rule. Bold marks ROI intervals that exclude zero. Lanes sorted by Brier ROI. The wide single-block version is in `sooth_tables/sooth_fullset_table3.tex`.

| Model | $\Delta S$ | $D$ | ROI | Events |
|---|---|---|---|---|
| *Brier rule* | | | | |
| Gemini 3.1 Pro (Variant B) | $-21.7$ [$-33.2$, $-10.8$] | $+43.1$ [$+36.7$, $+49.6$] | $+21.4$ [$+11.8$, $+31.5$] | 1,154 |
| GPT-5.5 (Variant B) | $-15.4$ [$-29.7$, $-0.6$] | $+32.2$ [$+25.4$, $+39.7$] | $+16.8$ [$+1.9$, $+32.6$] | 842 |
| Gemini 3.1 Pro (Variant A) | $-21.3$ [$-26.7$, $-16.0$] | $+31.3$ [$+28.1$, $+34.9$] | $+10.0$ [$+6.3$, $+13.8$] | 4,194 |
| GPT-5.5 (Variant D) | $-16.6$ [$-21.4$, $-11.8$] | $+24.8$ [$+22.8$, $+26.9$] | $+8.2$ [$+3.7$, $+12.8$] | 3,915 |
| Claude Sonnet 5 | $-17.5$ [$-29.1$, $-5.6$] | $+25.0$ [$+19.0$, $+31.7$] | $+7.4$ [$-3.1$, $+18.9$] | 1,094 |
| GPT-5.5 (Variant C) | $-15.6$ [$-26.2$, $-5.7$] | $+22.9$ [$+18.8$, $+27.3$] | $+7.3$ [$-3.3$, $+17.3$] | 979 |
| Claude Fable 5 | $-17.8$ [$-29.2$, $-6.9$] | $+25.1$ [$+19.7$, $+31.2$] | $+7.3$ [$-4.3$, $+19.0$] | 1,177 |
| Gemini 3.5 Flash | $-26.3$ [$-39.5$, $-13.9$] | $+32.2$ [$+25.6$, $+39.3$] | $+5.9$ [$-3.8$, $+15.5$] | 604 |
| Kimi K3 | $-15.0$ [$-29.1$, $-0.8$] | $+19.9$ [$+15.4$, $+25.4$] | $+4.8$ [$-7.8$, $+18.4$] | 749 |
| GPT-5.4 | $-61.7$ [$-75.4$, $-49.6$] | $+64.5$ [$+56.1$, $+75.4$] | $+2.8$ [$-3.8$, $+9.7$] | 3,615 |
| GPT-5.6 Sol | $-25.8$ [$-38.8$, $-12.6$] | $+27.8$ [$+20.5$, $+36.2$] | $+2.0$ [$-9.0$, $+13.6$] | 910 |
| Muse Spark 1.1 | $-35.3$ [$-50.4$, $-19.5$] | $+34.2$ [$+27.7$, $+41.1$] | $-1.1$ [$-15.2$, $+13.9$] | 756 |
| GLM 5.2 | $-50.3$ [$-62.1$, $-39.1$] | $+48.8$ [$+42.0$, $+55.9$] | $-1.4$ [$-9.4$, $+6.8$] | 2,244 |
| Qwen 3.5+ (Variant B) | $-64.9$ [$-75.3$, $-55.4$] | $+62.9$ [$+56.8$, $+69.4$] | $-2.0$ [$-9.6$, $+5.3$] | 2,211 |
| Claude Sonnet 4.6 (Variant A) | $-50.4$ [$-58.6$, $-42.3$] | $+45.3$ [$+41.0$, $+49.9$] | $-5.1$ [$-11.9$, $+1.7$] | 2,641 |
| Grok 4.5 | $-37.5$ [$-50.8$, $-24.8$] | $+32.0$ [$+24.9$, $+39.6$] | $-5.5$ [$-17.3$, $+6.6$] | 1,134 |
| Claude Sonnet 4.6 (Variant B) | $-38.7$ [$-52.8$, $-24.4$] | $+32.8$ [$+26.8$, $+39.6$] | $-5.8$ [$-19.0$, $+8.5$] | 731 |
| Gemma 4 31B | $-52.4$ [$-61.4$, $-43.3$] | $+45.6$ [$+40.7$, $+50.6$] | $-6.8$ [$-13.3$, $-0.2$] | 1,939 |
| *Logarithmic rule* | | | | |
| Gemini 3.1 Pro (Variant B) | $-10.8$ [$-18.2$, $-4.0$] | $+27.1$ [$+24.2$, $+30.1$] | $+16.3$ [$+8.9$, $+23.2$] | 1,154 |
| GPT-5.5 (Variant B) | $-9.5$ [$-19.9$, $+0.6$] | $+21.2$ [$+18.0$, $+24.9$] | $+11.7$ [$+1.0$, $+22.6$] | 842 |
| Gemini 3.1 Pro (Variant A) | $-12.6$ [$-15.9$, $-9.4$] | $+21.6$ [$+20.4$, $+22.9$] | $+9.1$ [$+6.1$, $+11.8$] | 4,194 |
| GPT-5.5 (Variant D) | $-10.2$ [$-14.1$, $-6.4$] | $+17.4$ [$+16.4$, $+18.4$] | $+7.2$ [$+3.5$, $+10.9$] | 3,915 |
| Claude Sonnet 5 | $-8.9$ [$-16.8$, $-0.8$] | $+15.5$ [$+13.2$, $+18.1$] | $+6.6$ [$-1.5$, $+15.2$] | 1,094 |
| GPT-5.5 (Variant C) | $-13.1$ [$-21.0$, $-5.2$] | $+18.4$ [$+15.4$, $+21.4$] | $+5.2$ [$-3.8$, $+13.8$] | 979 |
| Claude Fable 5 | $-12.9$ [$-20.8$, $-5.1$] | $+17.7$ [$+14.9$, $+20.7$] | $+4.8$ [$-3.3$, $+13.1$] | 1,177 |
| Gemini 3.5 Flash | $-17.4$ [$-26.9$, $-8.1$] | $+25.5$ [$+21.6$, $+29.7$] | $+8.1$ [$-0.5$, $+16.5$] | 604 |
| Kimi K3 | $-7.4$ [$-18.7$, $+4.5$] | $+13.7$ [$+11.7$, $+16.0$] | $+6.4$ [$-4.7$, $+18.2$] | 749 |
| GPT-5.4 | $-29.6$ [$-36.2$, $-23.0$] | $+33.2$ [$+31.3$, $+35.2$] | $+3.6$ [$-2.3$, $+9.7$] | 3,615 |
| GPT-5.6 Sol | $-15.4$ [$-24.8$, $-5.9$] | $+17.8$ [$+14.8$, $+21.1$] | $+2.5$ [$-6.5$, $+11.5$] | 910 |
| Muse Spark 1.1 | $-23.5$ [$-35.1$, $-11.4$] | $+24.4$ [$+20.7$, $+28.1$] | $+0.9$ [$-10.2$, $+12.3$] | 756 |
| GLM 5.2 | $-27.2$ [$-34.9$, $-19.5$] | $+27.9$ [$+25.4$, $+30.6$] | $+0.7$ [$-6.3$, $+7.7$] | 2,244 |
| Qwen 3.5+ (Variant B) | $-41.9$ [$-48.5$, $-35.5$] | $+40.7$ [$+37.9$, $+43.8$] | $-1.2$ [$-7.2$, $+5.2$] | 2,211 |
| Claude Sonnet 4.6 (Variant A) | $-35.9$ [$-41.9$, $-29.8$] | $+31.6$ [$+29.3$, $+34.0$] | $-4.3$ [$-9.9$, $+1.2$] | 2,641 |
| Grok 4.5 | $-25.0$ [$-33.8$, $-16.2$] | $+20.4$ [$+17.4$, $+23.6$] | $-4.6$ [$-13.7$, $+4.5$] | 1,134 |
| Claude Sonnet 4.6 (Variant B) | $-34.3$ [$-45.6$, $-22.7$] | $+28.0$ [$+24.6$, $+31.3$] | $-6.3$ [$-16.8$, $+4.2$] | 731 |
| Gemma 4 31B | $-33.6$ [$-40.1$, $-27.0$] | $+29.8$ [$+27.3$, $+32.5$] | $-3.8$ [$-9.5$, $+1.8$] | 1,939 |
| *Spherical rule* | | | | |
| Gemini 3.1 Pro (Variant B) | $-15.8$ [$-25.3$, $-6.4$] | $+34.5$ [$+29.4$, $+40.0$] | $+18.8$ [$+10.2$, $+27.3$] | 1,238 |
| GPT-5.5 (Variant B) | $-11.5$ [$-23.6$, $+0.5$] | $+24.5$ [$+19.3$, $+30.3$] | $+13.0$ [$+1.3$, $+25.8$] | 876 |
| Gemini 3.1 Pro (Variant A) | $-15.9$ [$-20.2$, $-11.4$] | $+22.7$ [$+20.4$, $+25.1$] | $+6.8$ [$+3.5$, $+10.1$] | 4,260 |
| GPT-5.5 (Variant D) | $-13.8$ [$-17.7$, $-9.8$] | $+18.7$ [$+17.2$, $+20.3$] | $+4.9$ [$+1.4$, $+8.7$] | 3,944 |
| Claude Sonnet 5 | $-15.9$ [$-26.9$, $-5.2$] | $+21.0$ [$+15.8$, $+27.1$] | $+5.1$ [$-5.2$, $+15.5$] | 1,122 |
| GPT-5.5 (Variant C) | $-14.7$ [$-23.7$, $-6.3$] | $+16.6$ [$+13.4$, $+20.1$] | $+1.9$ [$-7.2$, $+10.2$] | 1,092 |
| Claude Fable 5 | $-15.3$ [$-25.2$, $-5.7$] | $+19.1$ [$+15.2$, $+23.8$] | $+3.8$ [$-5.7$, $+13.8$] | 1,259 |
| Gemini 3.5 Flash | $-22.6$ [$-34.6$, $-11.1$] | $+27.5$ [$+22.1$, $+33.3$] | $+4.9$ [$-4.7$, $+14.8$] | 656 |
| Kimi K3 | $-10.4$ [$-23.5$, $+2.5$] | $+14.2$ [$+10.5$, $+18.9$] | $+3.8$ [$-8.3$, $+16.0$] | 771 |
| GPT-5.4 | $-48.4$ [$-59.6$, $-38.3$] | $+50.0$ [$+43.1$, $+58.4$] | $+1.6$ [$-4.3$, $+7.7$] | 3,621 |
| GPT-5.6 Sol | $-21.2$ [$-33.1$, $-9.9$] | $+21.4$ [$+15.7$, $+27.6$] | $+0.1$ [$-9.5$, $+9.7$] | 973 |
| Muse Spark 1.1 | $-31.1$ [$-44.6$, $-17.4$] | $+28.5$ [$+23.2$, $+34.6$] | $-2.6$ [$-14.7$, $+10.2$] | 781 |
| GLM 5.2 | $-38.8$ [$-48.0$, $-30.2$] | $+37.2$ [$+32.1$, $+42.7$] | $-1.6$ [$-8.5$, $+5.4$] | 2,301 |
| Qwen 3.5+ (Variant B) | $-54.2$ [$-62.7$, $-45.8$] | $+48.3$ [$+43.8$, $+53.5$] | $-5.9$ [$-12.5$, $+0.9$] | 2,378 |
| Claude Sonnet 4.6 (Variant A) | $-42.1$ [$-49.6$, $-35.0$] | $+35.8$ [$+32.1$, $+39.7$] | $-6.3$ [$-12.4$, $-0.3$] | 2,748 |
| Grok 4.5 | $-31.1$ [$-42.2$, $-20.3$] | $+25.5$ [$+20.1$, $+31.6$] | $-5.6$ [$-15.7$, $+5.0$] | 1,163 |
| Claude Sonnet 4.6 (Variant B) | $-35.1$ [$-47.4$, $-23.1$] | $+25.8$ [$+21.1$, $+31.2$] | $-9.3$ [$-20.9$, $+2.3$] | 758 |
| Gemma 4 31B | $-41.7$ [$-49.3$, $-34.2$] | $+35.8$ [$+31.8$, $+40.0$] | $-5.9$ [$-11.6$, $-0.3$] | 2,084 |

**Table 23:** ROI (%) of proper (Brier) betting and heuristic rules on the expanded dataset. $\Delta S$ is the per-forecast relative Brier score (market minus model; positive means the model is more accurate). Entropy and Conviction are the Divide-Entropy and Certainty rules of Appendix F.2, restricted to the positive-edge side. Bold marks the best ROI per row.

| | | ROI (%) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | $\Delta S$ | **Proper** | **Max (Mkt)** | **Max (Grp)** | **Inv-Margin** | **Kelly-Alike** | **Entropy** | **Conviction** | **Events** |
| Gemini 3.1 Pro (Variant B) | −0.0154 | **+21.4** | +5.9 | +4.9 | -0.4 | +9.0 | +3.4 | +4.7 | 1,154 |
| GPT-5.5 (Variant B) | −0.0059 | **+16.8** | +3.8 | -3.5 | -1.8 | +8.2 | +2.4 | +4.8 | 842 |
| Gemini 3.1 Pro (Variant A) | −0.0120 | **+10.0** | +3.1 | -7.1 | +3.9 | +5.8 | +2.0 | +3.5 | 4,194 |
| GPT-5.5 (Variant D) | −0.0047 | **+8.2** | +2.3 | -12.0 | -1.3 | +5.4 | +1.1 | +2.0 | 3,915 |
| Claude Sonnet 5 | −0.0055 | +7.4 | +1.7 | -34.6 | -0.1 | **+7.9** | +0.2 | -1.2 | 1,094 |
| GPT-5.5 (Variant C) | −0.0056 | +7.3 | -1.5 | +3.7 | +0.3 | +5.3 | **+7.4** | +2.0 | 979 |
| Claude Fable 5 | −0.0065 | **+7.3** | -1.1 | -21.8 | -5.0 | +4.2 | +0.9 | +0.0 | 1,177 |
| Gemini 3.5 Flash | −0.0188 | +5.9 | **+8.2** | +4.0 | +4.5 | +5.8 | +5.0 | +7.1 | 604 |
| Kimi K3 | −0.0030 | +4.8 | +2.9 | -35.5 | +0.3 | **+6.7** | -1.4 | -2.2 | 749 |
| GPT-5.4 | −0.0194 | +2.8 | -0.8 | -12.3 | -4.8 | **+4.2** | -0.1 | -2.1 | 3,615 |
| GPT-5.6 Sol | −0.0092 | +2.0 | -0.7 | -17.4 | -2.1 | **+3.6** | +0.6 | +0.3 | 910 |
| Muse Spark 1.1 | −0.0133 | -1.1 | **+3.1** | -17.6 | +0.8 | +1.2 | +0.8 | +0.4 | 756 |
| GLM 5.2 | −0.0194 | -1.4 | +0.9 | -24.2 | **+2.9** | +1.6 | -6.7 | -2.2 | 2,244 |
| Qwen 3.5+ (Variant B) | −0.0406 | -2.0 | -4.1 | -4.8 | -2.1 | +0.9 | **+4.1** | -1.1 | 2,211 |
| Claude Sonnet 4.6 (Variant A) | −0.0292 | -5.1 | -3.9 | -12.7 | -4.0 | -2.3 | **-2.2** | -2.8 | 2,641 |
| Grok 4.5 | −0.0115 | -5.5 | -1.0 | -29.5 | -1.1 | -0.4 | **+2.0** | -1.9 | 1,134 |
| Claude Sonnet 4.6 (Variant B) | −0.0236 | -5.8 | -4.9 | **+10.1** | +4.7 | -3.4 | -0.8 | -3.4 | 731 |
| Gemma 4 31B | −0.0339 | -6.8 | -3.5 | -16.9 | **-1.0** | -2.3 | -4.7 | -3.0 | 1,939 |

**Table 24:** Paired event-clustered 90% intervals on the ROI difference between proper betting and each heuristic in Table 23, evaluated on the same event resamples. Bold marks intervals that exclude zero.

| **Model** | − Max (Mkt) | − Max (Grp) | − Inv-Margin | − Kelly-Alike | − Entropy-Scaled | − Conviction-Wt |
|---|---|---|---|---|---|---|
| Gemini 3.1 Pro (Variant B) | +15.5 [+7.6, +24.1] | +16.5 [+4.3, +27.2] | +21.8 [+12.4, +32.2] | +12.4 [+6.2, +19.6] | +18.0 [+9.2, +27.7] | +16.7 [+8.9, +25.5] |
| GPT-5.5 (Variant B) | +13.0 [+0.6, +26.2] | +20.3 [+2.1, +37.4] | +18.6 [+3.0, +35.0] | +8.6 [−1.0, +19.2] | +14.5 [+0.5, +29.5] | +12.0 [−0.6, +25.5] |
| Gemini 3.1 Pro (Variant A) | +6.9 [+4.0, +10.2] | +17.1 [+12.2, +22.0] | +6.1 [−1.1, +12.0] | +4.2 [+1.5, +7.2] | +8.0 [+4.4, +11.8] | +6.5 [+3.4, +9.9] |
| GPT-5.5 (Variant D) | +6.0 [+2.6, +9.4] | +20.2 [+14.7, +25.6] | +9.5 [+4.8, +14.3] | +2.9 [−0.3, +6.1] | +7.1 [+2.7, +11.4] | +6.2 [+2.4, +9.9] |
| Claude Sonnet 5 | +5.7 [−2.9, +15.6] | +42.1 [+30.5, +53.8] | +7.5 [−3.6, +19.5] | −0.5 [−6.1, +6.0] | +7.3 [−1.1, +16.5] | +8.6 [+0.1, +18.0] |
| GPT-5.5 (Variant C) | +8.8 [+0.1, +17.3] | +3.6 [−23.3, +24.3] | +7.0 [−3.6, +17.6] | +2.0 [−5.1, +8.6] | −0.1 [−9.3, +9.4] | +5.3 [−3.5, +13.6] |
| Claude Fable 5 | +8.4 [−1.3, +19.4] | +29.1 [+17.4, +40.8] | +12.3 [+0.9, +24.9] | +3.0 [−4.7, +11.4] | +6.3 [−3.6, +17.1] | +7.3 [−2.5, +17.8] |
| Gemini 3.5 Flash | −2.4 [−10.3, +5.5] | +1.8 [−15.4, +17.1] | +1.3 [−8.7, +11.3] | +0.1 [−6.2, +6.4] | +0.8 [−7.8, +9.3] | −1.2 [−9.0, +6.5] |
| Kimi K3 | +1.9 [−8.6, +13.1] | +40.3 [+26.9, +53.9] | +4.5 [−9.6, +18.9] | −1.9 [−9.6, +6.4] | +6.2 [−6.1, +19.2] | +7.0 [−4.8, +19.5] |
| GPT-5.4 | +3.6 [−1.0, +8.2] | +15.1 [+6.1, +23.6] | +7.6 [+1.2, +14.5] | −1.4 [−4.0, +1.4] | +2.9 [−2.6, +8.7] | +4.9 [+0.0, +10.0] |
| GPT-5.6 Sol | +2.6 [−6.2, +12.6] | +19.4 [+5.9, +32.4] | +4.0 [−6.9, +15.7] | −1.7 [−9.4, +6.6] | +1.4 [−9.0, +12.0] | +1.6 [−8.2, +11.6] |
| Muse Spark 1.1 | −4.2 [−16.3, +8.9] | +16.5 [+0.0, +32.9] | −1.9 [−16.7, +13.5] | −2.3 [−9.9, +6.1] | −1.9 [−14.6, +11.3] | −1.5 [−13.0, +11.2] |
| GLM 5.2 | −2.3 [−8.7, +4.4] | +22.7 [+11.8, +33.3] | −4.3 [−13.0, +4.8] | −3.0 [−6.4, +0.7] | +5.2 [−5.8, +17.2] | +0.8 [−5.7, +7.4] |
| Qwen 3.5+ (Variant B) | +2.0 [−4.5, +8.4] | +2.8 [−10.8, +15.4] | +0.1 [−8.2, +7.9] | −3.0 [−7.3, +1.2] | −6.2 [−13.1, +0.6] | −0.9 [−7.1, +4.9] |
| Claude Sonnet 4.6 (Variant A) | −1.2 [−7.0, +4.3] | +7.6 [−3.6, +17.7] | −1.1 [−9.3, +7.0] | −2.8 [−6.0, +0.4] | −2.9 [−8.3, +2.8] | −2.4 [−7.7, +3.0] |
| Grok 4.5 | −4.6 [−14.0, +5.8] | +24.0 [+10.2, +37.2] | −4.4 [−16.4, +8.7] | −5.1 [−10.8, +1.3] | −7.5 [−16.2, +2.1] | −3.6 [−13.0, +6.9] |
| Claude Sonnet 4.6 (Variant B) | −0.9 [−11.8, +10.4] | −16.0 [−55.9, +15.5] | −10.5 [−33.1, +10.0] | −2.4 [−8.1, +4.5] | −5.0 [−16.0, +6.4] | −2.4 [−12.4, +8.7] |
| Gemma 4 31B | −3.3 [−8.4, +2.0] | +10.1 [+1.1, +18.7] | −5.8 [−12.6, +1.0] | −4.5 [−7.9, −1.0] | −2.1 [−8.8, +4.8] | −3.8 [−8.8, +1.4] |

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

1. **Claims**

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The abstract and Section 1 state the contributions: (i) a proper betting strategy whose expected profit decomposes into score gap, Bregman divergence, and liquidity loss, guaranteeing robust profitability (Theorem 1); (ii) the essential uniqueness of proper betting among robustly profitable strategies (Theorem 2) and the resulting characterization of strictly proper scoring rules (Proposition 2); and (iii) an empirical evaluation on thousands of AI forecasts in Section 4, including a month-long live pilot on Kalshi (Section 4.3). The abstract states that the result is a robust profitability guarantee, not profit maximization or an accuracy–profit equivalence, and presents the live pilot as an existence proof. Both cuts of the expanded dataset are reported (Section 4.1 and Appendix F.5), and the live pilot is described as an implementation proof.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: A consolidated Limitations paragraph is given in Section 5. Practical limitations are discussed throughout the paper. The dependence of the guarantee of Theorem 1 on sufficient market liquidity is stated in the theorem and discussed in Section 2; the effect of bid–ask spreads is analyzed in Section 3.3 and formalized in Appendix D.3; the guarantees concern expected rather than risk-adjusted profit (Section 5); and the live pilot is explicitly not intended to establish statistically reliable trading alpha (Section 4.3).

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Every theoretical result is stated with explicit assumptions in Section 3. Full proofs are deferred to Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 specifies the data sources (forecasts on Kalshi and Polymarket markets collected in collaboration with Sooth Labs, and the Prophet Arena pipeline), the construction of forecast–price–outcome tuples, and the evaluation pipeline. Implementation details for all betting strategies are in Appendix E.2; the synthetic-persona generation is described in Appendix E.4. Live-deployment details are documented in Appendix E.7.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release the betting data behind the main-text tables (linked in Section 4.1) and the trading history of the live pilot (linked in Section 4.3 and Appendix E.7). The underlying market data is available through the public APIs of Kalshi and Polymarket.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The data sources, forecast–price–outcome construction, and evaluation procedure are described in Section 4. Appendix E.4 describes the full synthetic-forecast generation process. The live-trading agent's parameters are documented in Appendix E.7.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Tables 2 and 3 and their full versions in Appendices E.2 and E.3 report paired, event-clustered 95% bootstrap confidence intervals. The live pilot reports the calendar-day Sharpe ratio with a 95% block-bootstrap confidence interval, which includes zero (Section 4.3, Table 9). Intervals for the submitted Prophet Arena tables and the live robustness statistics are in Appendix F.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All offline experiments run on pretrained-LLM forecast outputs and CPU-only post-processing, requiring no specialized compute. The live-trading worker described in Appendix E.7 runs on a single starter-tier cloud instance. No model training was performed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work involves prediction markets and automated trading strategies. Potential implications include market efficiency, financial risk, and misuse of automated agents. The paper focuses on improving understanding of forecasting and market behavior rather than promoting speculative trading.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The work has implications for prediction markets and AI forecasting. It highlights how accurate AI forecasts can be translated into economic value and may improve market efficiency.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not release any high-risk assets such as pretrained models, generative systems, or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper uses existing datasets and benchmarks, including Prophet Arena and prediction market data (Kalshi and Polymarket), with forecasts collected in collaboration with Sooth Labs. These sources are properly cited in the paper. For sourcing market data, we rely only on publicly available data and respect the terms of use of these platforms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces a live trading agent that utilizes our introduced proper betting strategies. Its design is documented in Appendix E.7. The trading logs are released in the same appendix section.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or experiments with human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human subjects or data collection from individuals.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are a core component of the empirical evaluation in Sections 4, 4.1 and 4.2. The paper analyzes forecasts produced by frontier LLMs and evaluates their performance against prediction-market prices.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.