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 a clean equivalence between forecasting accuracy and trading profit, but only for the specific automated market maker (AMM) design. However, the largest exchanges today are based on central limit order books in which informed forecasters routinely lose money while uninformed strategies can profit on simple heuristics. We resolve this discrepancy by establishing a formal equivalence between predictive accuracy and profitability. For any strictly proper scoring rule S, we exhibit a "proper" betting strategy that depends only on the forecaster's prediction $\mathbf{p}$ and the market price $\mathbf{q}$, and earns positive expected profit *whenever* $\mathbf{p}$ outperforms $\mathbf{q}$ under S and the market has sufficient liquidity. Moreover, this proper betting is essentially the only strategy with such robust profitability guarantee. The proof rests on a decomposition of expected profit that strictly generalizes the classical AMM guarantee and also explains how strategies can profit without an accuracy edge. Empirically, across thousands of forecasts 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. A month-long live deployment on Kalshi achieves $+80.33\%$ return on investment with a Sharpe ratio of 3.35.

1 Introduction

Prediction markets let people trade contracts whose prices reflect the market's estimated probability of an event, aggregating dispersed beliefs in the spirit of Hayek [1945]'s "marvel of the price system." Empirical studies confirm that real prediction-market prices are well calibrated and often outperform polls and expert panels [Berg et al., 2008, Rothschild, 2009, Wolfers 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 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 would 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

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

Preprint.

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 A). Prices arise from matching opposing limit orders, and a forecaster can pick any bet size but faces whatever price impact the order book happens to deliver, plus bid-ask spreads, finite liquidity, and platform fees. The link between accuracy and profit is no longer pinned down by the price-formation rule, and converting accuracy into profit becomes a separate problem.

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 leaderboards 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]. One may attribute the failure to the market frictions, but even in the idealized setting (with no price impact, bid-ask spread, or fees), we can observe that natural betting heuristics would produce negative profit from reasonably good forecasts (see the failure of Kelly in Example 1, the failure of max-margin-based betting in Example 2 and more empirical results in Section 4.1). Stranger still, the converse also fails: simple strategies earn positive expected profit even when the underlying forecaster has no accuracy edge at all (see Example 3). Translating accuracy into profit is therefore a problem in its own right, distinct from forecasting accuracy.

Main contributions. This paper establishes a link between predictive accuracy and trading profitability in general prediction markets with an arbitrary price-impact function. First, we exhibit a betting strategy that converts an accuracy edge over market into profits:

Theorem (Informal version of Theorem 1)**.** *If a forecaster can outperform a prediction market under a strictly proper scoring rule S, then there exists a specific "proper" betting strategy associated to S that 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). Furthermore, when instantiated in AMM markets, our proper betting strategy corresponds precisely to the canonic betting strategy in AMM that moves market price to the forecaster's belief. The above result then recovers the classic AMM guarantee as a boundary case where the Bregman divergence exactly absorbs liquidity loss.

Our second main result establishes that the above proper betting is essentially the only betting strategy that robustly guarantees profitability on any sufficiently liquid market — in the sense of ensuring positive profit *whenever* there is a score edge.

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

Our results also yield a new characterization of proper scoring rules: a scoring rule is proper *if and only if* there exists an associated betting strategy that is robustly profitable (Proposition 2). Finally, we further extend our results to three practically important settings: betting on a sequence of events based on empirical scores (instead of the theoretical expectation), betting under nonzero bid-ask spreads, and long-horizon trading in which the forecaster can buy and sell the same event's contracts at different time as both the forecaster's and market's predictions evolve.

Because testing our strategies requires vast amounts of prediction data, we experimentally evaluate our proper betting strategies on thousands of AI-generated forecasts and find a clear advantage over standard heuristic baselines. We identify systematic forecasting personas across models, with different personas favoring different proper betting strategies. Finally, to demonstrate the practical value of our theory, we deployed an AI forecasting agent with real capital on the Kalshi prediction market for a month. Our agent achieved an impressive +80.33% return on investment (ROI) with a Sharpe ratio of 3.35.

1.1 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. 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 margin. Our work inherently is also about betting strategy design, but ours differs from Kelly in a key aspect: our strategy is oblivious to the ground truth probabilities and comes with robustness guarantees, whereas 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].

2 Preliminaries

Notation. Define $[K] := \{1, \ldots, K\}$. 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 y-th entry, $\mathbf{1} \in \mathbb{R}^k$ the vector with every entry equal to 1, ∇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 forecaster's 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 $\mathbf{s} \cdot \mathbf{1}_y = s_y$.

For ease of presentation, we assume the price vector lies on the probability simplex, $\mathbf{q} \in \Delta_{[K]}$, equivalently $\sum_{k\in[K]} q_k = 1$. Under this condition, the price vector $\mathbf{q}$ admits a natural interpretation as the market's *consensus belief* [Arrow et al., 2008]. Thus a forecaster who believes that outcome k realizes with probability exactly q_k has zero expected return from this contract, i.e., $-q_k + q_k \cdot 1 = 0$. We note that any arbitrage-free market must have $\sum_{k\in[K]} q_k = 1$, otherwise a trader can obtain a riskless profit of $\left|1 - \sum_{k\in[K]} q_k\right|$ by buying or selling one share of every contract. Real markets, however, sidestep this by maintaining separate price for bid/buy and ask/sell, the gap of which is called the *spread*. Section 3.3 will discuss how our main results naturally extend to the case of non-zero bid-ask spread with minor modifications.

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 vector* $\mathbf{s}$, satisfying the boundary condition $\rho(\mathbf{0}) = \mathbf{q}$ and the monotonicity condition $(\rho(\mathbf{s})-\rho(\mathbf{s}'))\cdot(\mathbf{s}-\mathbf{s}') \geq 0, \forall \mathbf{s}, \mathbf{s}' \in \mathbb{R}^K$, i.e., walking the order book in any direction shall move 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 can be decomposed into the spot-price piece $\mathbf{s} \cdot \mathbf{q}$ plus a non-negative *liquidity loss* (slippage) capturing the price impact along the trajectory:[2]

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

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

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

Remark 1. *The form of ρ depends on the market design: the classic automated market makers (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 ρ is piecewise constant and read directly off the order book [Kalshi, 2025, Polymarket, 2023]. We discuss these two main prediction-market implementations in detail in Appendix A, but our theory is applicable to any design. We also note that ρ itself evolves over time as the order book updates with new resting orders and executions; we suppress this time dependence since each of our analyses concerns a single instantaneous 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 forecaster's 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 the forecast $\mathbf{p}$ based on y, with expected score $S(\mathbf{p}; \mathbf{p}^*) := \mathbb{E}_{y \sim \mathbf{p}^*}[S(\mathbf{p}, y)]$ under a ground truth $\mathbf{p}^*$. Scoring rules were introduced to truthfully elicit forecasts; a scoring rule is said to be *proper* if it incentivizes a forecaster to report their true distributional belief $\mathbf{p}$.

Definition 1 (Proper scoring rule)**.** *A scoring rule S is* proper *if a forecaster maximizes her expected score by reporting her true belief $\mathbf{p}$. Formally, for every $\mathbf{p}, \mathbf{p}' \in \Delta_{[K]}$, $S(\mathbf{p}; \mathbf{p}) \geq S(\mathbf{p}'; \mathbf{p})$. We say the rule is* strictly proper *if the inequality is strict for $\mathbf{p}' \neq \mathbf{p}$.*

A proper scoring rule lets us compare any two probabilistic forecasts under some ground truth $\mathbf{p}^*$. For example, 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, the comparison of scores depends on the ground truth $\mathbf{p}^*$, which can never be observed in most market applications. Section 3.3 discusses how to compare forecasts based on empirical score $S(\mathbf{p}; y)$ on realized outcome y, averaged across markets.

Every proper scoring rule corresponds to a convex *potential function* G as stated in a folklore result below.

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 $G : \Delta_K \to \mathbb{R}$ associated to S such that*

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

Meanwhile, the *Bregman divergence* is defined for every strictly convex differentiable function G as

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

where $D_G(\mathbf{q}, \mathbf{p}) \geq 0$ with equality if and only if $\mathbf{p} = \mathbf{q}$. Table 1 below lists examples of proper scoring rules along their associated convex functions (the last column s_G is introduced in Definition 2).

[2] Due to monotonicity condition of $\rho(\mathbf{s})$, $\int_0^1 \rho(t\mathbf{s}) \cdot \mathbf{s}\, dt$ is the least cost for purchasing $\mathbf{s}$ shares even $\rho(\mathbf{s})$ may not necessarily be a conservative vector field.

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

The goal of this paper is to translate a forecaster's prediction $\mathbf{p}$ 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 Kelly criterion [Kelly, 1956]. The *highest-margin* strategy places a bet on the outcome k for which $|p_k - q_k|$ is greatest, as this outcome intuitively has the most potential for profit: the difference between the perceived value p_k of the share and the actual price q_k is greatest. Perhaps surprisingly, we show below that both the Kelly criterion and the highest-margin strategy do *not* guarantee positive expected profit, thus demonstrating that betting optimization is challenging even for a good forecaster. In Section 4, we also empirically observe that all of these approaches 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 would spend* p_i *to purchase* p_i/q_i *shares of contract* i *for each* $i = 1, 2, 3$*,*[3] *and its expected profit is* $(\mathbf{p} \oslash \mathbf{q}) \cdot (\mathbf{p}^* - \mathbf{q}) < 0$*. The intrinsic reason of this loss is that the Kelly criterion is derived by assuming its forecast is the ground truth, which in practice can be very off from the forecast.*

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

This paper resolves this seemingly broken link between statistical accuracy and profitability in the examples above. We first define, for every proper scoring rule S, what we call a *proper* betting strategy corresponding to S.

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*

$$\mathbf{s}_G(\mathbf{p}, \mathbf{q}) := \nabla G(\mathbf{p}) - \nabla G(\mathbf{q}).$$

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

[3] This is the generalized solution for many-outcome Kelly criterion. The familiar two-outcome Kelly formula of spending $\frac{p_1 - q_1}{1 - q_1}$ to bet on *outcome* 1 (short 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* is equivalent to additionally buying $\frac{1 - p_1}{1 - q_1}$ shares of both *outcome 1* at price q_1 and *2* at price $1 - q_1$, leading to total bet amount p_1, p_2 on *outcome* $1, 2$ respectively.

3.1 Proper betting and its robust profitability

On robust profitability. We now establish a formal link between a good forecast that outperforms the market and its betting profitability. A fundamental challenge in designing good betting strategies is that we almost never observe the ground truth probability $\boldsymbol{p}^*$. Even when a forecast $\boldsymbol{p}$ is promised to have better score than the market $\boldsymbol{q}$ — in the sense that $S(\mathbf{p};\mathbf{p}^*) \geq S(\mathbf{q};\mathbf{p}^*)$ for some proper scoring rule S — this can hold true for infinitely many possible $\mathbf{p}^*$'s. Thus, a desirable property of a good betting strategy is *robust profitability*: it is guaranteed to have positive profit for any $\mathbf{p}, \boldsymbol{p}^*, \mathbf{q}$ satisfying $S(\mathbf{p};\mathbf{p}^*) \geq S(\mathbf{q};\mathbf{p}^*)$. Conversely, we say a strategy $\boldsymbol{s}(\mathbf{p},\mathbf{q})$ is not robustly profitable under S, if there exists some $\mathbf{p} \neq \mathbf{q}, \mathbf{p}^*$ such that $S(\mathbf{p};\mathbf{p}^*) \geq S(\mathbf{q};\mathbf{p}^*)$ but $\boldsymbol{s}(\mathbf{p},\mathbf{q})$ has non-positive expected profit under $\mathbf{p}^*$. Note that "profitability" here is a binary notion: it concerns only whether positive profit can be guaranteed, not the magnitude of that profit.

Our first main result shows that robustly profitable betting strategy exists whenever the market's liquidity loss is relatively small; in fact, the proper betting $\boldsymbol{s}_G(\mathbf{p},\mathbf{q})$ as defined in Definition 2 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 $\pi(\mathbf{s}^*,\mathbf{p}^*)$ of the proper betting strategy $\mathbf{s}^* = \mathbf{s}_G(\mathbf{p},\mathbf{q})$ admits the following decomposition:*

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

Therefore, if $D_G(\mathbf{q},\mathbf{p}) \geq L_\rho(\mathbf{s}^*;\mathbf{q})$ — i.e., the divergence offsets liquidity loss — then proper betting is robustly profitable *in the sense that it ensures positive profit whenever $S(\mathbf{p};\mathbf{p}^*) > S(\mathbf{q};\mathbf{p}^*)$.*

Proof. The proof hinges on an interesting profit decomposition lemma below that 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), we have*

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

Proof of Lemma 1. We express

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

where

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

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

Taking expectation of Equation (2) over $y \sim \mathbf{p}^*$ yields the (frictionless) expected profit of the proper bet,

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

Notably, the term $\mathbf{s}_G(\mathbf{p},\mathbf{q})\cdot(\mathbf{p}^* - \mathbf{q})$ is not the expected profit $\pi(\mathbf{s}^*,\mathbf{p}^*)$ yet since it counts the betting cost using the same market price $\mathbf{q}$ for purchasing $\mathbf{s}_G(\mathbf{p},\mathbf{q})$ contract shares, ignoring the slippage. This gap is captured precisely by the third term in Eq. (1) $L_\rho(\mathbf{s};\mathbf{q}) = \int_0^1 \rho(t\mathbf{s})\cdot\mathbf{s}\,dt - \mathbf{s}\cdot\mathbf{q}$. These together prove the theorem. □

Like the profit decomposition of Lemma 1, Theorem 1 also holds "point-wise" for each realized outcome $y \sim \mathbf{p}^*$. Indeed, this corresponds to the special case of $\mathbf{p}^* = \mathbf{1}_y$. These point-wise equalities will allow us to generalize the guarantee of proper betting to empirical scores as discussed in Section 3.3.3.

Remark 2. *The key insight from Theorem 1 is to realize 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 of Lemma 1 formalizes a natural intuition: a profitable forecast is not merely about being accurate (i.e., having big score gap), but is about sufficiently different from the market yet still remains sufficiently accurate. Conversely, positive profit does not imply accuracy: it can be attributed entirely to the Bregman bonus. This also explains the paradox in Example 3: when a forecaster's prediction is far from the market price* $\mathbf{q}$*, a large Bregman divergence can outweigh a negative score gap.*

Remark 3 (Beyond simplexes)**.** *Notably, Theorem 1 holds even when* $\mathbf{q} \notin \Delta_K$*. As an interesting application, we can also apply it when* $Q := \sum_{k=1}^K q_k < 1$*, i.e., there is trivial arbitrage opportunity by buying* $\mathbf{s} = \mathbf{1}$ *at cost* $Q(< 1)$ *with a deterministic return* 1 *(assuming little liquidity loss). Theorem 1 captures this situation as a special case. By choosing* $G(\mathbf{q}) = (\sum_{k=1}^K q_k)^2$*, we can verify that the proper betting* $\mathbf{s}_G = 2[\sum_{k=1}^K (p_k - q_k)] \cdot \mathbf{1}$*, leading to deterministic profit* $2(1-Q)^2$ *for any* $\mathbf{p}^*$*. It can be further verified that both Bregman divergence and score gap term equal* $(1-Q)^2$ *in this case, summing up to the* $2(1-Q)^2$ *profit.*

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 further shows that it is the only robustly profitable strategy, essentially — the reason of "essentially" is to rule out trivial transformations such as rescaling the strategy (i.e., $\lambda \boldsymbol{s}_G(\mathbf{p}, \mathbf{q})$) or adding a constant shift (i.e., $\boldsymbol{s}_G(\mathbf{p}, \mathbf{q}) + \lambda \mathbf{1}$), which will not change the profitability nature of the strategy.

Definition 3 (Essentially different betting)**.** *We say a betting strategy* $\boldsymbol{s}(\mathbf{p}, \mathbf{q})$ *is* essentially different *from another strategy* $\boldsymbol{s}'(\mathbf{p}, \mathbf{q})$ *if, after any rescaling of* $\boldsymbol{s}(\mathbf{p}, \mathbf{q})$ *via mapping* $\boldsymbol{s} \to \lambda \boldsymbol{s}$ *and constant shifting via* $\boldsymbol{s} \to \boldsymbol{s} + \lambda \mathbf{1}$*, there always exists an interior market price* $\mathbf{q} \in \mathrm{int}(\Delta_K)$ *such that* $||\boldsymbol{s}(\mathbf{p}_t, \mathbf{q}) - \boldsymbol{s}'(\mathbf{p}_t, \mathbf{q})|| \geq \epsilon$ *for some forecast sequence* $\{\mathbf{p}_t\}_{t=1}^\infty$ *converging to* $\mathbf{q}$*.*

If two strategies are not essentially different, we say they are essentially the same*.*

It is natural that rescaling and constant shifting should not lead to an intrinsically different betting strategy: rescaling only proportionally changes betting sizes, whereas placing the all-one bet $\mathbf{1}$ costs $\mathbf{1} \cdot \mathbf{q} = 1$ and pays off 1 deterministically. The only non-trivial requirement for essentially different betting strategies is that they need to have non-negligible difference (i.e., at least ϵ difference) for some sequence of forecasts $\{\mathbf{p}_t\}_{t=1}^\infty$ converging to $\mathbf{q}$. This is a natural (and also necessary) requirement because otherwise two robustly profitable betting strategies can be superficially different. For instance, given any robustly profitable strategy, e.g., our proper betting $\boldsymbol{s}_G(\mathbf{p}, \mathbf{q})$, one can construct $\boldsymbol{s}(\mathbf{p}, \mathbf{q}) = \boldsymbol{s}_G(\mathbf{p}, \mathbf{q}) + \epsilon\, \mathbf{e}_1$ simply by additionally buying ϵ amount of the first contract for some ϵ smaller than the profit of $\boldsymbol{s}_G(\mathbf{p}, \mathbf{q})$, which is at least the Bregman divergence $D_G(\mathbf{q}, \mathbf{p})$ as we will show. This negligible modification can be done even for infinitely many pairs of $\mathbf{p}, \mathbf{q}$ so long as they are all separated apart by at least a constant distance. Thus the real test of difference is at the limit when $\mathbf{p} \to \mathbf{q}$. Definition 3 is introduced precisely for that: two strategies $\boldsymbol{s}, \boldsymbol{s}'$ are essentially different if their difference $\boldsymbol{s}(\mathbf{p}_t, \mathbf{q}) - \boldsymbol{s}'(\mathbf{p}_t, \mathbf{q})$ does not converge to 0 – at least for some market price $\mathbf{q}$ and some sequence of forecasts $\{\mathbf{p}_t\}_{t=1}^\infty$ converging to $\mathbf{q}$.

Our second main theorem is stated below.

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 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 exists* $\mathbf{p}, \mathbf{p}^*, \mathbf{q}$ *such that* $S(\mathbf{p}, \mathbf{p}^*) > S(\mathbf{q}, \mathbf{p}^*)$ *but the expected profit of* $\mathbf{s}(\mathbf{p}, \mathbf{q})$ *is strictly negative. This holds even for frictionless markets, i.e.,* $L_\rho = 0$*.*

Proof sketch of Theorem 2. The proof is somewhat technical. We defer full proof to Appendix B.1 and only overview the proof ideas here. First, observe that it suffices to prove the theorem for frictionless markets since if any betting strategy is robustly profitable, then it must remain profitable in a frictionless market simply due to lower cost. Our proof thus argues that any robustly profitable betting strategy in a frictionless market is essentially the same as proper betting. Given a betting strategy $\mathbf{s}(\mathbf{p}_t, \mathbf{q})$ that is essentially different from proper betting $\mathbf{s}_G(\mathbf{p}_t, \mathbf{q})$, our proof here features an explicit construction of an infinite sequence of tuple $(\mathbf{p}_t, \mathbf{p}_t^*, \mathbf{q})$ and then leverages convergence properties to prove the existence of tuples in this sequences that satisfies $S(\mathbf{p}_t, \mathbf{p}_t^*) - S(\mathbf{q}, \mathbf{p}_t^*) > 0$ but the expected profit of $\mathbf{s}(\mathbf{p}_t, \mathbf{q})$ under the constructed $\mathbf{p}_t^*$ is strictly negative. □

A simple special case of Theorem 1 is for the $K = 2$ case, i.e., binary event betting. Note that rescaling via mapping $\boldsymbol{s} \to \lambda \boldsymbol{s}$ and constant shifting via $\boldsymbol{s} \to \boldsymbol{s} + \lambda \mathbf{1}$ will not change a betting strategy $\boldsymbol{s}$'s robust profitability. Then any betting strategy $\mathbf{s}$, normalized to have $||\boldsymbol{s}|| = 1$ and $\boldsymbol{s} \cdot \mathbf{1} = 0$, is either $s^+ = (1/\sqrt{2}, -1/\sqrt{2})^\top$ (i.e., buying $\sqrt{2}$ of *YES*) or $s^- = (-1/\sqrt{2}, 1/\sqrt{2})^\top$ (i.e., buying $\sqrt{2}$ of *NO*). If betting strategy $\boldsymbol{s}$ is essentially different from proper betting $\boldsymbol{s}_G$, then there exists $\boldsymbol{q}$ and a sequence $\{\boldsymbol{p}_t\}_{t=1}^\infty$ converging to $\boldsymbol{q}$ such that the *normalized* $\mathbf{s}(\mathbf{p}_t, \mathbf{q}), \mathbf{s}_G(\mathbf{p}_t, \mathbf{q})$ satisfy $||\mathbf{s}(\mathbf{p}_t, \mathbf{q}) - \mathbf{s}_G(\mathbf{p}_t, \mathbf{q})|| \geq \epsilon$ for some constant ϵ. This implies that $\mathbf{s}(\mathbf{p}_t, \mathbf{q}), \mathbf{s}_G(\mathbf{p}_t, \mathbf{q})$ must equal s^+, s^- respectively. Therefore, if $\mathbf{s}_G(\mathbf{p}_t, \mathbf{q})$ is profitable, its opposite purchase behavior $\mathbf{s}(\mathbf{p}_t, \mathbf{q})$ cannot be profitable, aligning with the theorem's statement. While this case with $K = 2$ is easy to see, the case with $K > 2$ becomes much more non-trivial since there are infinitely many normalized betting strategies. The formal proof in Appendix B.1 needs to leverage the geometry of the normalized strategy space.

3.3 Implications and real market considerations

3.3.1 A new characterization of proper scoring rules via betting profitability.

A keen reader might observe that the proper betting strategy $\mathbf{s}_G(\mathbf{p}, \mathbf{q})$ is well-defined for any G, not only those G that are convex and correspond to proper scoring rules. Is properness of the scoring rule necessary to guarantee the robust profitability of $\mathbf{s}_G(\mathbf{p}, \mathbf{q})$? Our next proposition show that properness of the scoring rule is indeed necessary, in a very strong sense. That is, if $\mathbf{p}, \mathbf{q}$ are compared by some scoring rule S satisfying $S(\mathbf{p}; \mathbf{p}^*) > S(\mathbf{q}; \mathbf{p}^*)$, the properness of S is necessary to guarantee the existence of a robustly profitable betting strategy (and when it exists, proper betting is one such 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 robustly profitable betting strategy $\mathbf{s}(\mathbf{p}, \mathbf{q})$ that guarantees positive profit in a frictionless market (i.e., $L_\rho = 0$) for any $\boldsymbol{p}^*$ and $\mathbf{p} \neq \mathbf{q}$ satisfying $S(\mathbf{p}; \mathbf{p}^*) \geq S(\mathbf{q}; \mathbf{p}^*)$.*

Proposition 2, proven in Appendix B.2, thus yields a novel characterization of proper scoring rules as precisely those rules that yield robustly profitable betting strategy when a forecast $\mathbf{p}$ outperforms market $\mathbf{q}$ under the scoring rule.

3.3.2 Proper betting is a strict generalization of canonical betting in AMMs

All our results thus far are applicable to any prediction market. Here we illustrate what proper betting strategy corresponds to in the special case of *automated market makers* (AMMs) and show that it is a strict generalization of the canonical betting strategy in AMMs.

An AMM is governed by a strictly convex cost function $C(\mathbf{x})$ such that the current market price $\mathbf{q} \in \Delta_K$ and net number of shares of each contract $\mathbf{x} \in \mathbb{R}^K$ always satisfies $\nabla C(\mathbf{x}) = \mathbf{q}$ [Hanson, 2007, Chen and Pennock, 2007]. A key property of AMMs is that if a forecaster's belief is some $\mathbf{p}$, then this forecaster maximizes their expected profit by purchasing shares $\mathbf{s}_{\mathbf{q}\to\mathbf{p}}$ that moves the market price from $\mathbf{q}$ to $\mathbf{p}$, i.e., $\nabla C(\mathbf{x} + \mathbf{s}_{\mathbf{q}\to\mathbf{p}}) = \mathbf{p}$ (see Appendix A for more details). Interestingly,

we observe that the above canonical betting strategy in AMMs corresponds precisely to the proper betting strategy of Definition 2 under the corresponding market scoring rule, or formally stated below

Fact 1. *In automated market makers, we have* $\mathbf{s}_{\mathbf{q}\to\mathbf{p}} = \mathbf{s}_G(\mathbf{p}, \mathbf{q})$.

To see this, we recall that the convex conjugate of the cost function $C(\mathbf{x})$ of an AMM, $G(\mathbf{p}) = \sup_{\mathbf{x}\in\mathbb{R}^K} \{\mathbf{p}\cdot\mathbf{x} - C(\mathbf{x})\}$, is precisely the convex potential that induces the market scoring rule $S(\mathbf{p}, y) = G(\mathbf{p}) + \nabla G(\mathbf{p})\cdot(\mathbf{1}_y - \mathbf{p})$. A basic fact about convex conjugates is that the inverse of the gradient function $\nabla G : \Delta_K \to \mathbb{R}^K$ is precisely the $\nabla C(\mathbf{x})$ function. Observing $\mathbf{x} + \mathbf{s}_{\mathbf{q}\to\mathbf{p}} = (\nabla C)^{-1}(\mathbf{p})$, we have the betting strategy

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

A basic fact in AMMs is that the profit $\pi(\mathbf{p}, \mathbf{q})$ for moving the market from current price $\mathbf{q}$ to new price $\mathbf{p}$ is $S(\mathbf{p}, \mathbf{p}^*) - S(\mathbf{q}, \mathbf{p}^*)$ for ground truth distribution $\mathbf{p}^*$ [Hanson, 2007], which corresponds precisely to the first term of Theorem 1. This is not a coincidence since, as it turns out, the divergence term and liquidity loss term in Theorem 1 cancel out precisely in any AMM, as stated below (proved in Appendix B.3).

Corollary 1 (Instantiation of Theorem 1 in AMMs)**.** *Consider any AMM governed by a strictly convex cost function* $C(\mathbf{x})$ *and let* S *denote its corresponding market scoring rule determined by* C*'s convex conjugate* $G(\mathbf{p})$*. Then we have* $D_G(\mathbf{q}, \mathbf{p}) = L_\rho(\mathbf{s}^*; \mathbf{q})$ *for the proper betting strategy* $\mathbf{s}^* = \nabla G(\mathbf{p}) - \nabla G(\mathbf{q})$*. Hence the profit of* $\mathbf{s}^*$ *is* $S(\mathbf{p}, \mathbf{p}^*) - S(\mathbf{q}, \mathbf{p}^*)$.

Remark 4 (Additional justification for the canonical betting strategy $\mathbf{s}_{\mathbf{q}\to\mathbf{p}}$)**.** *The foundational result of Hanson [2003, 2007] already implies that the betting strategy* $\mathbf{s}_{\mathbf{q}\to\mathbf{p}} (= \mathbf{s}_G(\mathbf{p}, \mathbf{q}))$ *that moves the market price from* $\mathbf{q}$ *to* $\mathbf{p}$ *in AMMs is robustly profitable in the sense that it guarantees positive profit whenever the forecast* $\mathbf{p}$ *beats the market under the ground truth* $\mathbf{p}^*$*. Our Theorem 2 provides an additional justification and shows that this is essentially the only robustly profitable strategy.*

3.3.3 Extensions to empirical scores

In practice the ground truth $\mathbf{p}^*$ is unobservable, therefore a forecaster's advantage over the market on a single event in expectation can never be validated. What is validatable however is the advantage on empirical scores. A practical question thus is whether proper betting can convert empirical score advantages to realized profit. The answer turns out to be *Yes*, inherently due to the fact that Theorem 1 holds also for realized outcome y and can then be averaged over a sequence of outcome realizations.

Suppose we only see realized outcomes $y_1, \ldots, y_n$ across a sequence of n events with predictions $\mathbf{p}_1, \ldots, \mathbf{p}_n$ and market prices $\mathbf{q}_1, \ldots, \mathbf{q}_n$. Let the empirical forecast and market scores be

$$\hat{S}_F := \frac{1}{n}\sum_i S(\mathbf{p}_i, y_i) \quad \text{and} \quad \hat{S}_M := \frac{1}{n}\sum_i S(\mathbf{q}_i, y_i).$$

Theorem 1 can be extended from a single-event to an empirical version below.

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

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

Hence the sequential proper betting has positive profit if the gap $\hat{S}_F - \hat{S}_M$ of empirical scores and average divergence $\frac{1}{n}\sum_{i=1}^n D_G(\mathbf{q}_i, \mathbf{p}_i)$ together offset the average liquidity loss $\frac{1}{n}\sum_{i=1}^n L_\rho(\mathbf{s}_i; \mathbf{q}_i)$. In Section 4, we experimentally observe a few forecasting agents profiting due to divergence from the market despite below-market accuracy.

Remark 5 (Bet sizes should vary across events)**.** *A natural intuition is that a forecaster whose accuracy is measured with equal weight across events should place equal bet size on these events. Corollary 2 says otherwise: the correct bet sizes are proportional to* $\mathbf{s}_i = \nabla G(\mathbf{p}_i) - \nabla G(\mathbf{q}_i)$*, with larger bets on events where the forecaster's prediction deviates more from the market.*

3.3.4 Extension to non-zero bid-ask spreads and correlated market outcomes.

In real prediction markets, factors such as low liquidity or platform transaction fees imply a positive difference between the lowest price at which someone will sell and the highest price at which someone will buy. For the k-th outcome of an event, a real market has $q_k^+ + q_k^- \geq 1$, where q_k^+ denotes the market price for buying the outcome (bid price) and q_k^- denotes the market price for selling the outcome (ask price). Theorem 1 is not directly applicable to this situation because when any k'th entry of the proper betting $\nabla G(\mathbf{p}) - \nabla G(\mathbf{q})$ is negative, its execution requires buying *NOT* k at price $1 - q_k^+$, which is not possible here since the price of *NOT* k is $q_k^- (> 1 - q_k^+)$.

The key idea to extend Theorem 1 to such real-world settings with $q_k^+ + q_k^- > 1$ is to decompose each outcome $k \in [K]$ into two *binary outcome* events, *YES* k and *NOT* k, that have prices equal to q_k^+ and q_k^- respectively. It can be shown that a proper betting will buy *YES* k contracts when $p_k > q_k^+$, buy *NO* k contracts when $p_k < 1 - q_k^-$, and not act when $1 - q_k^- \leq p_k \leq q_k^+$. Moreover, this can be further extended to situations when the K outcomes are not completely disjoint. For instance, a popular event on prediction markets is to predict bitcoin price where the outcome could be "price at least \$7K", "price at least \$8K", etc. In such cases, we treat each single outcome as a separate binary market. The full details of the reductions can be found in Appendix C.1.

3.3.5 Extension to sequential betting.

Events in prediction markets often take days or weeks to resolve, during which both the forecaster's prediction $\mathbf{p}^t$ and the market price $\mathbf{q}^t$ evolve in response to the revelation of new information. Appendix C.2 extends Theorem 1 to this dynamic setting via two strategies: a *fundamental*-driven strategy that executes the proper bet $\mathbf{s}_G(\mathbf{p}^t, \mathbf{q}^t)$ at each round and holds positions 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 profit decomposition of Theorem 1 summed across rounds, and the comparison reveals what each strategy needs for its performance guarantee: a per-round accuracy edge against the unobservable ground truth $\mathbf{p}^*$ (fundamental), versus a per-round accuracy edge against the next-step market price $\mathbf{q}^{t+1}$ (momentum) — an operationally observable condition.

4 Experiments

Testing betting strategies requires data that pairs forecasts with contemporaneous market prices and realized outcomes. Such data is hard to collect from human forecasters at scale, but recent AI forecasting benchmarks built on real-money prediction markets—most notably Prophet Arena [Yang et al., 2025]—provide a scalable alternative. We therefore use AI forecasts across thousands of archived markets to evaluate how different betting strategies translate predictive accuracy into realized returns. The evaluation has two stages: first, an offline study on the historical benchmark, where we simulate each market under zero price impact; second, a live deployment in real prediction markets, exposing our strategy to the full set of real-market frictions.

4.1 How good are proper betting strategies?

We use forecasts on 2,418 Kalshi markets [Kalshi, 2025] collected through Prophet Arena that span a variety of domains such as sports, politics, economics, and crypto. Each market includes a forecast, contemporaneous market price, and realized outcome (YES/NO). For each model and betting strategy, we apply the strategy to historical forecast–market (bid and ask) price pairs, simulate the resulting trades, and compute the realized return after market resolution, allowing us to compare proper betting strategies directly against standard heuristic baselines.

Specifically, each betting strategy is to determine a set of weights — how to allocate a fixed budget and on what direction — over a collection of prediction markets with forecasts and market prices. We implement the three proper betting strategies according to Table 1. We also consider a few baseline strategies derived from a few intuitive heuristics: (i) **Max-Margin**, which concentrates capital on

the largest disagreements with the market, either across all markets or by selecting 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 according to Kelly-style edge-to-odds ratios; and (iv) **Kelly Criterion** [Kelly, 1956], which maximizes asymptotic log-wealth growth[4]. Appendix D.2 expands on the formal setup and implementation details of the betting strategies.

Table 2 reveals why proper betting strategies warrant study: heuristic allocations fail to yield consistent 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 (i.e., minimal risk) bets, but as a consequence of its design, also limits opportunities for creating meaningful upside. The Kelly criterion performs particularly poorly because it is highly sensitive to miscalibrated probabilities. As a result, it becomes highly unstable under model error. More generally, these failures by heuristic betting strategies show that predictive accuracy alone is insufficient to create profit. How probabilities are translated into position sizes is a key determinant of returns, motivating the study of proper betting strategies. Appendix D.2 reports the results for all evaluated models.

Table 2: ROI (%) of baseline betting strategies and the proposed Proper (Brier) strategy on the standardized 200-event shared subset. ΔS is the Brier score gap: negative values indicate worse performance than the market. Bold marks the best ROI per row.

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

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

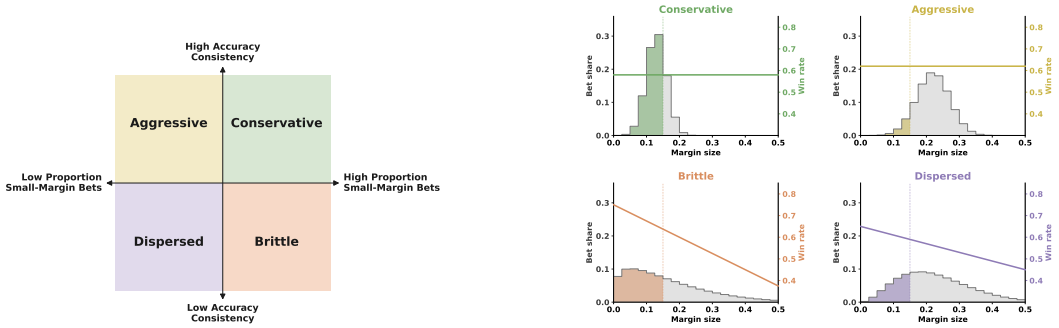
Score gap vs. divergence. Table 3 empirically illustrates the decomposition in Lemma 1. We report the proper score gap and divergence terms summed across bets and normalized by total spend, which exactly matches with the per-bet decomposition in Equation 1 in ROI units. This highlights how proper betting strategies induce trade-offs between accuracy and divergence. Profitability depends not only on the score gap (ΔS), but also on the divergence term (D), so models with worse accuracy can still achieve higher ROI if they generate sufficiently large divergence. For example, under the Log rule, GPT-5.2 (Base) achieves higher ROI than Gemini 3 despite a worse score gap, even though it is less accurate than both the market and Gemini 3. Moreover, a model's score under a given scoring rule does not by itself determine which rule yields the best returns, since divergence is rule-dependent. For instance, for Llama 4 Maverick, Brier yields a worse score gap than Spherical (-123.7 vs. -53.3) but higher ROI (-13.1 vs. -14.8), driven by a substantially larger divergence term. As such, identifying the most profitable models and betting strategies is non-trivial, as profitability depends not only on accuracy but also on how those strategies factor in disagreement with the market. Appendix D.3 reports the decomposition results for all models.

Forecaster personas. As seen in Table 3, different scoring rules yield materially different returns. To identify when each rule is preferable, we construct a controlled experiment using four synthetic model "personas" that all achieve the same Brier score on the same set of real prediction markets, but differ in how their predictions deviate from market prices. We focus on the margin range $|p - q| \in [0, 0.5]$, as higher-margin forecasts are too sparse in empirical models to support reliable or broadly applicable

[4]We use leverage if the Kelly ratio > 1 as explained in Appendix D.2.

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

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



(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.

analysis; although rare, such high-margin bets can occasionally be correct and highly profitable, which we examine further in Appendix D.8. We run the experiment in two regimes—one where the model beats the market and one where it loses—to examine how accuracy interacts with the choice of proper betting strategy and the resulting implications.

Persona design and motivation. We use 27,516 markets across 3,511 questions collected from the Prophet Arena pipeline between August 1, 2025 and April 15, 2026. Models differ in how aggressively they deviate from market prices, and in whether those deviations tend to pay off. To capture this variation, we construct personas defined by two features: (i) the proportion of small-margin bets (i.e., $|p - q| \leq 0.15$), and (ii) the accuracy drop-off across margins, measured by the slope of the directional win rate (the fraction of bets whose chosen side matches the realized outcome) as a function of margin size $|p - q|$. Both regimes are calibrated to a fixed ± 0.05 relative quadratic scoring rule gap. Appendix D.4 outlines the details of the synthetic forecast generation process.

We define four personas spanning representative combinations of these dimensions, as shown in Figure 1. **Conservative** models concentrate most forecasts in the small-margin region and exhibit flat accuracy across margins. **Aggressive** models place fewer small-margin bets, with accuracy remaining flat across margins. **Dispersed** models spread bets broadly across margin sizes and show a gradual decline in win-rate as margin increases. **Brittle** models have a significant proportion of small-margin forecasts but exhibit a sharply declining win-rate as margin increases. For further details, see Appendix D.4.

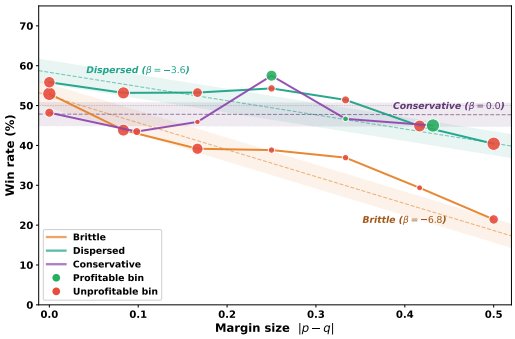
We evaluate each persona under two regimes: Regime A, where the model outperforms the market and all strategies yield positive ROI, and Regime B, where it underperforms and all strategies incur

Table 4: Simulated persona ROI (%) under proper betting strategies. S-M Share denotes the share of forecasts with small margin: $|p - q| \leq 0.15$. Consistency is the OLS slope of per-bet directional win rate against margin size.

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

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

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



(b) Binned per-bet win rate vs. margin $|p - q|$. Each dot has area proportional to profit for their persona. Dashed lines are OLS fits where β denotes the slope. Overlapping points are jittered horizontally for clarity.

Figure 2: Mapping real models to personas.

losses. A symmetric ± 0.05 quadratic score gap (market quadratic score = 0.839) isolates the effect of persona and strategy while holding average accuracy constant.

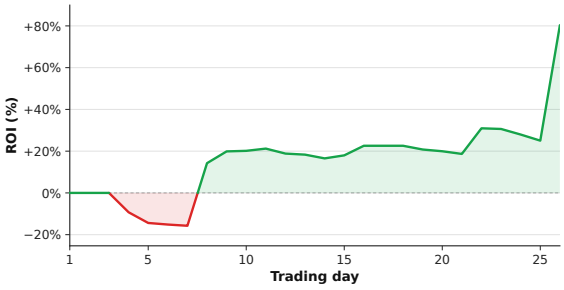
As shown in Table 4, Brier is optimal in Regime A for all personas, while Log is optimal in Regime B. In Regime A, the positive Brier gap implies that expected returns are positive across margin bins, so linear scaling in $|p - q|$ is optimal. In Regime B, losses are dominated by high-margin errors, making Log's sublinear weighting preferable as it attenuates exposure to large, costly mistakes. 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, leading to broad losses. Additional analysis and weighting visualizations are provided in Appendix D.5.

Mapping real models to personas. To map synthetic personas to real model behavior, we evaluate each model along the two defining dimensions: small-margin forecasts and accuracy consistency. Because these patterns only emerge reliably at scale, we evaluate each model on its full event history rather than the standardized subset of events. Figure 2 summarizes these results, with Figure 2a showing models' proportion of small-margin forecasts and Figure 2b showing the win-rate profiles as a function of margin size. For classification of all evaluated models and additional analysis, see Appendix D.6.

Along the margin dimension, models separate cleanly into the three persona groups: Brittle models concentrate most bets at small margins, Dispersed models spread forecasts more broadly across margins, and Conservative models place nearly all bets at small margins. Figure 2b further confirms the taxonomy along the accuracy-consistency dimension: Brittle models exhibit a steep decline in win rate as margins increase, Dispersed models decline more gradually, and Conservative models remain roughly flat across margins. Taken together, these results show that personas provide a simple

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

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



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

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

way to summarize how model behavior translates into returns. This abstraction lets us reason about profitability and select or adapt betting strategies based on these behavioral patterns.

4.3 Live deployment under active portfolio management

We further evaluate proper betting in a live portfolio setting using the momentum-driven strategy described in Section 3.3 (with details deferred to Appendix C.2). Unlike the previous backtests, live execution must contend with bid–ask spreads, limited liquidity, slippage, and discrete tick sizes. As a result, even accurate forecasts may not translate cleanly into realized returns, since trades must be executed at available prices rather than idealized market probabilities.

We deploy Gemini 3 using the Brier-derived betting strategy on Kalshi events for 26 days (April/May 2026) with an initial budget of \$200. Operating on a fixed two-hour cadence, the agent evaluates all open markets satisfying an ex-ante eligibility filter that excludes contracts with ambiguous resolution criteria and settings known to impair LLM forecasting performance (see Appendix D.7.2). Summary statistics and cumulative ROI from the live deployment are shown in Figure 3, whereas the trading history **is documented here**, which logs performance of the trading agent over the trading period. Over the trading period, the agent executed 236 orders across 129 markets, achieving an ROI of **+80.33%** (decomposed into $\Delta S = +0.7205$ and $D = +0.0828$ under the Brier decomposition) and a Sharpe ratio of **3.35**, indicating that nearly all gains arise from predictive accuracy rather than divergence, consistent with the Conservative persona identified in Section 4.2. Note that we report the Sharpe ratio only for the live deployment, where the daily returns form a continuous time series amenable to risk-adjusted analysis; the previous offline experiments resolve each event independently and most models post negative returns, so a Sharpe ratio there would be uninformative. We document the full implementation, analysis of representative trades, and trading logs in Appendix D.7.

5 Conclusion

We investigate the puzzle that forecasters can outperform prediction markets in accuracy yet still lose money, by establishing a formal equivalence between predictive accuracy and trading profitability. Key to our result is to identify a "proper" betting strategy for arbitrary prediction markets. Moreover, we demonstrate that this strategy strictly generalizes the canonical betting strategy only applicable to the special AMM markets, is essentially the only robustly profitable strategy, and has deep connection to proper scoring rules. Through this lens, it is unsurprising that proper betting emerges as the only strategy reliably converting accuracy into ROI across thousands of archived AI forecasts; what is more striking is both the magnitude and the practical viability in real markets — a month-long live deployment achieved $+80.33\%$ ROI with a Sharpe ratio of 3.35.

Several future directions remain open. First, our persona analysis in Section 4.2 hints at the potential of *data-adaptive scoring-rule selection*: identifying the most suitable rule for each forecaster. Second, the framework opens *market-design* questions on which scoring rules and market structures best elicit accurate forecasting from participants at scale. Finally, our guarantees concern *expected* profit; extending them to risk-adjusted criteria would narrow the gap between theory and live deployment.

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. In *Proceedings of the 5th ACM Conference on Electronic Commerce (EC)*, pages 41–47, 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*. Holt, Rinehart and Winston, 1971.

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 Implementations of prediction markets

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

Automated market makers. A classic implementation of a prediction market is an *automated market maker* (AMM) governed by a cost function [Hanson, 2003, 2007, Chen 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 strictly convex, differentiable *cost function* $C : \mathbb{R}^K \to \mathbb{R}$. A forecaster who executes the position vector $\mathbf{s} \in \mathbb{R}^K$ pays the total cost $C(\mathbf{x}+\mathbf{s}) - C(\mathbf{x})$ of moving the share state to $\mathbf{x}+\mathbf{s}$, and receives the payout $\mathbf{s} \cdot \mathbf{1}_y$ for a realized outcome y, with expected profit $\pi^{\text{AMM}}(\mathbf{s}, \mathbf{p}^*) := \mathbf{s} \cdot \mathbf{p}^* - (C(\mathbf{x}+\mathbf{s}) - C(\mathbf{x}))$. The 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, ρ is the piecewise-constant function read directly off the order book, and L_ρ captures the depth-induced slippage as the trader walks the book.

B Omitted proofs

B.1 Proof of Theorem 2

For any strictly proper scoring rule S, we show that if a betting strategy $\mathbf{s}(\mathbf{p}, \mathbf{q})$ is intrinsically different from $\mathbf{s}_G(\mathbf{p}, \mathbf{q})$, then we can find some $\mathbf{p} \neq \mathbf{q}$ and $\mathbf{p}^*$ such that $S(\mathbf{p}; \mathbf{p}^*) \geq S(\mathbf{q}; \mathbf{p}^*)$ but $\boldsymbol{s}(\mathbf{p}, \mathbf{q})$ has non-positive expected profit under $\mathbf{p}^*$.

Given any two intrinsically different betting strategies $\mathbf{s}(\mathbf{p}, \mathbf{q})$ and $\mathbf{s}_G(\mathbf{p}, \mathbf{q})$, we first apply constant shifting and rescaling to both strategies to normalize them so that $\mathbf{s}(\mathbf{p}, \mathbf{q}) \cdot \mathbf{1} = \mathbf{s}_G(\mathbf{p}, \mathbf{q}) \cdot \mathbf{1} = 0$ (via constant shifting) and $||\mathbf{s}(\mathbf{p}, \mathbf{q})|| = ||\mathbf{s}_G(\mathbf{p}, \mathbf{q})|| = 1$ (via rescaling). No transformations are needed if $\mathbf{s}(\mathbf{p}, \mathbf{q}) = \mathbf{0}$. Since $\mathbf{s}(\mathbf{p}, \mathbf{q})$ and $\mathbf{s}_G(\mathbf{p}, \mathbf{q})$ are intrinsically different, after normalizations there still exists an interior market price $\mathbf{q} \in \text{int}(\Delta_{[K]})$, a sequence of forecasts converging to $\mathbf{q}$, i.e., $\{\mathbf{p}_t\}_{t=1}^\infty \to \mathbf{q}$, yet a small constant ϵ such that $||\mathbf{s}(\mathbf{p}_t, \mathbf{q}) - \mathbf{s}_G(\mathbf{p}_t, \mathbf{q})|| \geq \epsilon$ for every $t = 1, 2, \cdots$. The sequences $\{\mathbf{s}(\mathbf{p}_t, \mathbf{q})\}_{t=1}^\infty$ and $\{\mathbf{s}_G(\mathbf{p}_t, \mathbf{q})\}_{t=1}^\infty$ are all bounded, so there exists subsequences that converge. Slightly overloading the notation, let

$$\{\mathbf{s}(\mathbf{p}_t, \mathbf{q})\}_{t=1}^\infty \to \hat{\mathbf{s}} \quad \text{and} \quad \{\mathbf{s}_G(\mathbf{p}_t, \mathbf{q})\}_{t=1}^\infty \to \mathbf{s}^*.$$

We have $||\hat{\mathbf{s}} - \mathbf{s}^*|| \geq \epsilon$. Next we shall construct a $\mathbf{p}_t^*$ for every t such that $S(\mathbf{p}_t, \mathbf{p}_t^*) \geq S(\mathbf{q}, \mathbf{p}_t^*)$ but the expected profit of $\mathbf{s}(\mathbf{p}_t, \mathbf{q})$ is negative for large enough t, hence disproving the robust profitability of $\mathbf{s}(\mathbf{p}_t, \mathbf{q})$.

We start by handling the corner case with $K = 2$. In this case, after normalization via rescaling and constant shifting, the betting strategy is either $(1/\sqrt{2}, -1/\sqrt{2})^\top$ or $(-1/\sqrt{2}, 1/\sqrt{2})^\top$ or $\mathbf{0}$, which essentially buys $\sqrt{2}$ of *YES*, or buys $\sqrt{2}$ of *NO*, or buys nothing. Since $||\mathbf{s}(\mathbf{p}_t, \mathbf{q}) - \mathbf{s}_G(\mathbf{p}_t, \mathbf{q})|| \geq \epsilon$, they must correspond to these two of these three different strategies respectively. Since $\mathbf{p}_t \neq \mathbf{q}$, $\mathbf{s}_G(\mathbf{p}_t, \mathbf{q})$ cannot be $\mathbf{0}$. If $\mathbf{s}_G(\mathbf{p}_t, \mathbf{q})$ buys *YES*, then $\mathbf{p}_t^* = (1, 0)^\top$ induces higher score for $\mathbf{p}_t$ than that of the market $\mathbf{q}$ but induces non-positive profit for the other betting strategy $\mathbf{s}(\mathbf{p}_t, \mathbf{q})$, as desired. The case when $\mathbf{s}_G(\mathbf{p}_t, \mathbf{q})$ buys *NO* is symmetric.

In the remainder of the proof, we consider the non-trivial case $K > 2$. We claim that there exists a constant $\mathbf{v}$ such that (1) $\mathbf{v} + \mathbf{q} \in \Delta_k$ or equivalently $\mathbf{v} \cdot \mathbf{1} = 0$, and (2) $\mathbf{v} \cdot \mathbf{s}^* > c > 0 > -c > \mathbf{v} \cdot \hat{\mathbf{s}}$ for some small positive constant c. To see this, let $\mathbf{1}^\perp = \{\mathbf{v} \in \mathbb{R}^K : \mathbf{v} \cdot \mathbf{1} = 0\}$ denote the subspace that is perpendicular to the all one vector $\mathbf{1}$. Since $\mathbf{q} \in \mathrm{int}(\Delta_{[K]})$, $\Delta_{[K]}$ contains a full-dimensional unit ball within the subspace $\mathbf{1}^\perp$ that is centered around $\mathbf{q}$. Due to normalization of betting strategy $\mathbf{s}, \mathbf{s}_G$, we know $\bar{\mathbf{s}}, \bar{\mathbf{s}}_G \in \mathbf{1}^\perp$ and so is their difference. Since $||\hat{\mathbf{s}} - \mathbf{s}^*|| \geq \epsilon$, there must exist some $\mathbf{v} \in \mathbf{1}^\perp$ such that $\mathbf{v} \cdot \mathbf{s}^* > c > 0 > -c > \mathbf{v} \cdot \hat{\mathbf{s}}$, as desired.

We construct $\mathbf{p}_t^* = \mathbf{q} + \mathbf{v} \in \Delta_K$ (the same for every t), with the $\mathbf{v}$ as chosen above. The expected profit of betting strategy $\mathbf{s}(\mathbf{p}_t, \mathbf{q})$ under this $\mathbf{p}_t^*$ satisfies

$$\begin{aligned}
(\mathbf{p}_t^* - \mathbf{q}) \cdot \mathbf{s}(\mathbf{p}_t, \mathbf{q}) &= \mathbf{v} \cdot \mathbf{s}(\mathbf{p}_t, \mathbf{q}) \\
&= \mathbf{v} \cdot [\mathbf{s}(\mathbf{p}_t, \mathbf{q}) - \hat{\mathbf{s}}] + \mathbf{v} \cdot \hat{\mathbf{s}} \\
&< 0 \quad \text{for any } t \text{ large enough,}
\end{aligned}$$

because the first term goes to 0 due to $\mathbf{s}(\mathbf{p}_t, \mathbf{q}) \to \hat{\mathbf{s}}$, whereas the second term $\mathbf{v} \cdot \hat{\mathbf{s}}\,(< -c)$ is a negative constant.

However, we show that $\mathbf{p}_t$'s expected score under ground truth probability $\mathbf{p}_t^*$ is positive. Let G be the associated potential function of strictly proper scoring rule S, i.e., $S(\mathbf{p}, y) = G(\mathbf{p}) + \nabla G(\mathbf{p}) \cdot (\mathbf{1}_y - \mathbf{p})$. Then we have

$$\begin{aligned}
& S(\mathbf{p}_t, \mathbf{p}_t^*) - S(\mathbf{q}, \mathbf{p}_t^*) \\
= \;& G(\mathbf{p}_t) + \nabla G(\mathbf{p}_t) \cdot (\mathbf{p}_t^* - \mathbf{p}_t) - G(\mathbf{q}) - \nabla G(\mathbf{q}) \cdot (\mathbf{p}_t^* - \mathbf{q}) \\
= \;& G(\mathbf{p}_t) - G(\mathbf{q}) - \nabla G(\mathbf{q}) \cdot (\mathbf{p}_t^* - \mathbf{q}) + \nabla G(\mathbf{p}_t) \cdot (\mathbf{p}_t^* - \mathbf{p}_t) \\
= \;& G(\mathbf{p}_t) - G(\mathbf{q}) - \nabla G(\mathbf{q}) \cdot (\mathbf{p}_t - \mathbf{q}) - \nabla G(\mathbf{q}) \cdot (\mathbf{p}_t^* - \mathbf{p}_t) + \nabla G(\mathbf{p}_t) \cdot (\mathbf{p}_t^* - \mathbf{p}_t) \\
= \;& D_G(\mathbf{p}_t, \mathbf{q}) + [\nabla G(\mathbf{p}_t) - \nabla G(\mathbf{q})] \cdot (\mathbf{p}_t^* - \mathbf{p}_t) \\
= \;& D_G(\mathbf{p}_t, \mathbf{q}) + \mathbf{s}_G(\mathbf{p}_t, \mathbf{q}) \cdot (\mathbf{q} + \mathbf{v} - \mathbf{p}_t) \\
= \;& D_G(\mathbf{p}_t, \mathbf{q}) + \mathbf{s}_G(\mathbf{p}_t, \mathbf{q}) \cdot (\mathbf{q} - \mathbf{p}_t) + (\mathbf{s}_G(\mathbf{p}_t, \mathbf{q}) - \mathbf{s}^*) \cdot \mathbf{v} + \mathbf{s}^* \cdot \mathbf{v} \\
> \;& 0 \quad \text{for any } t \text{ large enough,}
\end{aligned}$$

because the Bregman divergence $D_G(\mathbf{p}_t, \mathbf{q}) \geq 0$ for the convex G, the last term $\mathbf{s}^* \cdot \mathbf{v}$ is larger than the positive constant c due to our choice of $\mathbf{v}$, whereas the two middle terms both go to 0 since $\mathbf{p}_t \to \mathbf{q}$, $\mathbf{s}_G(\mathbf{p}_t, \mathbf{q}) \to \mathbf{s}^*$ whereas $\mathbf{s}_G(\mathbf{p}_t, \mathbf{q}), \mathbf{v}$ are all bounded vectors.

This concludes the proof of the second part. That is, we have found $\mathbf{p}_t, \mathbf{p}_t^*, \mathbf{q}$ for large enough t such that $S(\mathbf{p}_t, \mathbf{p}_t^*) - S(\mathbf{q}, \mathbf{p}_t^*) > 0$ but the expected profit of the betting strategy $\mathbf{s}(\mathbf{p}_t, \mathbf{q})$ under this $\mathbf{p}_t^*$, i.e., $(\mathbf{p}_t^* - \mathbf{q}) \cdot \mathbf{s}(\mathbf{p}_t, \mathbf{q})$, is strictly negative.

B.2 Proof of Proposition 2

The "only if" direction is precisely Theorem 1 since proper betting guarantees robust profitability. We thus only need to prove the "if" direction below. The challenge here is to show that if G is not strictly convex, any betting strategy – not just proper betting – cannot be robustly profitable. That is,

for any not strictly convex G and any betting strategy $\mathbf{s}(\mathbf{p}, \mathbf{q})$, there exists $\boldsymbol{p}^*$ and $\mathbf{p} \neq \mathbf{q}$ that satisfy $S(\mathbf{p}; \mathbf{p}^*) \geq S(\mathbf{q}; \mathbf{p}^*)$ [5] but yield non-positive profit under $\mathbf{s}$.

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

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

Given the $\mathbf{p}, \mathbf{q}$ pair satisfying the above condition, we now construct a corresponding $\mathbf{p}^*$. Let $\mathbf{s}$ be any betting strategy, $\hat{\mathbf{s}} = \frac{\mathbf{s}(\mathbf{p},\mathbf{q})}{||\mathbf{s}(\mathbf{p},\mathbf{q})||}$ denote the direction of the betting strategy at the above $\mathbf{p}, \mathbf{q}$, and $L = ||\mathbf{s}_G(\mathbf{p}, \mathbf{q})||$ denote 2-norm of the proper betting strategy. We construct $\mathbf{p}^* = \mathbf{q} - \frac{\kappa}{2L}\hat{\mathbf{s}}$.

Next we show that the constructed $\mathbf{p}, \mathbf{q}, \mathbf{p}^*$ have negative profit but satisfy $S(\mathbf{p}; \mathbf{p}^*) \geq S(\mathbf{q}; \mathbf{p}^*)$. Notably, this is an interesting situation where the market price $\mathbf{q}$ may be much closer to the ground truth $\mathbf{p}^*$ than the forecast $\mathbf{p}$, but has worse score than $\mathbf{p}$. This is intrinsically due to the non-convexity of G.

The difference of the expected score is analyzed as follows:

$$\begin{aligned}
S(\mathbf{p}, \mathbf{p}^*) - S(\mathbf{q}, \mathbf{p}^*) &= G(\mathbf{p}) + \nabla G(\mathbf{p}) \cdot (\mathbf{p}^* - \mathbf{p}) - G(\mathbf{q}) - \nabla G(\mathbf{q}) \cdot (\mathbf{p}^* - \mathbf{q}) \\
&= G(\mathbf{p}) - G(\mathbf{q}) + \nabla G(\mathbf{p}) \cdot (\mathbf{q} - \mathbf{p}) - [\nabla G(\mathbf{q}) - \nabla G(\mathbf{p})] \cdot (\mathbf{p}^* - \mathbf{q}) \\
&= -D_G(\mathbf{q}, \mathbf{p}) + \mathbf{s}_G(\mathbf{p}, \mathbf{q}) \cdot (\mathbf{p}^* - \mathbf{q}) \\
&\geq \kappa - L \cdot \kappa/(2L) > 0
\end{aligned}$$

The profit of $\mathbf{s}(\mathbf{p}, \mathbf{q})$ is analyzed as follows:

$$(\mathbf{p}^* - \mathbf{q}) \cdot \mathbf{s}(\mathbf{p}, \mathbf{q}) = -\frac{\kappa}{2L}\hat{\mathbf{s}} \cdot \mathbf{s}(\mathbf{p}, \mathbf{q}) = -\frac{\kappa}{2L}||\mathbf{s}(\mathbf{p}, \mathbf{q})|| < 0$$

B.3 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 strictly convex cost function $C(\mathbf{x})$, 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 betting. We compute

$$\begin{aligned}
L_\rho(\mathbf{s}^*; \mathbf{q}) &= C(\mathbf{s}^* + \mathbf{x}) - C(\mathbf{x}) - \nabla C(\mathbf{x}) \cdot \mathbf{s}^* \\
&= D_C(\mathbf{s}^* + \mathbf{x}, \mathbf{x}) \\
&= D_G((\nabla C)^{-1}(\mathbf{s}^* + \mathbf{x}), (\nabla C)^{-1}(\mathbf{x})) \\
&= D_G(\mathbf{p}, \mathbf{q})
\end{aligned}$$

where the third equation is due to the fact that if C, G are convex conjugates, then $D_C(\mathbf{x}', \mathbf{x}) = D_G((\nabla C)^{-1}(\mathbf{x}'), (\nabla C)^{-1}(\mathbf{x}))$.

C Extensions of Theorem 1

C.1 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.

[5] Scoring rule S is defined as $S(\mathbf{p}, y) = G(\mathbf{p}) + \nabla G(\mathbf{p}) \cdot (\mathbf{1}_y - \mathbf{p})$.

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

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

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

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

To see this is a valid betting strategy under non-zero bid/ask spread, we observe that $G(p_k) - \nabla G(q_k^+) \ge 0$ when $p_k > q_k^+$ due to convexity of G and $G(p_k) - \nabla G(1 - q_k^-) > 0$ for similar reasons. Hence, s_k never needs to buy the other side (which may have inconsistent price due to non-zero 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 $q_k^+ + q_k^- \ge 1$). 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 bid-ask-aware proper bet has expected profit admits the following decomposition*

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

Proof. For each coordinate k, $s_k = \nabla G(p_k) - \nabla G(\tilde{q}_k)$ is the binary proper bet with respect to the reference $\tilde{q}_k$. Apply the binary version of Lemma 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 $y \sim \mathbf{p}^*$, minus the liquidity loss, yields Eq. (5). □

Remark 6 (Correlated outcomes.)**.** *Corollary 3 holds even for outcomes $\{1, \cdots, K\}$ that need not to 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

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

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

C.2 Proper betting for long time horizon

For events that take days or weeks to resolve, both forecaster and market predictions evolve over time. A forecaster can make different prediction $\mathbf{p}^t$ at different time $t = 1, \cdots, T$ and makes trades based on $\mathbf{p}^t$ as well as then market price $\mathbf{q}^t$. We extend Theorem 1 to two natural multi-period strategies, each compounding the per-round proper bet but differing in what they hold and what accuracy edge they require. 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. Then the cumulative expected profit at resolution decomposes as*

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

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* $\mathbf{x}^0 := \mathbf{0}$*. Then the cumulative mark-to-market return decomposes as*

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

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:

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

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

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

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

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

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

where the inner sum telescopes by

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

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

Table 5: 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 ± 0.005	0.916 ± 0.006	-7.8% ± 3.6 (Log)
GPT-5.2 (Base)	11,789	0.013 ± 0.003	0.916 ± 0.005	-4.4% ± 3.9 (Log)
Grok 4.1 Fast	11,407	0.027 ± 0.004	0.913 ± 0.005	-9.1% ± 4.5 (Log)
Claude Sonnet 4.5	11,785	0.023 ± 0.003	0.911 ± 0.005	-8.1% ± 3.5 (Log)
Kimi K2 Thinking	11,811	0.024 ± 0.004	0.912 ± 0.005	-8.7% ± 3.8 (Log)
DeepSeek V3.2	11,071	0.031 ± 0.005	0.899 ± 0.006	-12.8% ± 3.4 (Log)
Minimax M2	11,688	0.028 ± 0.005	0.900 ± 0.006	-9.4% ± 3.8 (Log)
Brittle (avg)	**11,620**	**0.024**	**0.910**	$\mathbf{-8.6\%}$ **(Log)**
LLaMA 4 Maverick	29,613	0.069 ± 0.004	0.871 ± 0.005	-4.9% ± 2.2 (Log)
Qwen 3 235B	29,923	0.078 ± 0.004	0.868 ± 0.004	-11.2% ± 1.9 (Log)
DeepSeek R1	29,759	0.100 ± 0.005	0.848 ± 0.005	-5.4% ± 1.7 (Log)
Dispersed (avg)	**29,765**	**0.082**	**0.862**	$\mathbf{-7.2\%}$ **(Log)**
Gemini 3	9,721	0.017 ± 0.003	0.936 ± 0.004	$+14.2\%$ ± 8.8 (Brier)
Claude Opus 4.6	3,172	0.011 ± 0.004	0.949 ± 0.007	$+17.5\%$ ± 7.2 (Brier)
Conservative (avg)	**6,446**	**0.014**	**0.942**	$\mathbf{+15.8\%}$ **(Brier)**

D Additional experiments

D.1 Model details

Table 5 reports the LLMs evaluated across three metrics – expected calibration error (ECE), the score under the quadratic scoring rule, and ROI under each model's best proper betting strategy. GPT-5.2 (Base/High) correspond to different reasoning levels of the same underlying model. The number of markets varies across forecasters because models were released at different times, resulting in differing amounts of collected data through the Prophet Arena pipeline; to ensure comparability, we standardize evaluations in Section 4.

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

D.2 Implementation details of betting strategies

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

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

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

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

1. **Max-Margin:** Strategies that bet the margin.
   - *Market-Level:* $w_i \propto \text{sign}(p_i - q_i), \forall i \in [n]$. This strategy places a unit bet on every market, with direction determined by whether $p_i > q_i$.
   - *Grouped:* Let $\{\mathcal{I}_k\}_{k=1}^m$ denote a partition of markets into groups corresponding to the same underlying question. For each group k, define

     $$i^\star(k) = \arg\max_{i \in \mathcal{I}_k} |p_i - q_i|.$$

     This strategy allocates the budget uniformly across groups, placing a single bet per group on the market with the largest margin:

     $$w_{i^\star(k)} = \frac{\text{sign}(p_{i^\star(k)} - q_{i^\star(k)})}{m}, \quad w_i = 0, \quad \forall i \notin \{i^\star(k)\}_{k=1}^m.$$

2. **Inverse-Margin:** $w_i \propto \frac{\text{sign}(p_i - q_i)}{|p_i - q_i|}, \forall i \in [n]$. This heuristic places more weight on markets with smaller margins, under the hypothesis that forecasts are more reliable when deviations from the market are modest.
3. **Kelly-Alike:** $w_i \propto \frac{p_i - q_i}{1 - q_i}, \forall i \in [n]$. This heuristic strategy mimics the ratio in the Kelly criterion.
4. **Kelly Criterion [Kelly, 1956]:** Unlike the previous strategies, which allocate a fixed budget B across markets via weights w_i with the goal of maximizing expected profit, Kelly sizing determines the fraction of wealth to wager on each individual market sequentially to maximize the expected log-growth rate. For a binary market with forecast p_i and market price q_i, a unit stake on YES pays $1/q_i$, giving net odds $b_i = (1 - q_i)/q_i$. Maximizing

   $$\mathbb{E}[\log W_{i+1}] = p_i \log(1 + f_i b_i) + (1 - p_i) \log(1 - f_i)$$

   over the wealth fraction f_i yields the closed form $f_i = \frac{|p_i - q_i|}{1 - q_i}$. If f_i exceeds 1, we allow for the use of leverage to purchase additional shares.

Table 6 reports the full comparison of evaluated models across all baseline betting strategies and the proposed Proper (Brier) strategy.

Table 6: ROI (%) of baseline betting strategies and the proposed Proper (Brier) strategy. ΔS is the proper (Brier) score gap: negative values indicate worse performance than the market. Bold marks the best ROI per row.

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

D.3 Empirical decomposition

Table 7 reports the full decomposition across all models and all evaluated proper betting strategies on the standardized 2,418 market dataset. Divergence varies substantially across both models and rules, and often offsets large negative score gaps, meaning models less accurate can still yield positive ROI (e.g., Claude Sonnet 4.5 under Log). Rules such as Log tend to amplify divergence, leading to higher upside but also greater dispersion in outcomes, while Brier and Spherical are comparatively more stable. This reinforces that the mapping from forecasts to returns is highly rule-dependent, and that profitability hinges as much on how aggressively a strategy exploits disagreement as on the underlying predictive accuracy.

Table 7: ROI (%) decomposition of proper betting strategies. ΔS is the aggregate score difference between the forecast and the market across all bets placed, and D is total Bregman divergence. All quantities are summed across bets, normalized by total cost staked under each rule and multiplied by 100 so $\Delta S + D = \text{ROI}$. Values are rounded to 1 decimal place.

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

D.4 Synthetic persona generation

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

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

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

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

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

For Brittle, matching the same ΔS (Brier) cannot be achieved with a single linear win-rate schedule: slopes steep enough to capture the empirical tail decline reduce mean accuracy below the Regime A

target, while flatter slopes eliminate the pattern. We therefore use a two-piece schedule

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

with $\tau = 0.20$, $s = 1.0$, and $m \sim \mathrm{Beta}(0.12, 20)$, calibrated to $a_0 = 0.999/0.223$ for Regimes A/B. In Figure 1b, we visualize this fitting a linear regression (OLS) for comparability with the other personas.

D.5 Proper betting rule capital allocation surfaces

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

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

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

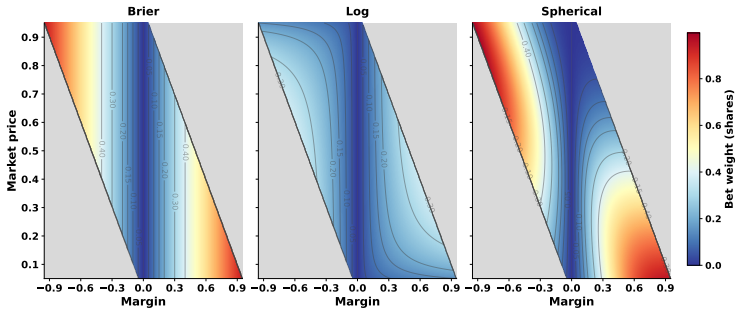


Figure 4: Capital allocation (bet weight in shares) as a function of signed margin and market price (q, y-axis) for the three proper betting strategies. Warmer colors indicate larger bets. Grey regions are infeasible ($p \notin (0, 1)$). Contour lines mark constant weight levels.

D.6 Mapping empirical forecasters to personas

As shown in Table 8, along the margin dimension, models separate into three clear groups: seven place 75%–85% of bets below 0.15 (Brittle), three distribute more mass across margins (Dispersed, $<$ 70% below 0.15), and two concentrate over 85% at small margins (Conservative). Second, Figure 2b 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 ($\beta = 0$).

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

Model	**[0,.05)**	**[.05,.15)**	**[.15,.25)**	**[.25,.50)**	**[.50,.70)**	**[.70,.90)**	**[.90,1.0)**	$\boldsymbol{\Delta S}$
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**	$\mathbf{+0.001}$

D.7 Details about live deployment experiment

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

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.

D.7.1 Prediction context and methodology

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

Whenever a market is traded, we rebalance the position toward the new target τ on each cycle, as is detailed in Appendix C.2.

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 3a 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.

D.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[6], creating a mismatch between the LLM's belief about market closure and the actual outcome realization.

In addition to eligibility criteria, we also drop a market from consideration due to an underspecified resolution rule. We have contacted Kalshi about contracts of this kind, and we view such situations as an artifact of the novelty of prediction markets that will phase out as venues mature and resolution-rule conventions standardize; this is not a limitation of our betting strategy but of the underlying contract specification. The event was titled *"What will be the top AI model this month?"* and the published resolution rule reads in full: *"If `claude-opus-4-6-thinking` is the top-ranked AI model on Apr 30, 2026 at 10:00 AM ET, then the market resolves to Yes."* The rule does not specify which leaderboard, benchmark, or evaluation suite determines "top-ranked," nor does it name the operator that produces such a ranking. Furthermore, plausible benchmarks — LMArena, Artificial Analysis, the OpenLLM Leaderboard, MMLU, etc. — can even disagree on a given day, so the market's outcome is determined by the market operator's discretionary choice of ranking source rather than by any externally verifiable fact. We therefore treat this market as failing the resolution-clarity precondition for forecasting and exclude it from the agent's eligible set [7].

D.7.3 Case studies

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

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

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

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

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

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

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

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

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

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

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

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

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

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

D.8 Profitability at extreme margins

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

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

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

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

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

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

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

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

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

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

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

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

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

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