Kalshi Research

Calibration in Prediction Markets: Theory and Evidence

Nicole Kagan and Rubens Baiocchi

Working Paper

August 2026

Abstract

Prediction markets are commonly described as producing prices that admit a literal probabilistic interpretation, but this empirical claim has yet to be evaluated at scale on any U.S.-regulated exchange. This paper represents the first such exercise, using the complete history of resolved markets on Kalshi—the world's largest prediction market—comprising 2,243,741 markets across eleven categories from the platform's launch in 2021 through mid-2026, to evaluate calibration and accuracy as functions of time-to-resolution, trading volume, and the number of participating traders. We document that Kalshi's prices are, in aggregate, extremely well calibrated as markets approach resolution. Assessing all markets (excluding Exotics), we record a decline in Brier score from approximately 0.08–0.09 at a 3-Month horizon to roughly 0.02 at Close. Further, reliability diagrams track the 45° line closely across nearly every category examined. In parallel, we note that naive accuracy rises from 88.3% at a 3-Month horizon to 97.2% at Close. Beneath these strong aggregate results lie more granular conclusions. First, calibration quality improves near-monotonically with both event-level trading volume and the number of unique traders, which we interpret as preliminary evidence that deeper participation may sharpen prices rather than distort them. Second, calibration is stronger and accuracy higher for high-attention categories listed well in advance, such as Economics, than for categories for which the information discovery period is limited and outcomes widely dispersed, such as Mentions. Third, though it is a natural assumption that calibration is a linear function of time-to-resolution, we find that anchoring to the true timing of the underlying event's occurrence rather than to raw closing timestamps, which may be subject to significant delay, unexpectedly improves calibration in Sports and de-biases results in Elections. To our knowledge, this represents the first attempt to re-scale expiration times in the study of prediction markets. We interpret these findings, taken together, as variation around a strong baseline rather than as a qualification of it: Kalshi's prices behave like genuine probabilities, and increasingly so as resolution approaches.

Keywords: Prediction Markets, Calibration, Accuracy, Market Efficiency, Kalshi

1 Introduction

A central claim of prediction market proponents is that these markets provide meaningful insight into future events. As the argument goes, by aggregating dispersed information from a heterogeneous population of traders, such markets harness a form of collective intelligence to produce an accurate view of how likely an event is. This idea is hardly new. It traces back to Hayek's (1945) observation that markets are uniquely effective at soliciting and aggregating information that is held by many individuals, no one of whom possesses the complete picture. Applied to forecasting, the argument is that a market price accomplishes something a single expert, poll, or model cannot easily replicate. It compels participants with genuinely different information and beliefs to translate their views into a single number, with real money disciplining that translation. A trader who believes the market price is wrong has a direct financial incentive to correct it, and in doing so to move the price toward the truth.

That background raises a natural question. Does the information encoded in a prediction market price actually behave like a probability, and if so, under what conditions does that behavior hold? To answer that question, we draw on a sample of 2,243,741 resolved markets from Kalshi, the world's largest prediction market, spanning from its launch in 2021 through the middle of 2026. We evaluate the informational value of displayed prices through two complementary lenses: calibration and accuracy.

Calibration analysis assesses whether the price of a given market at a given time indicates the true probability of the corresponding event. Formally, it tests whether quoted probabilities match observed outcome frequencies. Taken to its extreme, a perfectly calibrated market is one in which a contract trading at 70 cents resolves to "Yes" exactly 70% of the time. This analysis speaks directly to whether prices can be interpreted literally as probabilities, with broad implications for anyone using Kalshi pricing as an input to trading, decision-making, or forecasting.

We complement calibration with a measure of accuracy, which rounds any price above 50 cents to a prediction of "Yes" and evaluates whether that prediction was correct. Additionally, we produce a lead-market measure, which asks whether, in markets with several mutually exclusive outcomes, the most-favored outcome was the one that occurred.

At a high level, we find that Kalshi's prices are well calibrated and accurate overall, a marked improvement on random guessing that grows as resolution approaches. Zooming in, a few key features become apparent. First, calibration is a function of time-to-resolution, trading volume, and the number of participating traders. Second, meaningful variation appears when assessed on a per-category level, and sensitivity to end-date measurement (e.g. market resolution time vs. event-end time) becomes pronounced. The remainder of this paper is structured as follows: Section 2 reviews the relevant literature to date; Section 3 sets out the definitions and methodological framework used; Section 4 presents calibration results, both for the full sample and a series of informative subsamples; Section 5 presents accuracy results; Section 6 discusses these results; Section 7 addresses the study's limitations; and Section 8 concludes, drawing out the implications of these findings and outlining the natural next steps for this research agenda.

2 Context and Literature Review

2.1 Kalshi

Kalshi began operating in 2021, following approval from the U.S. Commodity Futures Trading Commission (CFTC) in November 2020 as a "Designated Contract Market"—the same regulatory category that governs traditional futures exchanges such as the Chicago Mercantile Exchange. It

builds upon a longer tradition stemming from the Iowa Electronic Markets first introduced in 1988. Kalshi's core product is the *event contract*, an Arrow–Debreu security that pays \$1 if a specified outcome occurs and \$0 otherwise, quoted in cents between 1 and 99, with trades matched between a "Maker," who posts the first offer, and a "Taker," who accepts it. Kalshi publishes its trade data through a public API at no cost.

2.2 Literature Review

This paper sits at the intersection of three literatures that have largely developed on separate tracks: the forecast-verification tradition built up in meteorology and statistics, the economics literature on prediction-market efficiency, and a fast-growing set of studies examining Kalshi specifically.

Improving upon earlier calibration measures used in meteorology, Brier's (1950) proposal to score probability forecasts on the squared distance between the stated probability and the outcome created a rule that rewards honest reporting over hedging or boldness. Later, Murphy (1973) partitioned the Brier score into components attributable into reliability, resolution, and uncertainty, while Murphy and Winkler (1987) codified a general framework for forecast verification. These studies, in addition to Gneiting et al. (2007), who introduce the concept of sharpness, anticipate the language used throughout this paper.

The property that carries the idea from individual forecasters to markets is propriety, formalized by Gneiting and Raftery (2007): a scoring rule is proper if a forecaster who privately believes $\mathbb{P}(Y = 1) = q$ can do no better, in expectation, than to report $p = q$. This is the theoretical bridge between calibration as a statistical criterion and market efficiency as an economic one: a market of self-interested traders has reason to produce calibrated prices only because a payoff-linked contract behaves, in effect, like a proper scoring rule.

The empirical case for markets as aggregators of dispersed information traces to Hayek (1945). Wolfers and Zitzewitz (2004) formalize it for prediction markets, cataloguing applications from the Iowa Electronic Markets to internal forecasting exercises at Google, Hewlett-Packard, and Intel. Cowgill and Zitzewitz (2015), examining corporate markets at Google, Ford, and others, confirm a high degree of forecast reliability and show that individual accuracy within these markets varies systematically with a trader's experience and social proximity to the event being forecast. This is evidence that the aggregation mechanism, and not simply the presence of a large trader pool, is doing real work. Bassamboo et al. (2015) report a similar result in a corporate operations setting, where internal prediction markets used for demand forecasting produced accurate, well-calibrated forecasts once a sufficiently large and diverse pool of employees participated, a precondition echoed directly in the volume- and trader-count results of Sections 4.5–4.8 below.

Manski (2006) offers an important caveat: a market price is not, in general, a sufficient statistic for the average belief of participants without further assumptions on risk preferences and trading behavior. In response, Gjerstad (2004) and Wolfers and Zitzewitz (2006) show that log-utility position sizing, together with a symmetric distribution of beliefs, is sufficient to recover the average-belief interpretation exactly. The strongest direct evidence for the aggregation story comes from Berg et al. (2007), who found that Iowa Electronic Markets vote-share prices were closer to the eventual outcome than contemporaneous national polls in 74% of comparisons across five U.S. presidential elections, an advantage that held even more than 100 days before the election, when polls still had ample time to close the gap and did not.

A related literature documents a favorite-longshot bias in market prices, in which longshots trade too rich and favorites too cheap (Ottaviani and Sørensen, 2008; Snowberg and Wolfers, 2010). Page and Clemen (2013) trace a version of the bias in InTrade data to ordinary time discounting of a distant payout, and Bürgi et al. (2026) document a comparable pattern on Kalshi, where contracts

priced below 10 cents underperform what a break-even trader would require. This bias is not the focus of the present paper, so we largely set it aside.

Closest to the present study in both data and spirit is Diercks et al. (2026), who use Kalshi's macroeconomic markets on inflation, payrolls, GDP, and the federal funds rate to build a real-time benchmark of market expectations and compare it against the New York Fed's Survey of Market Expectations and Bloomberg consensus forecasts. They report that Kalshi's median and modal federal funds rate forecasts achieve a perfect record the day before each FOMC meeting, a statistically significant improvement over fed funds futures, and that its core CPI and unemployment forecasts are statistically indistinguishable from Bloomberg consensus while its headline CPI forecasts significantly outperform that consensus. Those findings are consistent with the smooth, near-linear improvement in Economics-category calibration we document in Section 4.3.2, and gives that category a significance for researchers and policymakers well beyond its modest share of total platform volume. Taken together, this literature suggests that good calibration close to resolution, and its gradual sharpening with time, participation, and category, is less a property of any one exchange than a recurring feature of how prediction markets process information. With that background in place, we turn in Section 3 to the formal definitions of calibration and the Brier score that structure the empirical results that follow.

3 Setup and Definitions

Before moving to the empirical results of our calibration study, it is first useful to provide the methodological background that underpins the study of calibration.

3.1 The Basics of Calibration

Let $Y \in \{0, 1\}$ be a binary outcome (e.g., "candidate wins," "contract resolves YES"). A forecaster issues a probability forecast $p \in [0, 1]$ before Y is realized. Over many forecasting instances, we observe pairs (p_i, Y_i) for $i = 1, \ldots, N$. A forecaster is *perfectly calibrated* if, among all instances where the quoted forecast p, the outcome occurs with empirical frequency p:

$$\mathbb{P}(Y = 1 \mid \hat{p} = p) = p \qquad \text{for all } p \in [0, 1]$$

In practice, since p is continuous, calibration is assessed over bins $B_k = [p_k, p_k + \delta)$:

$$\frac{1}{|B_k|} \sum_{i \in B_k} Y_i \;\approx\; \bar{p}_k, \qquad \text{where } \bar{p}_k = \frac{1}{|B_k|} \sum_{i \in B_k} p_i$$

Plotting observed frequency (y-axis) against forecast probability (x-axis) gives the reliability diagram; perfect calibration is the 45° line.

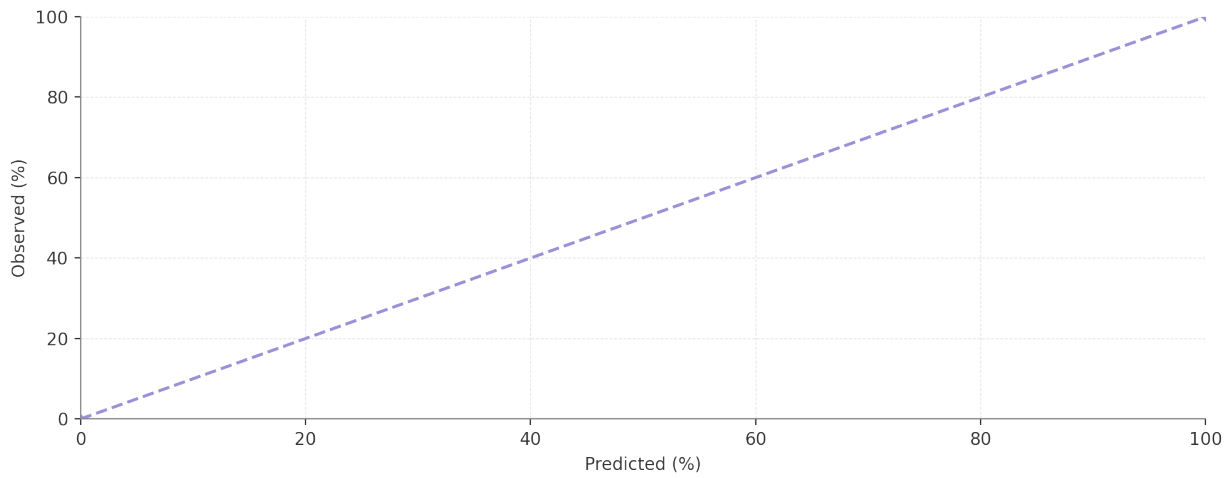

Figure 1: Calibration Diagram

3.2 The Brier Score

The Brier score (Brier, 1950) is the standard proper scoring rule for binary probability forecasts:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - Y_i)^2$$

Lower is better; $BS = 0$ is a perfect forecast, and $BS = 0.25$ is what results from always guessing $p = 0.5$ against a 50/50 base rate. We can effectively consider this to be the distance between the estimate made and the observed frequency.

3.3 The Probability-Theory Justification: The Law of Large Numbers gives calibration empirical content

Probability is only testable through repetition. If a forecaster claims $\mathbb{P}(Y = 1) = p$ for a class of events, the only way to check this claim empirically is to pool many instances where p was forecast and check the realized frequency, because by the Law of Large Numbers:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{\text{a.s.}} \mathbb{P}(Y = 1 \mid \hat{p} = p) \qquad \text{as } n \to \infty.$$

Calibration is precisely the statement that this limiting frequency equals the stated p. Without calibration, probability forecasts are unfalsifiable opinions; with it, they make a testable prediction about long-run frequencies.

4 Calibration Results

The proceeding Section walks through the results of our calibration analysis. We first present full-sample calibration results, namely the Brier scores for all markets excluding Exotics. Then as a result of the potential sampling bias introduced by the late listing of markets we display results with markets held constant from 1-month out, and then present an additional sub-sample that excludes markets where information revelation happens "in-event" (e.g. sports and mentions) and further exclude short-dated daily, hourly, and 15-minute markets. We then present tail-truncated and time-truncated results.

4.1 Full Sample Calibration

We begin with the most basic test available: pooling every contract on Kalshi and examining how the Brier score evolves as resolution approaches. Figure 2 presents this evolution under two alternative constructions of the data. The first ("Per-horizon") includes every contract with a price observation at a given horizon, so that the 3-Month bar reflects only the considerably smaller set of markets that were already open three months before they resolved, while the Close bar reflects nearly the full sample. This is a by-product of markets being created at varying times to resolution. The second ("Fixed at 1-Month") restricts every horizon to the same fixed cohort of 35,562 markets that existed a full month before resolution, and asks how the calibration of that specific set of markets evolves as their own resolution approaches. The distinction is material at the 1-Day horizon: the per-horizon Brier score of 0.126 is driven by the approximately 492,000 short-fused daily, hourly, and sub-hourly markets—predominantly sports and crypto markets—that exist only in the final day before resolution and are inherently more difficult to forecast. The fixed-cohort Brier score for the same 35,562 longer-dated markets, observed exactly one day before they resolve, is a substantially lower 0.039. Both constructions agree, however, on the overall shape of the result: Brier scores decline steadily as resolution nears, from approximately 0.087 at three months to below 0.02 at close.

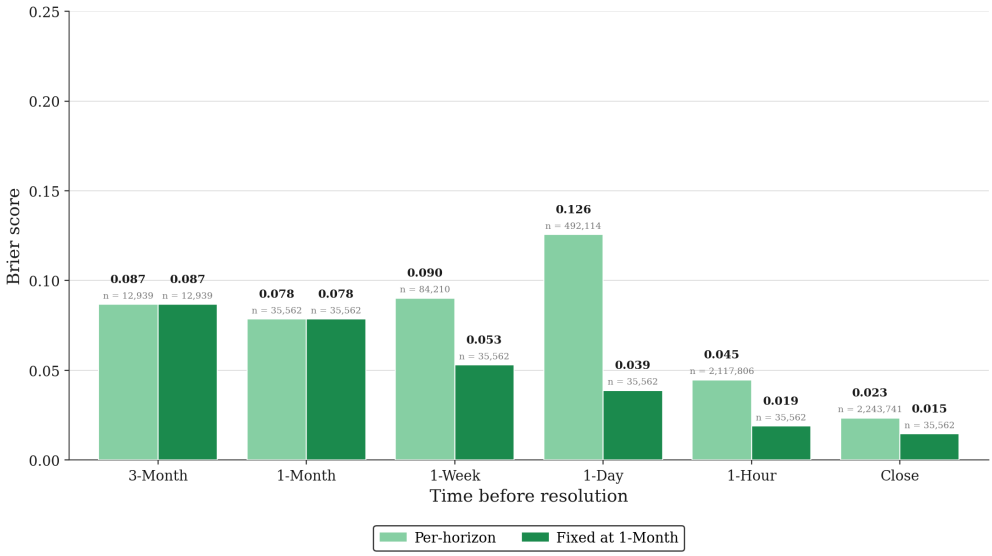

Figure 2: Brier score per horizon versus fixed cohort.

Figure 3 repeats this exercise on the subsample of markets used throughout Sections 4.2–4.3, which excludes Sports, Mentions, and purely intraday (daily, hourly, and sub-hourly) markets. With the noisiest and shortest-fused categories removed, the anomaly at the 1-Day horizon disappear. Brier scores decline monotonically, and the gap between the per-horizon and fixed-cohort constructions narrows substantially (0.052 versus 0.035 at 1-Day, and 0.030 versus 0.016 at Close), confirming that most of the earlier discrepancy was compositional in origin rather than evidence of a failure of calibration to improve with time.

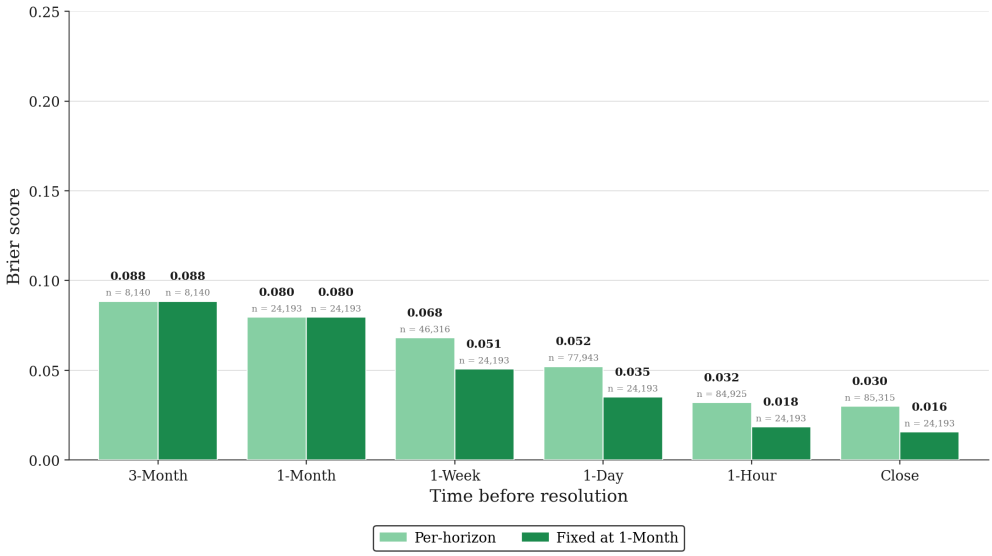

Figure 3: Brier scores with Sports, Mentions, and intraday markets removed.

4.1.1 Calibration Plots

Brier scores summarize calibration in a single number, whereas reliability diagrams reveal its shape. Figure 4 plots observed frequency against predicted probability for four horizons, pooling all markets outside of Exotics (otherwise known as "Combos"). Even at the 3-Month horizon the curve tracks the diagonal reasonably closely, although with visible bowing in the middle bins (40–60%), where bin populations are smallest and most sensitive to sampling noise. By the 1-Day horizon, the fit is nearly exact, and the largest bins—near 0 and near 100 cents—each contain tens of thousands of markets.

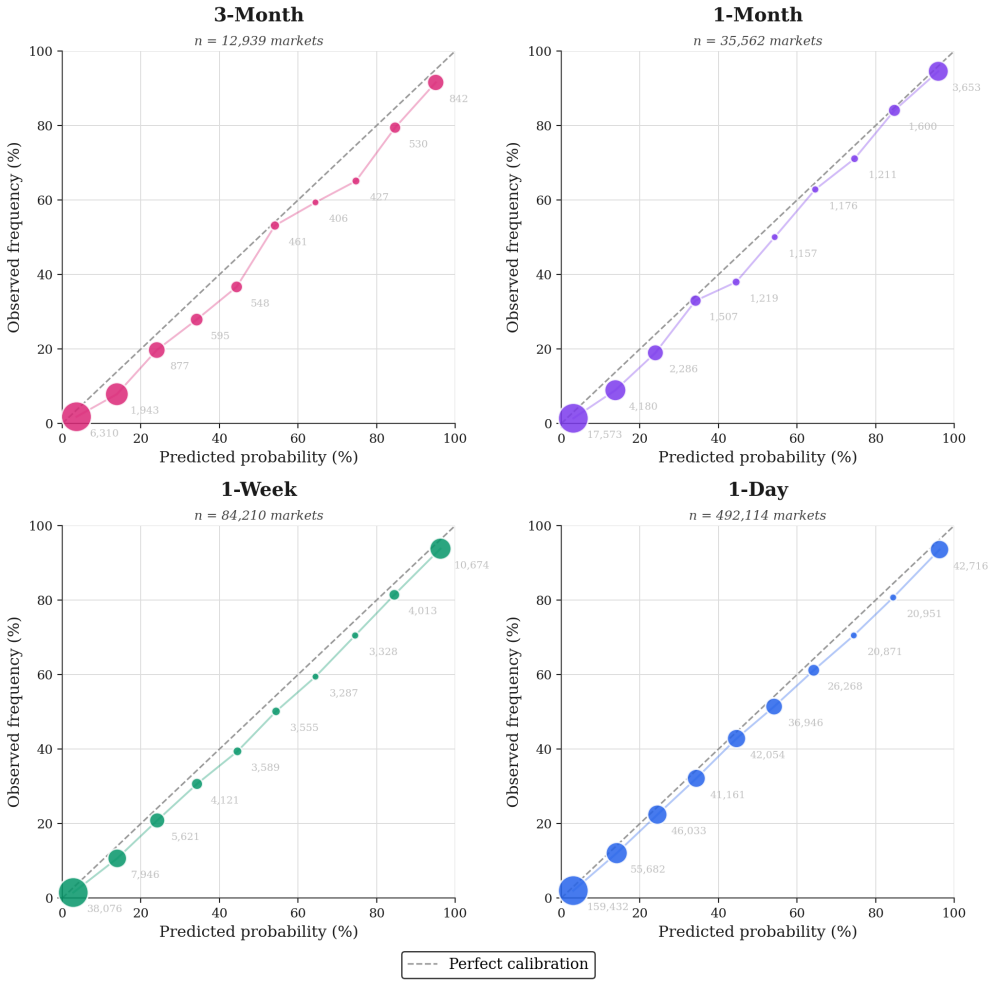

Figure 4: Calibration plots by time horizon on all markets.

4.1.2 Overall Results - Holding 1-Month Constant

Figure 5 repeats the diagram while holding the same 35,562-market cohort fixed across all four horizons, thereby isolating how the calibration of a specific set of markets sharpens as their own resolution approaches, independent of any change in the underlying composition of what is being measured. The improvement is visible directly: bin populations barely change between panels, since the same markets are tracked throughout, while the fit to the 45° line visibly tightens moving from 3-Month to 1-Day.

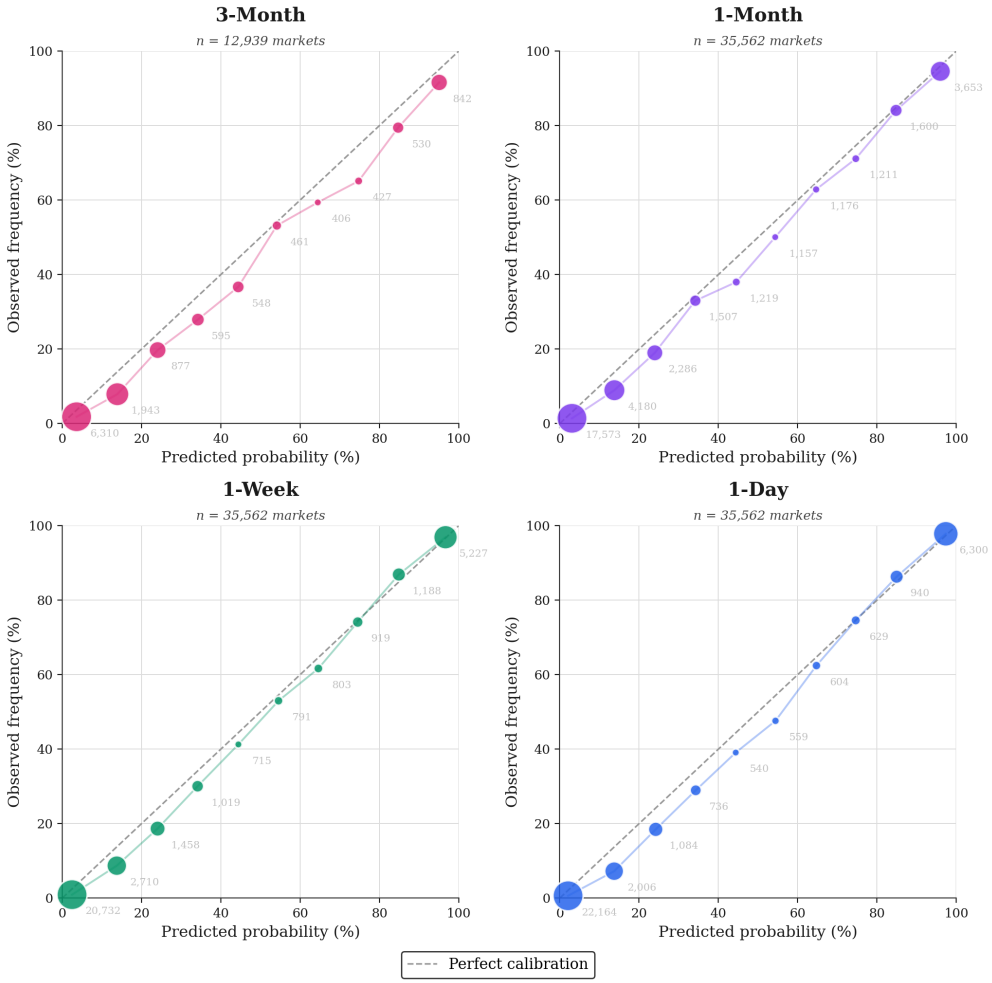

Figure 5: Calibration plots by time horizon on all markets with fixed 1-month cohort.

4.1.3 Subsample Calibration

Figure 6 removes Sports, Mentions, and purely intraday markets altogether. The resulting curves are the tightest reported in this paper, essentially indistinguishable from perfect calibration by the 1-Week and 1-Day horizons. This documents heterogeneity in the calibration between categories.

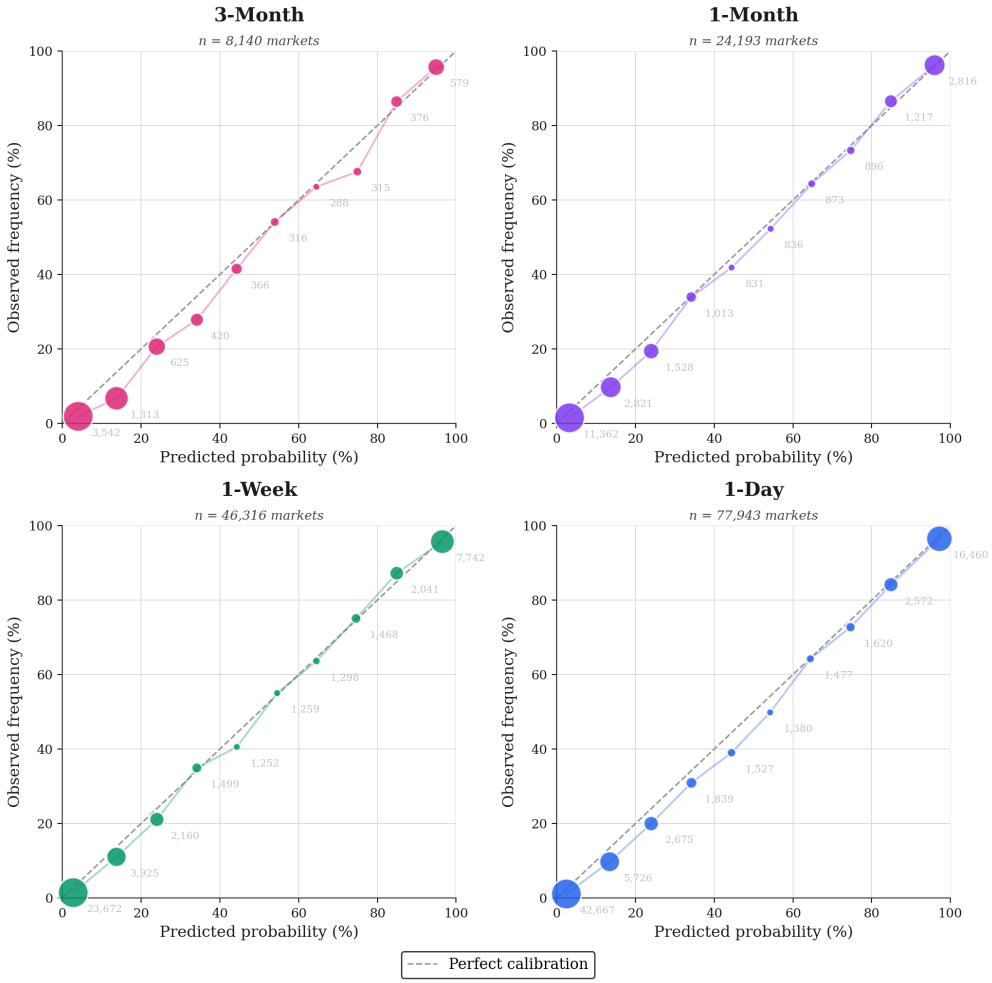

Figure 6: Calibration plots for market subsample.

4.2 Tail Truncated Results

A natural question that may arise about any aggregate Brier score is how much weight it is placing on markets that are already close to certain in resolution (i.e. a contract trading at 1 cent or 99 cents). Figure 7 examines this directly by comparing the subsample introduced above to a tail-truncated version that retains only markets priced above 2 and below 98 cents at the relevant horizon, that is, only the subset of markets that remained genuinely uncertain at that moment. Of course, tail truncation has no natural limits or enforced parameters on extremes, so this is a first-degree approximation of drawing a line regarding market uncertainty (though some may justify other thresholds they believe more relevant also).

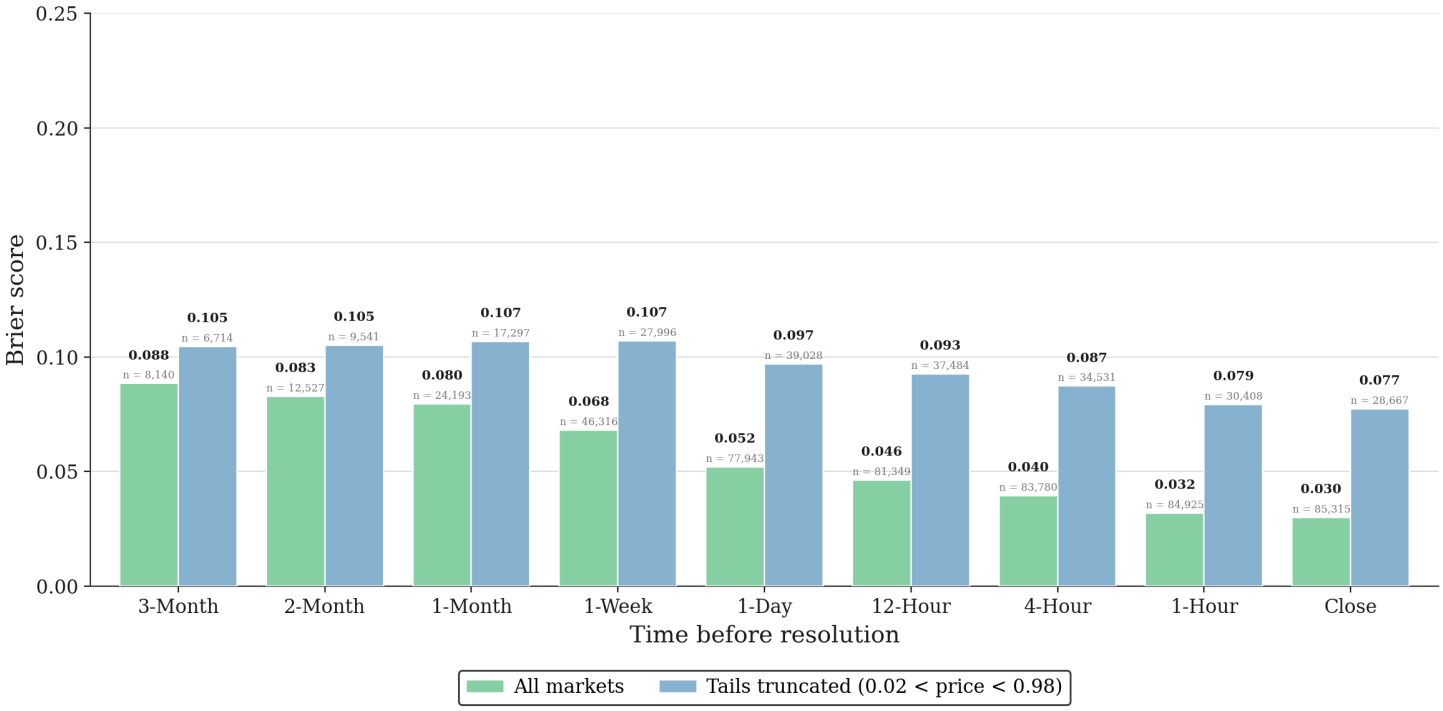

Figure 7: Brier score per horizon versus tail truncated cohort on market subsample

As expected, the tail-truncated Brier score is higher than the corresponding full-sample figure at every horizon: 0.077 versus 0.030 at market close; 0.097 versus 0.052 at the 1-Day horizon; and 0.105 versus 0.088 at the 3-Month mark. To some extent, this should not be surprising: Removing well-calibrated samples is likely to make calibration appear worse. Though this is a popular method by which to establish resolution (i.e. how discerning a market is), there are two additional considerations to keep in mind. First, pricing an event with high confidence is not a free statistical artifact. Establishing that an outcome is genuinely 99% likely, or symmetrically that it is only 1% likely, requires real information. That these confident prices go on to resolve correctly at the rates implied by the reliability diagrams in Sections 4.1 is exactly the kind of resolution that the governing principle of the discipline – maximizing sharpness subject to calibration – identifies as a hallmark of forecasting skill rather than a loophole in the Brier score. Second, and more directly, even the harder and more contested subset of markets remains well calibrated in an absolute sense: a Brier score of 0.077 at close, or 0.105 at three months, is still far superior to the 0.25 that a coin-flip forecaster would earn against a 50/50 base rate. This is achieved on precisely the set of markets that were more difficult to price.

4.3 Category Calibration

We now present calibration results separately by category to convey the potential structural differences between market types.

4.3.1 Overall

Table 1 reports Brier scores separately for each of the eleven categories in the sample, at every horizon from three months to market close.

Category	N	3-Mo	2-Mo	1-Mo	1-Wk	1-Day	1-Hr	Close
Climate	112,242	0.069	0.065	0.128	0.051	0.112	0.013	0.013
Commodities	32,473	0.155	0.133	0.098	0.076	0.104	0.034	0.031
Crypto	686,055	0.132	0.107	0.081	0.094	0.082	0.026	0.024
Economics	16,701	0.108	0.109	0.100	0.089	0.073	0.066	0.066
Elections	6,064	0.073	0.072	0.054	0.033	0.022	0.006	0.004
Entertainment	47,399	0.088	0.084	0.080	0.073	0.056	0.037	0.034
Finance	217,250	0.075	0.079	0.075	0.069	0.085	0.032	0.032
Mentions	58,294	0.060	0.192	0.149	0.180	0.160	0.014	0.007
Politics	11,312	0.085	0.066	0.068	0.069	0.053	0.016	0.014
Sports	1,133,757	0.084	0.076	0.069	0.107	0.155	0.063	0.022
Tech & Science	2,528	0.108	0.117	0.083	0.056	0.035	0.016	0.015

Table 1: Brier score by category and time horizon before resolution. N is the total number of resolved markets per category.

Two patterns stand out. First, all market types report calibration substantially better than random chance throughout the entire sample period. Second, there exist differences in between-category calibration that may be caused by one of three key factors that we identify: (1) differences in times of market listing that increase sample size only at certain time horizons, and in so doing, adversely affect calibration outcomes as price discovery for these shorter-lived markets is yet to have substantially occurred, (2) Election markets are prone to significant settlement delay, making their Brier scores biased downward (we address this in Section 4.4), and (3) certain market categories are structurally more difficult to predict than others. For instance, Mentions report worse calibration than Politics, but (a) much of the information about the contents of a given speech or address are only available as of the beginning of that speech or address as it sets the tone for points of discussion, and (b) identifying singular words likely to be uttered is on-face more difficult than determining reasonable positioning in markets with a plethora of existing pricing information able to be cross-referenced (e.g. news, party ideology, etc.).

4.3.2 Double-clicking into Economics

Unlike others, Economics markets exhibit monotonicity. Brier scores decline steadily and almost linearly from 0.108 at three months to 0.066 at close. This pattern is consistent with the fact that most Economics markets resolve against a pre-scheduled, unambiguous release (a CPI print, a payrolls report, a GDP estimate) rather than against a live, continuously evolving event whose outcome crystallizes only gradually, as in a sports game. It is also consistent with the broader case made by Diercks, Katz, and Wright (2026) that Kalshi's macroeconomic markets constitute a genuinely informative, high-frequency benchmark of market expectations: a category that resolves against scheduled, well-defined releases is precisely the setting in which a price-discovery mechanism should be expected to sharpen smoothly as the release date approaches, and the absence of any hump or reversal in Figure 8 is consistent with that expectation.

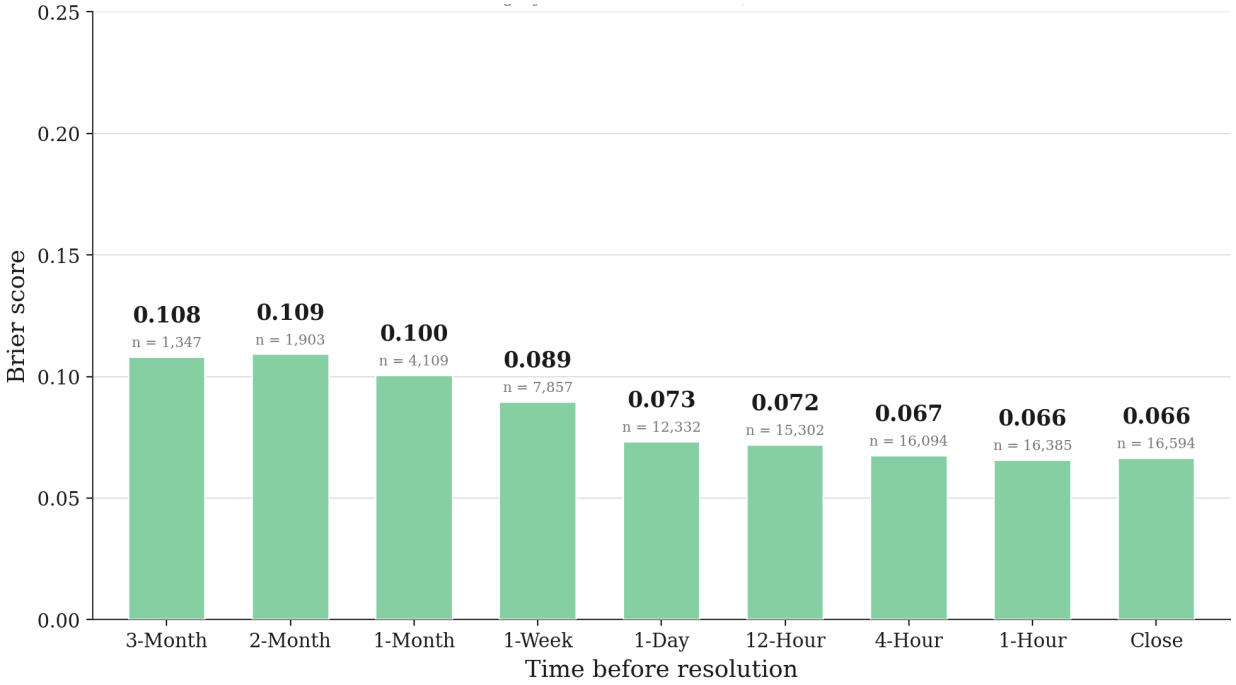

Figure 8: Brier scores of Economics markets.

4.4 Date Truncated

Every horizon label used in this paper, including those in Table 1, is computed relative to the timestamp at which Kalshi formally closes a market. While a common practice, it also represents an administrative clock, rather than a real-world one. The moment a platform closes a contract can lag (and structurally must lag), the moment at which the underlying real-world event that the contract is actually about becomes known. In some cases this is not applicable, such as in Economics markets where the market itself closes for trading prior to a statistical print, but in others, such as Sports or Elections, it may have a significant impact. Truncating a market's history to the actual date on which the underlying event occurred, rather than to the date the platform recorded as closed, can therefore add a further, complementary layer of information beyond what the close timestamp alone provides. This section examines both cases in the data where this distinction is most salient and shows that re-anchoring to the true occurrence of the underlying event moves the calibration picture in opposite directions for the two categories.

We first move to an analysis of Sports markets. Sports market closing times often lag the games' end by a few minutes. For the first time in the study of prediction market calibration, we use a proprietary dataset containing to-the-second game-end timestamps to re-base our analysis. If a substantial amount of the deterioration in calibration during the final minutes of a game that some of the literature attributes to Sports markets actually extends beyond the end of the game, into a "waiting for resolution" period, further research should focus on that window, since it could explain why calibration deteriorates at the close. Figure 9 tests this hypothesis by re-anchoring the clock to the moment the underlying game actually concludes, rather than to Kalshi's market close timestamp.

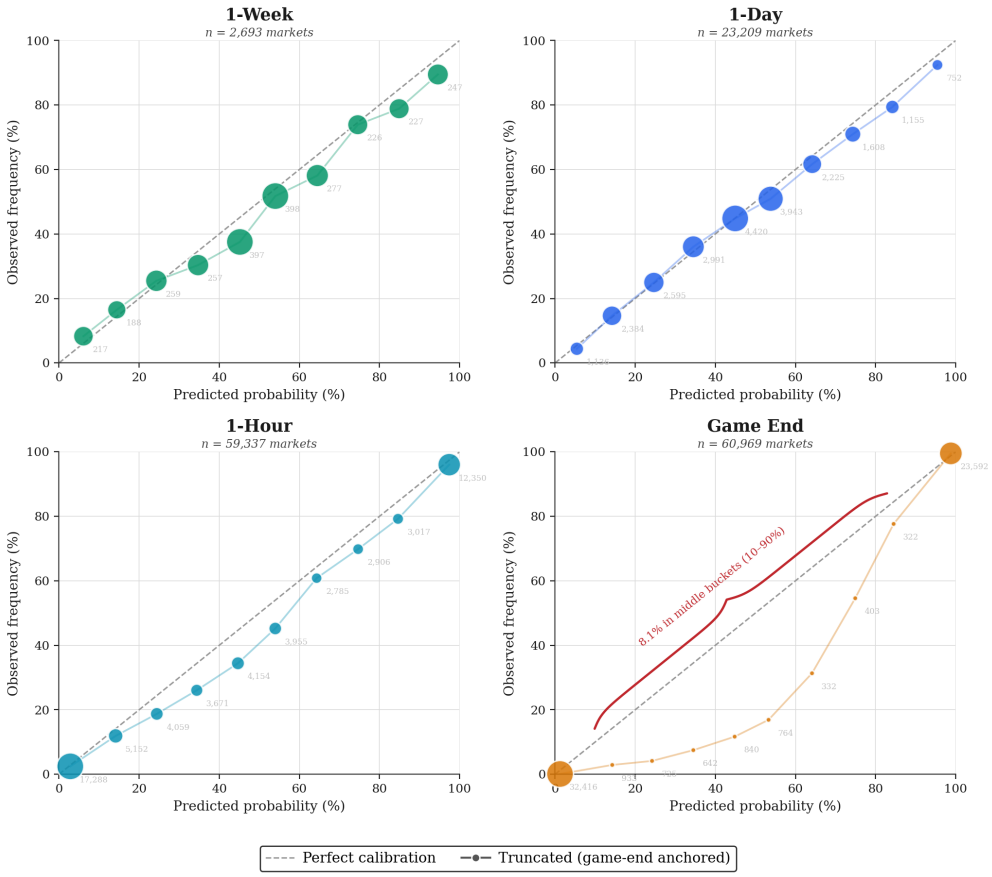

Figure 9: Calibration of game-end anchored Sports markets.

Anchored in this manner, Sports calibration remains excellent at every true horizon: the 1-Week, 1-Day, and 1-Hour panels all track the 45° line closely, with the largest bins (near-certain markets) situated almost exactly on the diagonal. In the instant the game ends, we see immense concentration at the tails. This is more likely than not largely a mechanical property – that there is a brief information-propagation lag between a game's true conclusion and the moment at which Kalshi's order book has fully absorbed that outcome. In some cases, these are also constituted by cases where the markets in question, asymmetrically multi-outcome bracket events rather than standalone binary outcomes, are themselves thinly traded and/or lower liquidity and as a result are vulnerable to stale pricing.

On the other hand, Election market closing times can lag by days or weeks. Figure 10 compares the original, market-close-anchored Brier scores already reported in Table 1 to two alternative constructions: prices measured relative to the actual date of the election itself ("Election-day anchored"), and a further restriction to markets asking specifically which candidate wins a given race ("Candidate victory markets only").

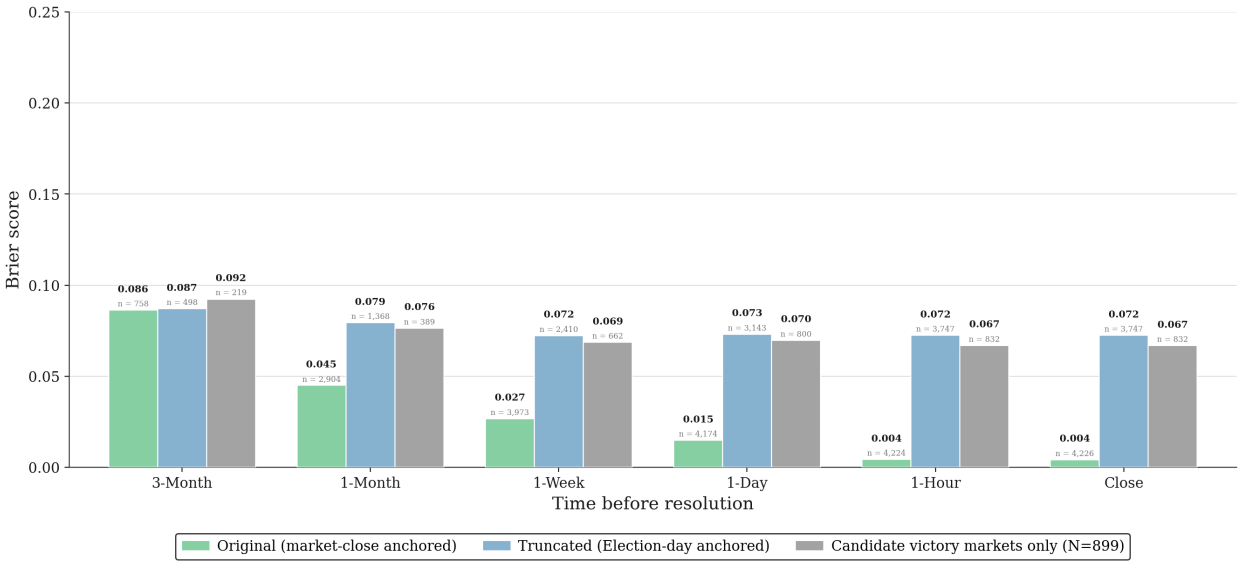

Figure 10: Brier scores of Election markets using election day as anchor.

Anchoring to election day renders calibration extremely strong, but naturally *worse*, not better, at every horizon: the 1-Month Brier score rises from 0.045 (market-close-anchored) to 0.079 (election-day-anchored) and 0.076 (candidate-victory-only). The most plausible explanation is a mirror image of the Sports case. Many Election-category markets with high certainty resolve at calendar dates well after the substantive election-night uncertainty has already been settled by voters because the markets were contractually, until November 2025, unable to be resolved earlier than the date of vote certification, or, in some cases, inauguration, which may occur weeks to months after an election is held. Once the clock is anchored to election day itself, the sample is dominated by the genuinely more difficult pre-election forecasting problem, and the true difficulty of that problem is revealed to be greater than the raw, close-anchored Brier scores suggested. Though novel here, we believe that this approach represents a more faithful interpretation of Election calibration and hope for it to become the new industry reporting standard.

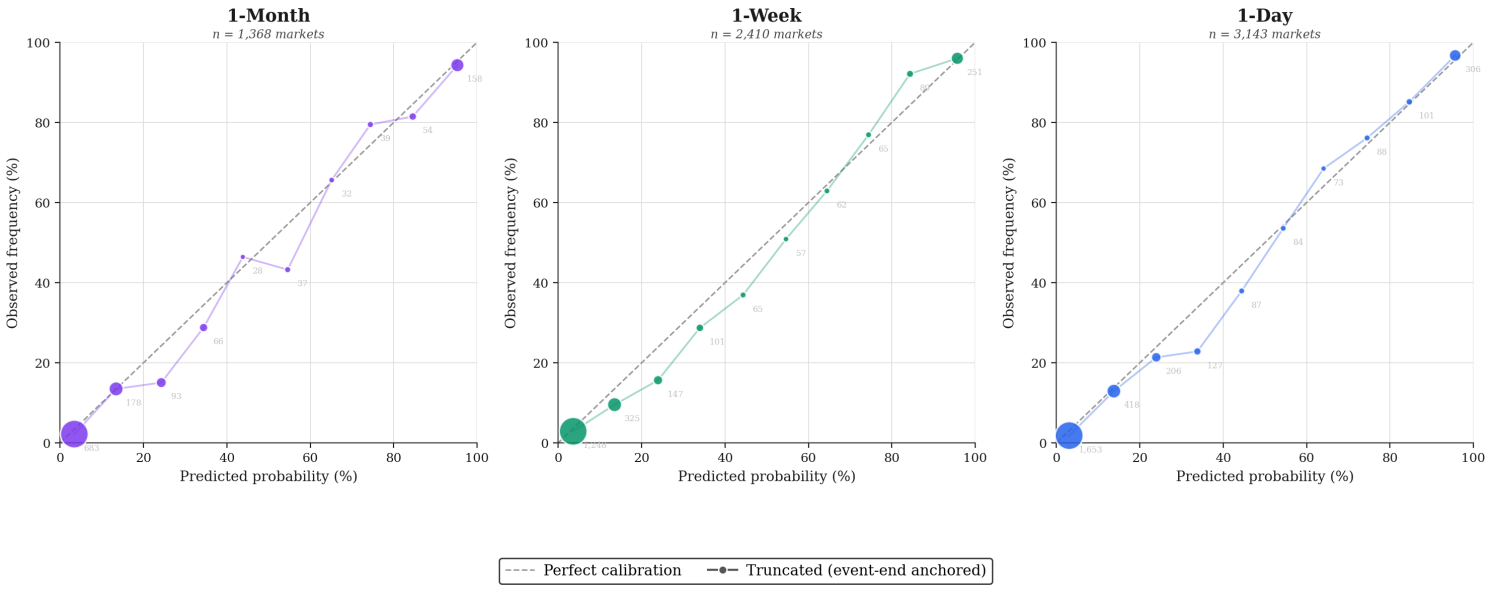

Figure 11: Calibration of Election markets using election day as anchor.

Figure 11's reliability diagrams for the election-day-anchored sample show a reasonable, if noisier, fit to the 45° line at the 1-Month, 1-Week, and 1-Day horizons, with the additional noise attributable to comparatively small sample sizes (1,368 to 3,143 markets, as against tens of thousands for other categories at similar horizons).

4.5 Calibration by volume

The above analysis explored calibration as a function of time. We now turn to a second dimension: the depth of participation in a given market, measured either by total dollar volume traded or by the number of unique traders. Figures 12–13 plot Brier score against an event-volume threshold on filtered markets (excluding Sports, Mentions, and purely intraday markets), separately by time horizon and, in Figures 14–15, by category. We find that Brier scores decrease as a function of volume, though non-monotonically across some categories, and largely monotonically as a function of trader number.

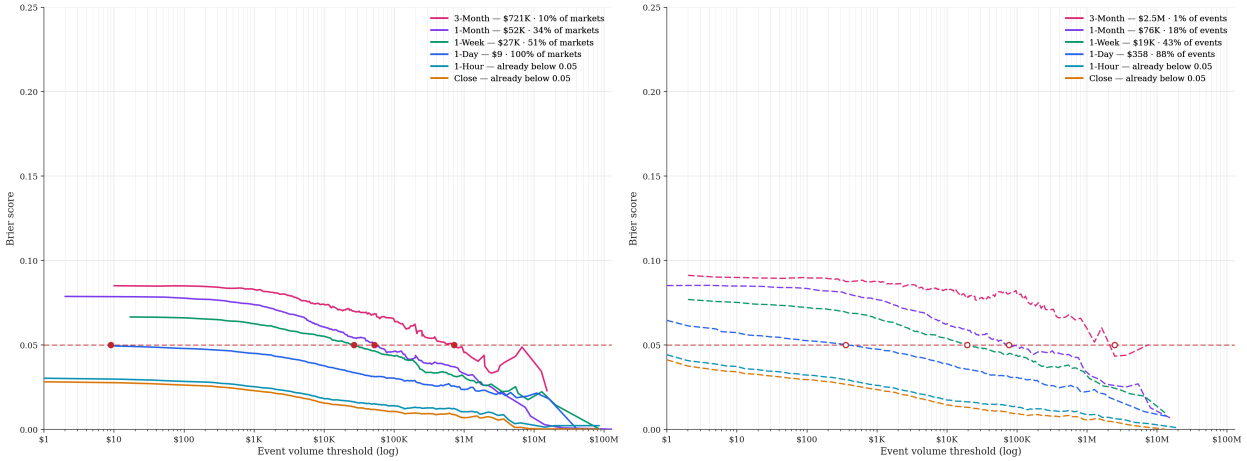

Figure 12: Brier scores by volume on market subsample. The chart on the left represents one observation per market, while the dashed lined chart on the right represents one observation per event.

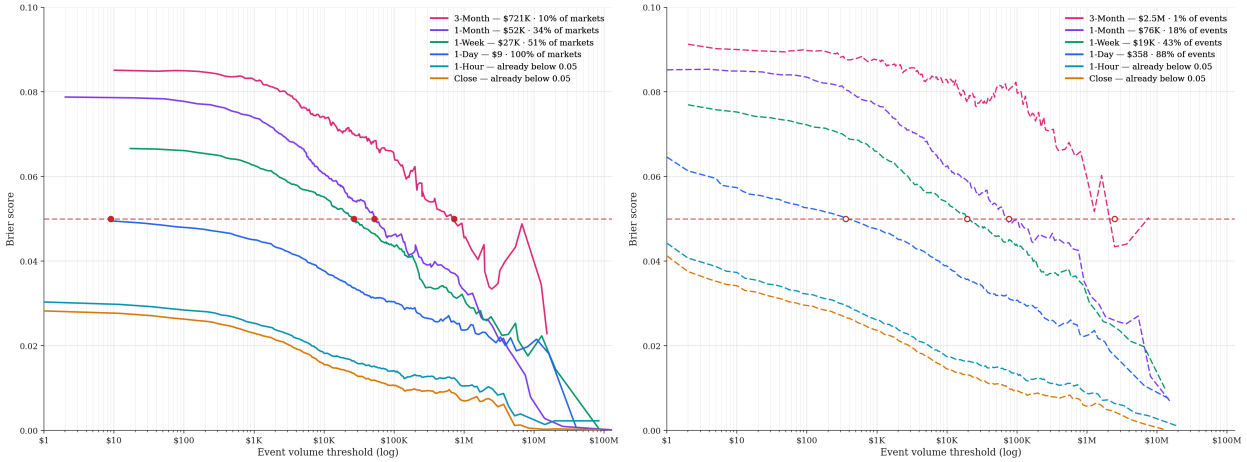

Figure 13: Brier scores by volume on market subsample (enlarged).

The chart is unambiguous and highly robust: within every time horizon, Brier scores fall monotonically as the volume threshold rises, with noisier behavior confined to the extreme right tail of the distribution, where sample sizes shrink to a handful of markets. This pattern is suggestive of something like a learning curve for a market, where accuracy improves as more trading activity accumulates, much as a neural network's error falls with more training data.

4.6 Brier Score by Event Volume Threshold by Category

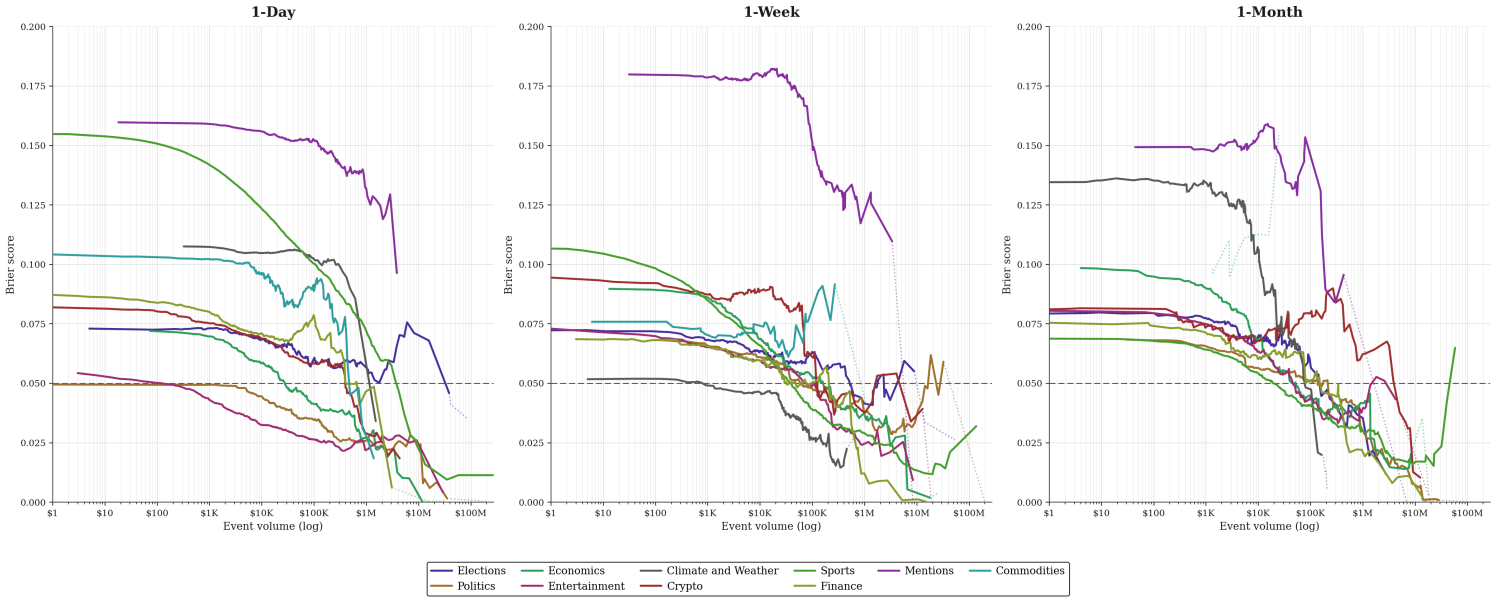

Figure 14: Brier scores by volume on all categories and horizons.

Breaking these curves down into their categorical components reveals further insights. Different categories cross the 0.05 Brier-score reference line at very different volume levels. Figure 14 shows that, at the 1-Day horizon, Politics and Economics markets are already below 0.05 at volumes of a few thousand dollars, while Sports and Mentions cross the threshold at roughly two orders of magnitude more volume. This represents divergent underlying dynamics in the difficulty of prediction that require further analysis. However, one could say that, on average, it is more challenging to forecast Mentions, Sports, and Climate markets compared to Entertainment, while Elections offer a sort of middle ground. There are also likely confounding variable driving volume – Mentions markets may require more volume to be predictive, but Sports markets are likely to present higher volumes due to trader interest, not due to a requirement of such volumes.

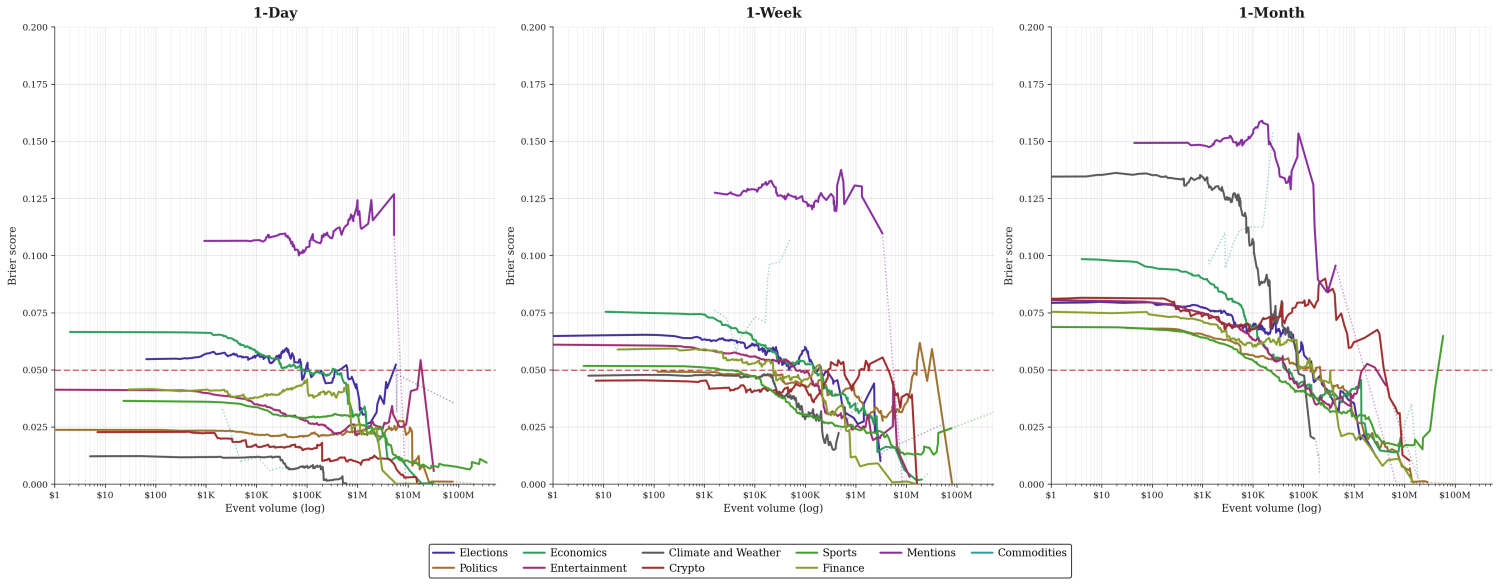

Figure 15: Brier scores by volume on all categories fixed constant at 1-month and horizons.

Figure 15 repeats the volume analysis while holding the market cohort fixed at the 1-Month mark across all three horizon panels, isolating the pure cross-sectional relationship between volume and calibration from any horizon-driven confound. The monotonic decline survives this stricter test, though selection effects alter the data significantly.

4.7 Calibration by trader number

Figure 16 repeats the exercise using the number of unique traders rather than dollar volume, on the same filtered sample. Long-horizon markets (3-Month) never cross the 0.05 Brier-score threshold regardless of how many traders participate; 1-Month markets reach it only at approximately 1,900 traders, covering just 1% of markets at that horizon; 1-Week markets cross the threshold at only 901 traders; and 1-Day markets cross it at a mere 20 traders, a level already exceeded by 40% of markets at that horizon. Depth of participation, in other words, improves calibration, but only up to a limit set by how much genuine uncertainty remains at a given horizon: no amount of additional trading fully substitutes for the passage of time.

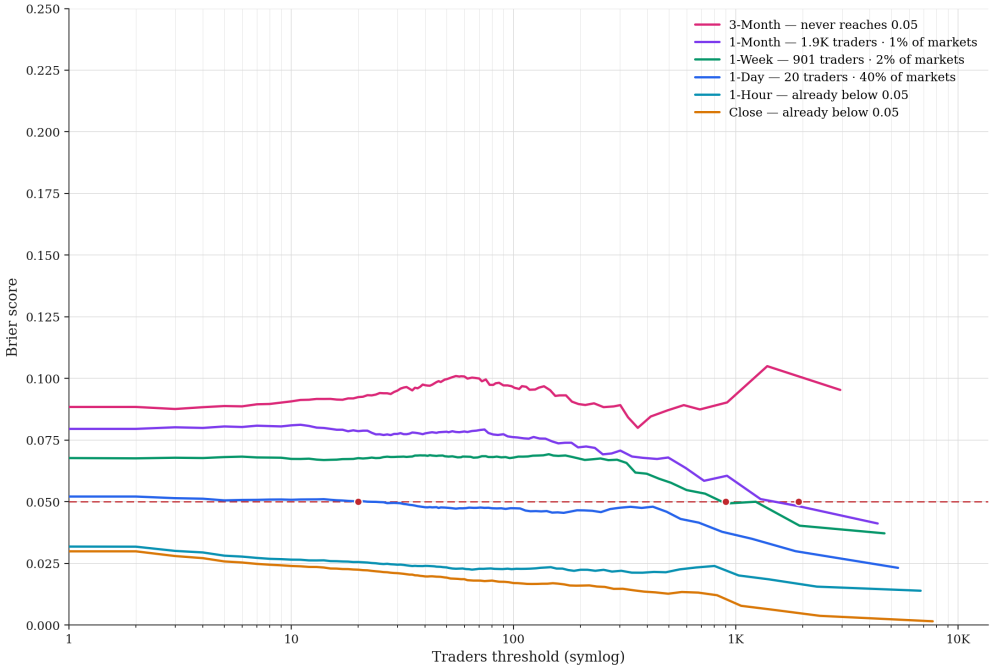

Figure 16: Brier scores by number of traders on subsample.

4.8 Event volume x Time

Tables 2 and 3 make the volume and trader-count relationships precise, showing that time and volume both matter for calibration. Per the table and chart below, calibration improves as a function of volume across time horizons, moving from 0.0635 at Close for markets with less than $10k in total event volume to 0.0099 for those with volume greater than $200k. This effect is essentially monotonic with one exception at the 1-Month mark for the lowest-volume markets, likely attributable to the significant addition of new markets at that time.

Event volume bucket *Brier · n*	Close	1-Hour	1-Day	1-Week	1-Month	3-Month
< $10K	**0.0635** 22,547	**0.0645** 22,237	**0.0815** 21,888	**0.0944** 14,131	**0.1087** 8,758	**0.1042** 2,989
$10K – $50K	**0.0211** 22,719	**0.0229** 22,663	**0.0468** 20,682	**0.0670** 12,648	**0.0732** 6,238	**0.0821** 1,949
$50K – $200K	**0.0140** 16,070	**0.0173** 16,056	**0.0343** 13,799	**0.0536** 8,240	**0.0608** 3,657	**0.0751** 1,148
≥ $200K	**0.0099** 18,168	**0.0132** 18,164	**0.0290** 16,000	**0.0403** 9,231	**0.0426** 4,304	**0.0627** 1,532

Table 2: Brier scores by event volume bucket and time horizon.

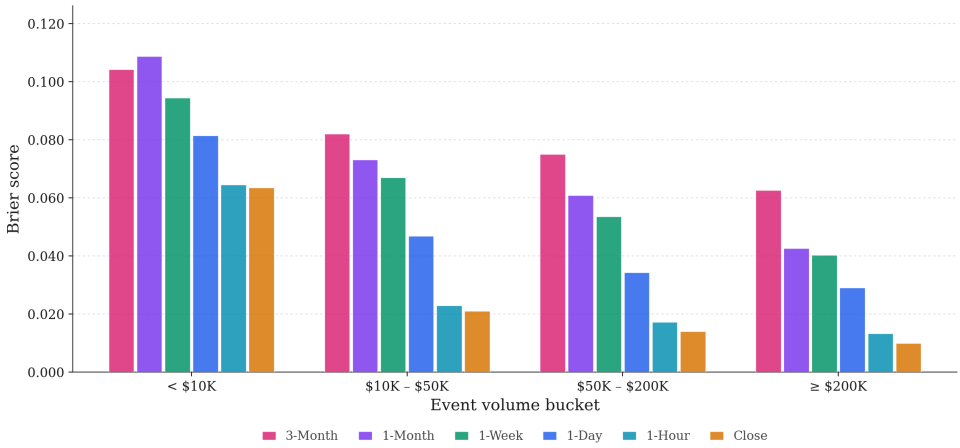

Figure 17: Brier scores by volume buckets and time horizons.

A comparable fourfold gap (0.0354 to 0.0089) appears across trader-count buckets when moving from lowest trader count to highest at Close. One relationship reverses direction at long horizons, however: at the 3-Month mark, markets with at least 1,000 traders are slightly *worse* calibrated (Brier score 0.0929) than markets with fewer than 20 traders (0.0855). The most plausible explanation is a confound between participation and genuine event salience rather than a causal effect of crowding. The markets that have already accumulated a thousand traders three months before they resolve tend to be the platform's highest-profile events—national elections, championship games—which are, almost by construction, intrinsically more difficult to forecast than an average, low-attention three-month-out market, irrespective of how many participants are trading it.

Trader-count bucket *Brier · n*	Close	1-Hour	1-Day	1-Week	1-Month	3-Month
< 20	**0.0354** 49,036	**0.0366** 49,125	**0.0536** 46,682	**0.0685** 26,925	**0.0801** 15,035	**0.0855** 4,724
20 – 100	**0.0253** 23,805	**0.0270** 23,506	**0.0512** 20,825	**0.0676** 12,757	**0.0798** 5,972	**0.0902** 2,298
100 – 1K	**0.0181** 10,993	**0.0231** 10,855	**0.0487** 9,317	**0.0700** 5,910	**0.0788** 2,793	**0.0974** 1,030
$\geq$ 1K	**0.0089** 1,481	**0.0208** 1,439	**0.0359** 1,119	**0.0508** 724	**0.0587** 393	**0.0929** 88

Table 3: Brier scores by trader-count bucket and time horizon.

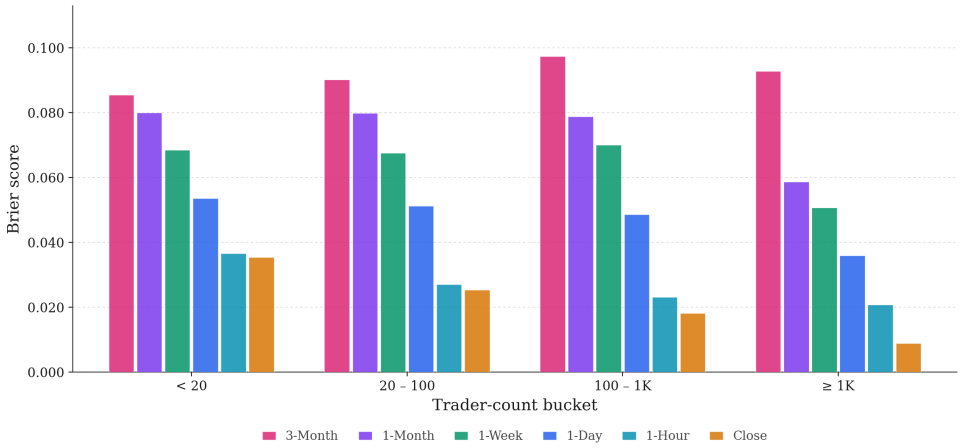

Figure 18: Brier scores by volume buckets and time horizons.

5 Accuracy

5.1 Naive/Lead Accuracy

Section 4 evaluated calibration. We now turn to two coarser but more readily interpretable measures: *naive accuracy* and *lead-market accuracy*. *Naive accuracy* is defined as whether the higher-priced side of a contract turned out to be correct, with 50 cents understood to be rounded to "Yes" (analogously - do the outcomes priced as more likely than not actually happen?). *Lead-market accuracy* is the analogue of naive accuracy for events with more than two mutually exclusive outcomes (i.e. does the single most-favoured candidate among many actually win?). We use our proprietary dataset which separates the two for evaluation.

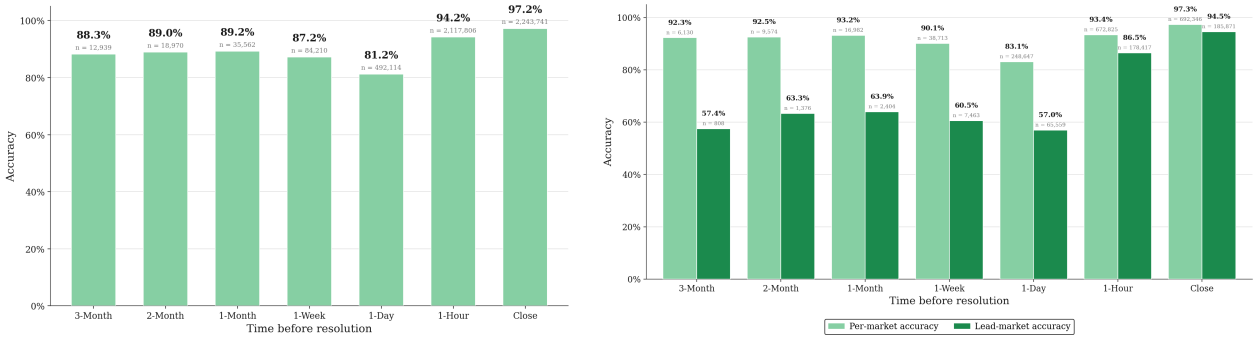

Figure 19: Naive accuracy and lead accuracy by time horizon for all markets, and mutually-exclusive markets.

Figure 19 shows naive accuracy for the full sample rising from 88.3% at three months to 97.2% at close, but with the same sample-driven hump seen earlier; accuracy actually dips to 81.2% at the 1-Day mark before recovering to 94.2% (1-Hour) and 97.2% (Close) largely due to the introduction of short-dated and more difficult-to-predict markets (e.g. Sports and Mentions). Assessed naive accuracy improves over the sample of markets that are mutually exclusive by construction.

Figure 20 confirms this is compositional: on the filtered sample, naive accuracy improves smoothly and monotonically at every horizon, from 87.9% at three months to 96.2% at close, with no such dip.

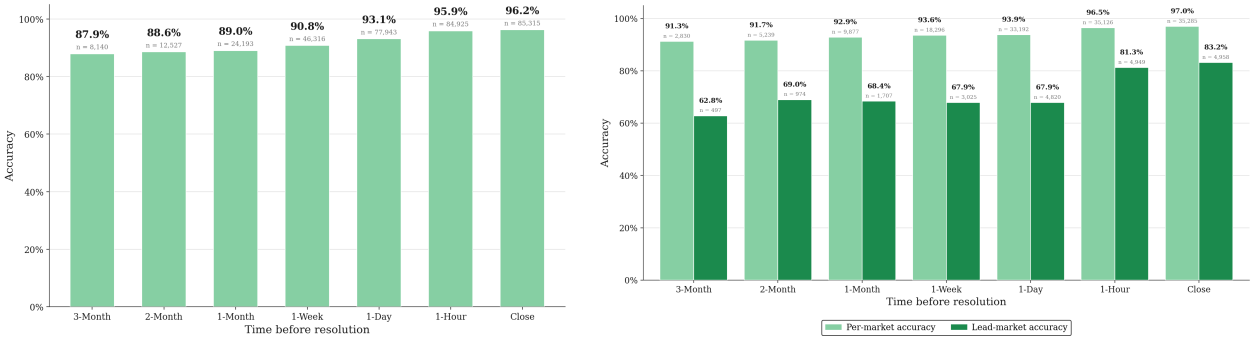

Figure 20: Naive accuracy and lead accuracy by time horizon for mutually-exclusive market subsample.

Lead-market accuracy is a considerably more demanding benchmark than it might first appear, and the levels reported should not be judged against an implicit target of near-certainty. For events with several mutually exclusive outcomes, correctly identifying the single eventual winner is a fundamentally harder task than correctly pricing any one constituent Yes/No contract. If the leading candidate in a typical multi-way race is priced at, say, 60 cents, a well-calibrated market should see that candidate win only about 60% of the time, not systematically more—a market that produced lead-market accuracy in the high 90s across genuinely contested multi-way races would, if anything, be evidence of underpriced favorites rather than of superior forecasting. The relevant comparison is therefore not to 100%, but to a baseline set by the true dispersion of outcomes in a given race, a standard familiar from outside prediction markets as well: a hedge fund whose directional calls are correct only 52% of the time on a coin-flip proposition is ordinarily regarded as highly successful, precisely because the appropriate benchmark is 50%, not 100%.

Figure 19 shows lead-market accuracy on the full sample ranging between approximately 57% and 66% at every horizon from three months through one day, before converging toward per-market accuracy levels by the 1-Hour and Close horizons, as the field of realistic contenders narrows and genuine uncertainty resolves. The filtered sample in Figure 20 exhibits the same qualitative pattern at a somewhat higher level, approximately 63–69% at long-to-medium horizons, reaching the low 80s by close. Read against the correct baseline, these figures represent a meaningful and growing edge over chance in races that, by construction, remain genuinely contested well before they conclude, rather than a shortfall relative to a standard of near-certainty. A forthcoming companion study develops this point further, outlining several alternative ways of conceiving of predictive accuracy for multi-way events that depart from the naive, all-or-nothing criterion used here and that more directly account for the number of live contenders and the degree of genuine uncertainty remaining in a race at a given horizon.

Category	N	3-Mo	2-Mo	1-Mo	1-Wk	1-Day	12-Hr	4-Hr	1-Hr	Close
Climate	112,242	90.9	92.4	80.9	92.9	83.8	88.6	97.8	98.4	98.4
Commodities	32,473	—	—	—	90.6	85.7	87.6	93.0	96.0	96.4
Crypto	686,055	80.2	85.9	89.3	86.2	88.7	89.7	92.7	96.8	97.1
Economics	16,701	86.0	84.9	86.3	87.7	90.1	90.3	91.0	91.2	91.2
Elections	4,226	87.3	88.7	89.5	90.6	90.5	90.3	90.6	90.6	90.6
Entertainment	47,399	88.3	88.2	88.7	90.3	92.5	92.9	94.1	95.2	95.6
Finance	217,250	90.5	89.2	89.6	91.0	88.8	88.0	89.7	96.2	96.2
Mentions	58,294	—	71.0	78.4	72.6	76.1	81.6	86.9	98.1	99.1
Politics	11,312	87.8	91.2	90.6	90.4	92.5	95.0	96.4	98.0	98.4
Sports	1,133,757	88.8	90.0	90.8	84.5	76.1	74.0	75.0	91.6	97.4
Tech & Science	2,528	85.2	83.6	88.5	92.2	95.7	96.5	98.2	98.2	98.3

Table 4: Market accuracy (%) by category and time horizon before resolution, for cells with $n \geq 50$ markets. N is the total number of resolved markets per category.

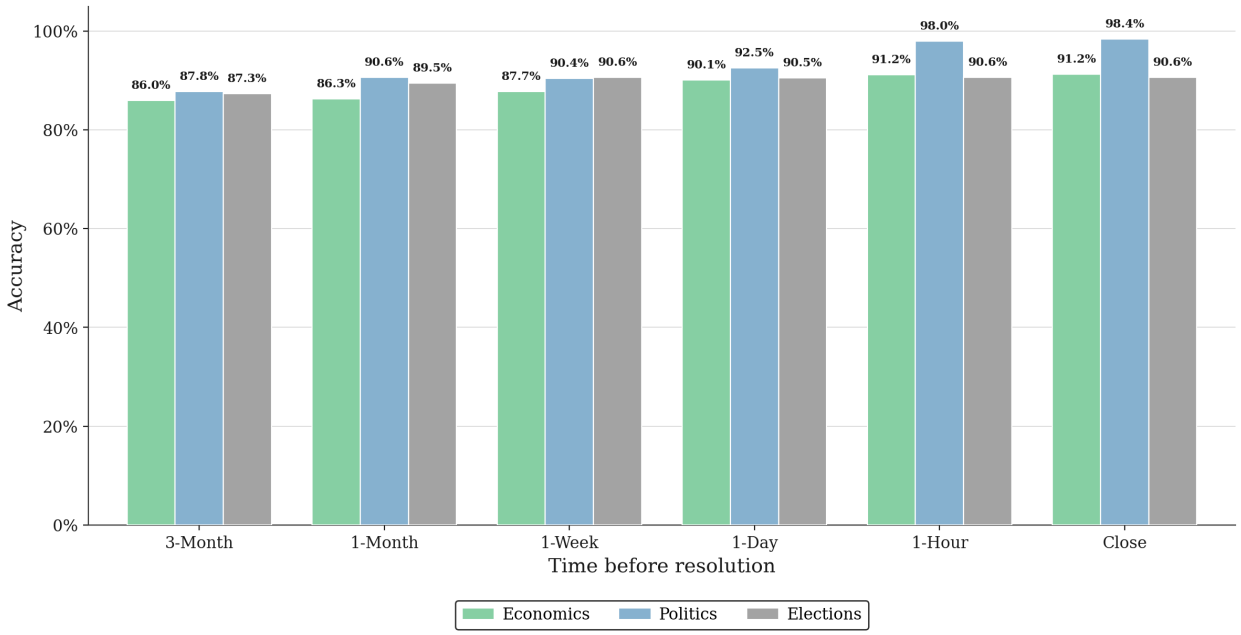

Figure 21: Accuracy scores by category.

Table 4 and Figure 21 disaggregate accuracy by category for the three most closely related domains—Economics, Politics, and Elections. We show that naive accuracy is high across all categories, clear of 85% throughout the entirety of the considered sample. Lead accuracy results by category are reported in Appendix A.

6 Discussion

Considered as a whole, the results in the sections above support a single unambiguous headline finding. **Kalshi's prices are consistently better calibrated than a naive benchmark** across

every dimension examined in this paper, including time-to-resolution, trading volume, trader count, and category. Even the least favorable cell reported in Table 1, a Brier score of 0.192 for Mentions markets at the 2-Month horizon, remains well below the 0.25 benchmark of randomness. Taken together with the reliability diagrams of Section 4.1 and the accuracy figures of Section 5.1, this strongly affirms the claim examined in Section 1: Kalshi's prices behave like genuine probabilities, and increasingly so as resolution approaches. The heterogeneity documented across Sections 4 and 5—by time, by category, by volume, by trader count—is therefore best read as variation *around* a consistently favorable baseline, rather than as evidence that calibration is fragile or occasional.

Two questions, in our view, organize what remains open and define the natural next steps for this research agenda.

1. **Calibration is a joint outcome of time, depth, and category, but is it also a joint outcome of trader skill?** In this paper, we establish that calibration quality is not a fixed property of the platform. It varies systematically, and in every case examined it improves with time, volume, and trader count. These results demonstrate clearly that participation improves calibration; they are considerably less informative about who, specifically, supplies that improvement. The reversal documented in Table 3, in which markets with the most traders are, at the longest horizon, very slightly worse calibrated than markets with the fewest, suggests that the raw number of participants is an imperfect proxy for the quantity that plausibly matters most: The presence of traders with a genuine, correctly expressed informational edge. Determining whether Kalshi's calibration is produced broadly, through the ordinary operation of a large and diverse trader base, or is instead concentrated among a comparatively small number of well-informed or well-resourced traders (as Tetlock would term them, superforecasters), is not a question this paper's data can settle on their own, since we observe only prices and outcomes in this study. It is, however, a well-posed empirical question, with direct implications for how well the calibration should hold as Kalshi's trader base continues to grow and diversify.

2. **Markets are well calibrated, but through what mechanism?** This paper's results are considerably more informative about *when* and *where* Kalshi is well calibrated than about *why*. Whether one truncates the price distribution to exclude near-certain markets, or anchors the resolution clock to an event's true occurrence rather than to a settlement timestamp, changes the measured calibration curve in ways that are themselves informative about the underlying market microstructure, even though, as shown in Sections 4.2 and 4.4, the resulting calibration remains good in absolute terms under every construction examined. A complete account of why Kalshi calibrates as well as it does would need to apportion credit among genuine information aggregation, the specific mechanics of the maker/taker order book, and the incentives facing informed traders operating within it—a decomposition the data assembled here are well suited to support but do not, on their own, resolve.

We regard these two questions less as gaps in the account offered here than as its natural continuation, a productive, well-defined agenda for future research building on the dataset and measurement framework developed in this paper.

7 Limitations

Before concluding, we note several features of this study's design that may bound the conclusions we can draw from it.

First, the relationships documented between calibration and both trading volume and trader count are correlational rather than causal. As we note above, higher-profile events plausibly attract both more participants and better information independently of one another, so the extent to which participation causally improves calibration, rather than simply tracking events that were always going to be easier to price, is not something this dataset can distinguish on its own.

Second, some measures are taken to a reasonable first degree approximation, but may not in and of themselves be objectively superior. For example, the 2–98 cent tail-truncation threshold is a defensible boundary for genuine uncertainty, but a discretionary one. Similarly, the date-truncation exercise that re-anchors the resolution clock to the true occurrence time of the underlying event was done only for some categories viewed to likely be most affected and for which matching time data was available. Whether the same exercise would move the calibration picture for other categories with a settlement lag remains an open question.

Third, the accuracy measures in Section 5 are also worth qualifying on their own terms. Naive and lead-market accuracy are binary right-or-wrong criteria that do not account for how confident a given price was. A contract that traded at 51 cents and resolved Yes is scored identically to one that traded at 95 cents and resolved Yes, even though the latter reflects a considerably stronger claim. As a result, the accuracy measures in Section 5 are not quite answering the same question as the calibration measures in Section 4: a market could show improving accuracy over time simply because more markets drift toward the extremes, independent of whether that added confidence was itself well-placed. A conviction-weighted accuracy measure, one that credits a market in proportion to how far its price moved beyond 50 cents, would complement the present analysis. Future work from Kalshi Research shall propose a novel accuracy measure that incorporates conviction weighting into results.

8 Conclusions

This study set out to evaluate directly the central claim of prediction markets' proponents, that a price on Kalshi can be interpreted literally as a probability. Using the full history of resolved Kalshi markets, we find that this claim holds up well, and increasingly well as resolution approaches, across every category, time horizon, and level of participation examined. That average result is not uniform. It combines large and systematic differences across categories, trading volume, and trader participation, together with the more subtle finding that seemingly innocuous choices about how time itself is measured can shift the calibration picture in economically meaningful ways. Sports and Election markets each illustrate a version of this last point. Apparent miscalibration in the raw data resolves in opposite directions once the resolution clock runs from the true occurrence of the underlying event rather than from Kalshi's administrative market close timestamp.

To our knowledge, this is the first broad-scale study of calibration conducted on the complete resolved history of a major prediction-market exchange, and the first to enrich that analysis with a joint disaggregation by trading volume and trader count, in addition to category and time-to-resolution. The scale of the exercise, more than 2.2 million resolved markets spanning the platform's full operating history, is what makes it possible to document patterns, such as the specific volume and trader-count thresholds identified in Sections 4.6 through 4.8, that would not be visible in a smaller or more narrowly scoped sample, and that we hope will serve as a useful empirical benchmark as both Kalshi and the broader prediction-market industry continue to grow.

None of this is in tension with the basic case for prediction markets as forecasting instruments; if anything, the scale of the calibration achieved close to resolution—Brier scores below 0.02 at Close across more than two million markets—offers a substantial confirmation of the theoretical

argument, developed in Section 2, that a strictly proper scoring rule, which a well-functioning market approximates, renders honest reporting the dominant strategy for participants with money at stake. At the same time, the results show that "prediction markets are well calibrated" is not a fixed, timeless property of a platform, but a property that is earned, market by market, through sufficient time, sufficient participation, and sufficiently unconflicted information among the traders setting the price. Understanding precisely how participation, trader skill, and incentives jointly translate into calibration, the subject of the two open questions raised in Section 6, is the natural next step for this research agenda.

References

Bassamboo, A., Cui, R., & Moreno, A. (2015). Wisdom of Crowds in Operations: Forecasting Using Prediction Markets. *Working paper*.

Berg, J., Nelson, F., & Rietz, T. (2007). Prediction Market Accuracy in the Long Run. Working paper, University of Iowa. Published (2008) in *International Journal of Forecasting*, 24(2), 285–300.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1), 1–3.

Bürgi, C., Deng, W., & Whelan, K. (2026). Makers and Takers: The Economics of the Kalshi Prediction Market. University College Dublin working paper.

Decomposing Crowd Wisdom: Domain-Specific Calibration Dynamics in Prediction Markets (2026). arXiv:2602.19520.

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

Dawid, A. P. (1982). The Well-Calibrated Bayesian. *Journal of the American Statistical Association*, 77(379), 605–610.

DeGroot, M. H., & Fienberg, S. E. (1983). The Comparison and Evaluation of Forecasters. *The Statistician*, 32(1–2), 12–22.

Diercks, A. M., Katz, J. D., & Wright, J. H. (2026). Kalshi and the Rise of Macro Markets. Working paper.

Gjerstad, S. (2004). Risk Aversion, Beliefs, and Prediction Market Equilibrium. University of Arizona Working Paper 04-17.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic Forecasts, Calibration and Sharpness. *Journal of the Royal Statistical Society: Series B*, 69(2), 243–268.

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

Gu, A., Kagan, N., Sun, A., Wu, J., & Xu, H. (2026). When Do Prophets Profit in Prediction Markets? arXiv:2607.06166.

Hayek, F. A. (1945). The Use of Knowledge in Society. *American Economic Review*, 35(4), 519–530.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of Probabilities: The State of the Art to 1980 (Technical Report ADA101986).

Manski, C. F. (2006). Interpreting the Predictions of Prediction Markets. *Economics Letters*, 91(3), 425–429.

Moshrefi, N. (2026). Prices, Probabilities, and Parlays: Systematic Bias in Sports Prediction Markets. arXiv:2607.14430.

Murphy, A. H. (1973). A New Vector Partition of the Probability Score. *Journal of Applied Meteorology*, 12(4), 595–600.

Murphy, A. H., & Winkler, R. L. (1987). A General Framework for Forecast Verification. *Monthly Weather Review*, 115(7), 1330–1338.

Ottaviani, M., & Sørensen, P. N. (2008). The Favorite–Longshot Bias: An Overview of the Main Explanations. In *Handbook of Sports and Lottery Markets*, Elsevier.

Page, L. (2012). "It Ain't Over Till It's Over": Yogi Berra Bias on Prediction Markets. *Applied Economics*, 44(1), 81–92.

Page, L., & Clemen, R. T. (2013). Do Prediction Markets Produce Well-Calibrated Probability Forecasts? *The Economic Journal*, 123(568), 491–513.

Sanders, F. (1963). On Subjective Probability Forecasting. *Journal of Applied Meteorology*, 2(2), 191–201.

Snowberg, E., & Wolfers, J. (2010). Explaining the Favorite–Longshot Bias: Is It Risk-Love or Misperceptions? *Journal of Political Economy*, 118(4), 723–746.

Thaler, R. H., & Ziemba, W. T. (1988). Anomalies: Parimutuel Betting Markets: Racetracks and Lotteries. *Journal of Economic Perspectives*, 2(2), 161–174.

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

Wolfers, J., & Zitzewitz, E. (2006). Interpreting Prediction Market Prices as Probabilities. NBER Working Paper No. 12200.

Yates, J. F. (1982). External Correspondence: Decompositions of the Mean Probability Score. *Organizational Behavior and Human Performance*, 30(1), 132–156.

9 Technical Appendix

The Mathematical Motivation for Meaningful Forecasts

Strictly proper scoring rules make honesty the dominant strategy

A scoring rule $S(p, Y)$ is *strictly proper* if a forecaster who genuinely believes $\mathbb{P}(Y = 1) = q$ minimizes their expected score by reporting $p = q$:

$$q = \arg\min_{p} \; \mathbb{E}_{Y \sim \text{Bernoulli}(q)} \big[S(p, Y) \big].$$

The Brier score is strictly proper:

$$\mathbb{E}[(p - Y)^2] = (p - q)^2 + q(1 - q),$$

which is uniquely minimized at $p = q$ (differentiate with respect to p and set to zero: $2(p - q) = 0 \Rightarrow p = q$). The log score

$$S(p, Y) = -\big[Y \log p + (1 - Y) \log(1 - p) \big]$$

is also strictly proper, and is the unique (up to affine transformation) proper scoring rule with the local/information-theoretic properties tied to Shannon entropy.

This is the formal reason prediction markets are theoretically well-motivated: a market price is a payoff-linked forecast, and if traders are risk-neutral expected-value maximizers, the price that clears the market is exactly the price a strictly-proper-scoring-rule-style incentive would elicit — traders profit by moving the price toward their true belief and are penalized for misreporting. Calibration testing on market prices is therefore a direct empirical test of whether the market is functioning as an efficient information-aggregation device, in the Hayek / Grossman–Stiglitz sense, or whether structural frictions (risk aversion, the favorite-longshot bias, fee structure) are driving a wedge between price and true probability.

10 Brier Scores by Volume and Category

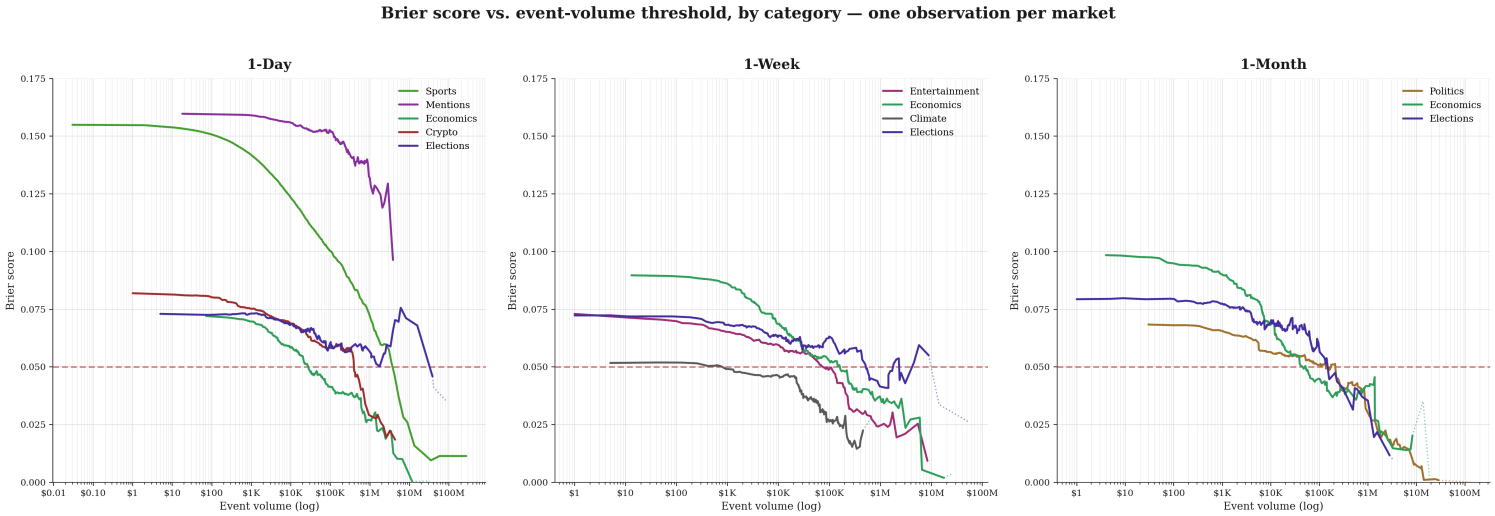

Figure 22: Brier scores by volume on selected categories and horizons.

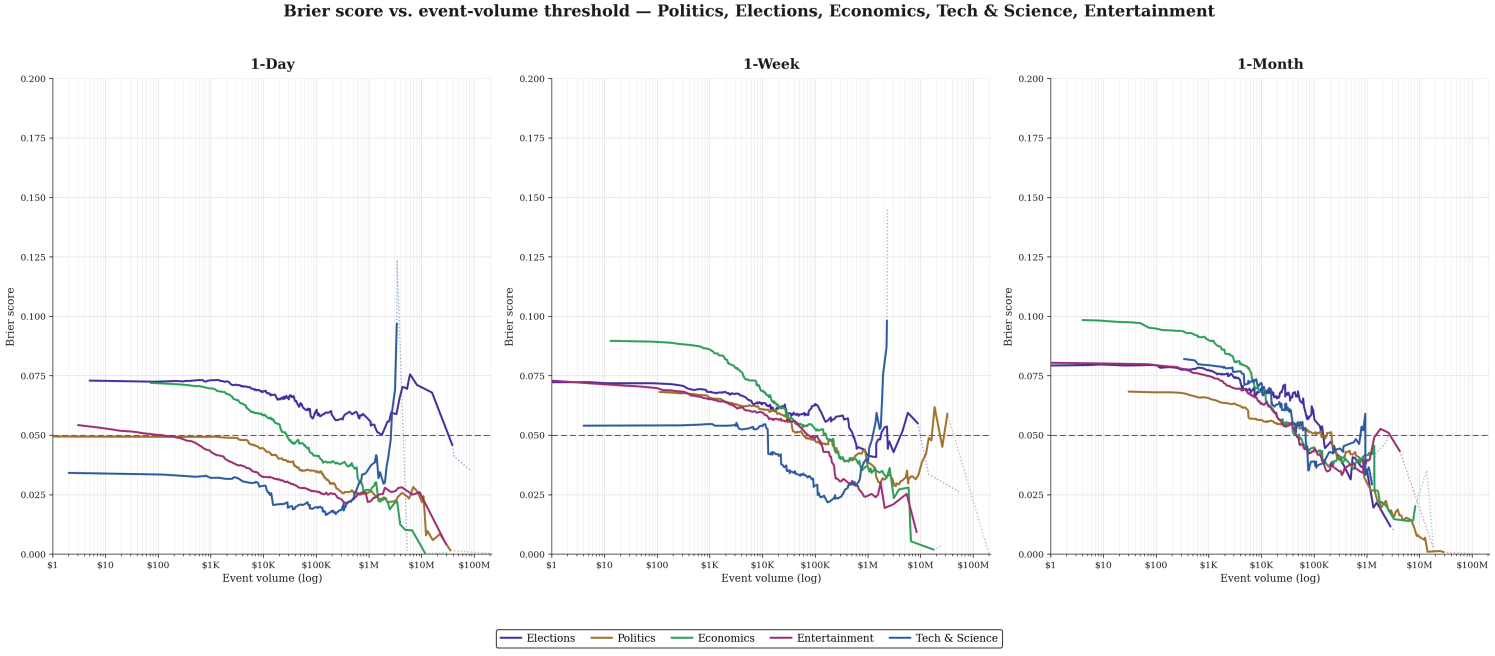

Figure 23: Brier scores by volume on all categories and horizons.

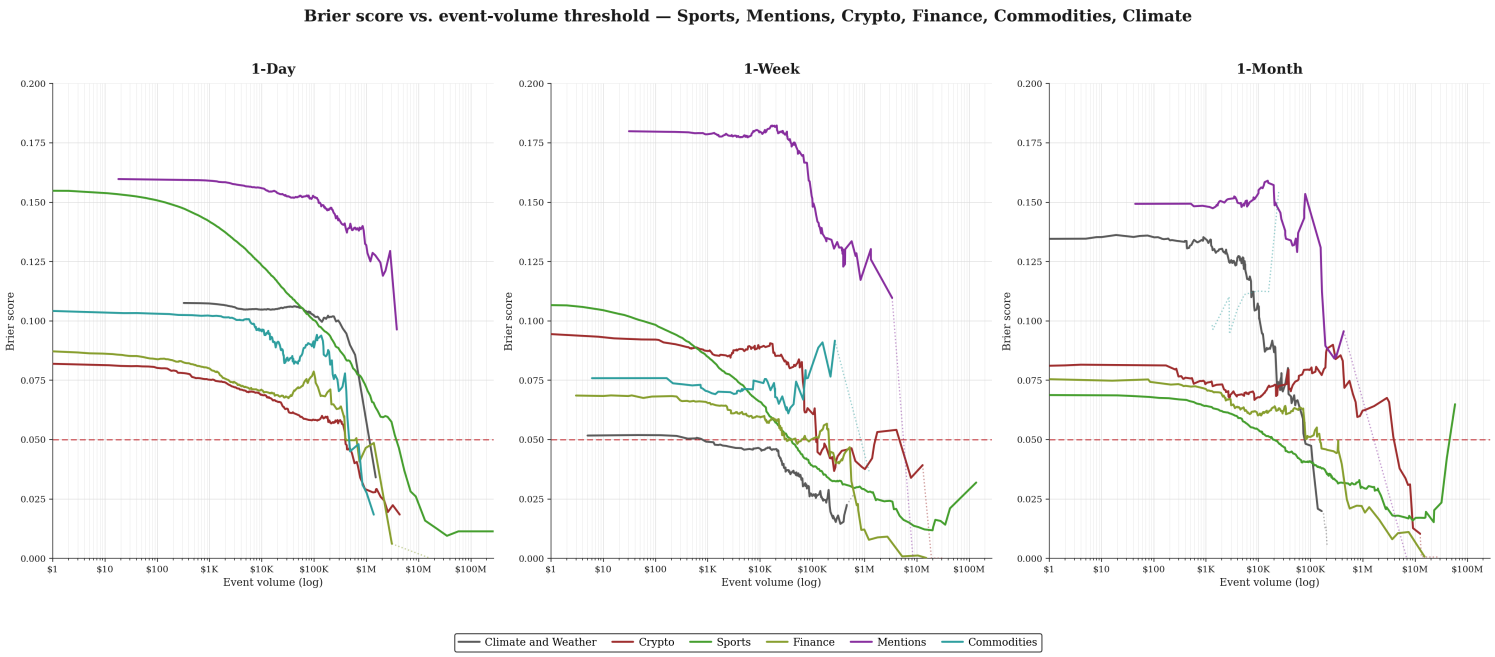

Figure 24: Brier scores by volume on all categories and horizons.

11 Lead-Market Accuracy - Category Extension

Table 5 and Figure 25 show the corresponding lead-market accuracy rising from approximately 55-66% at long horizons to 68 - 94% for Elections and Politics by market close, but remaining comparatively low for categories such as Crypto (11.4% at one week) and Finance (25.5% at one day), where, unlike a two-candidate election, there is frequently no single heavily favored outcome well before resolution.

Category	Events	3-Mo	2-Mo	1-Mo	1-Wk	1-Day	12-Hr	4-Hr	1-Hr	Close
Climate	12,936	—	—	—	—	48.6	64.3	91.7	94.1	94.3
Commodities	606	—	—	—	—	26.6	26.9	38.8	65.2	72.4
Crypto	40,383	—	—	—	11.4	20.8	25.1	44.3	85.4	85.4
Economics	166	—	—	72.3	75.3	79.4	80.0	78.9	77.7	80.1
Elections	934	66.4	72.3	72.6	72.3	69.9	67.7	69.1	68.6	68.6
Entertainment	5,068	60.2	57.9	54.6	51.4	58.8	65.8	72.9	76.5	78.0
Finance	4,607	—	—	—	37.6	25.5	24.8	42.5	77.5	78.4
Politics	785	55.6	80.7	77.6	77.3	69.6	79.0	86.1	93.1	93.8
Sports	125,200	48.9	49.5	53.3	56.9	60.9	61.8	64.0	86.8	98.9
Tech & Science	118	—	—	—	90.6	95.8	96.6	97.5	97.5	97.5

Table 5: Lead-market accuracy (%) by category and time horizon before resolution, for cells with $n \geq 50$ events. Events is the total number of mutually-exclusive events per category.

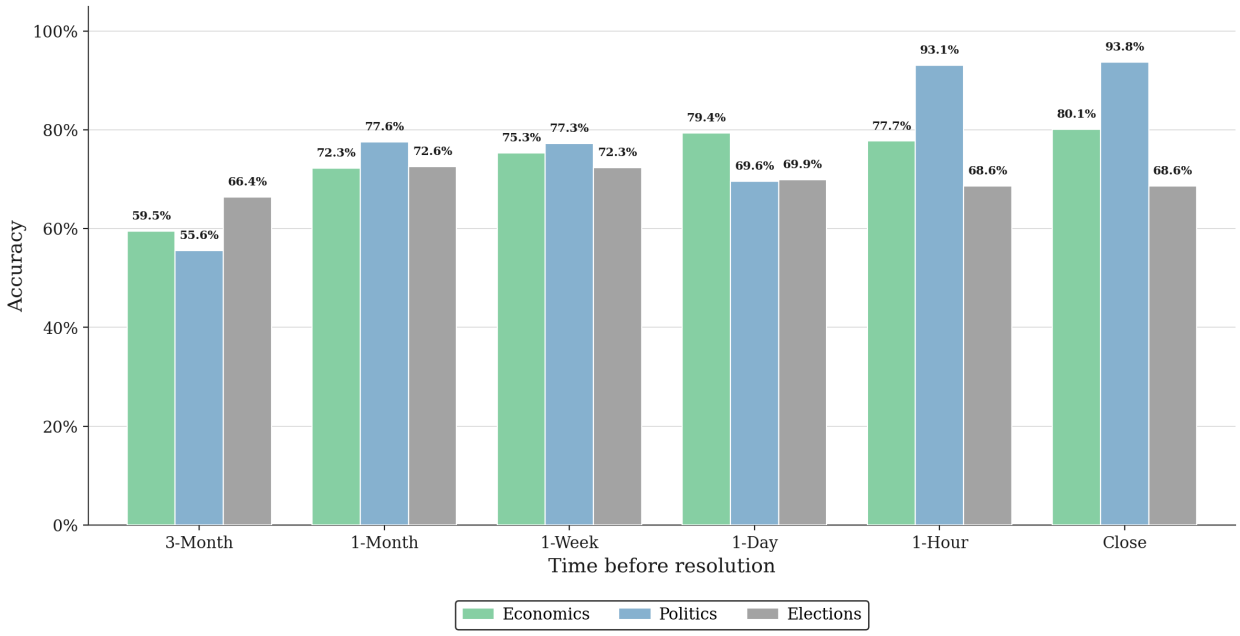

Figure 25: Lead accuracy scores by category.