Kalshi Research

Kalshi and the 2026 Midterms: Calibration and Accuracy

Nicole Kagan and Rubens Baiocchi

Working Paper

November 2026

Abstract

A study released earlier this year assessed over 2.2 million Kalshi markets over the history of the Exchange and found them to be both well-calibrated and accurate (Kagan and Baiocchi, 2026). In this study, we apply a similar analytical lens to markets in the 2026 U.S. Midterm elections. Using data from Kalshi's Exchange and that sourced from alternative election model projections, we find that for races called by 6:26 a.m. ET on November 4, 2026, directional market accuracy improved from 93.8% (3 months out) to 98.2% (polls close), and Brier scores fell from 0.040 to 0.016. Further, Kalshi matched or exceeded polling averages at every horizon where they existed, and on the 9 a.m. Election Day snapshot recorded the lowest Brier score of all forecasts compared.

JEL classification: C53, D72, D84, G14
Keywords: prediction markets, elections, calibration, Brier score, Kalshi

1 Introduction

Prediction markets use the views of many traders to create a single, continuously updating forecast of a future event. In the best case, prices bounded between $0 and $1 admit a literal probabilistic interpretation because of participants' incentive to be correct (Wolfers and Zitzewitz, 2004, 2006; Snowberg, Wolfers, and Zitzewitz, 2013). The record for election forecasting goes back more than a century. From 1884 to 1940 the mid-October favorite in the New York election markets won 11 of 15 presidential elections, and the underdog won once (Rhode and Strumpf, 2004). In the five presidential elections from 1988 to 2004, the Iowa Electronic Markets came closer to the final vote share than the contemporaneous poll in 74% of 964 comparisons (Berg, Nelson, and Rietz, 2008). Research on markets with little real-money volume finds mixed results relative to polls, particularly once poll forecasts are debiased (Erikson and Wlezien, 2008; Rothschild, 2009; Page and Clemen, 2013).

That said, little evidence has been produced to date prior to this year on contemporary prediction markets. Crane and Strumpf (2026) find that in its first month of listing election contracts in 2024, Kalshi's prices were more accurate than every poll-based forecast in the national presidential race, the seven battleground states, and nineteen Senate races. For control of Congress in 2026, Gruca and Rietz (2026) compare Kalshi with the Iowa Electronic Markets and find that Kalshi's prices show day-to-day momentum that the Iowa prices do not.

Given this gap in the literature, this paper scores Kalshi's race-level win probabilities for the Senate, House, and governor races in the 2026 Midterm elections. We compare them with polling averages and with forecast models such as those of The Economist, Silver Bulletin, Race to the WH, and Decision Desk HQ.

We first report accuracy, drawing a strict directional comparison between Kalshi markets and their counterparts across polling and forecast models. We then present calibration results from the available sample at the time of publication.

This paper is structured as follows. Section 2 presents the methodology and data, Section 3 the results, and Section 4 concludes. This is a preliminary working paper and will be updated as further races are called.

2 Methodology and Data

2.1 Kalshi and Event Contracts

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 on a longer tradition of political prediction markets 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. Contracts are 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. In an election market, the contract on a candidate pays $1 if that candidate wins the race, so its price in cents is read directly as the market's win probability for that candidate, as in Section 1. The

companion calibration study (Kagan and Baiocchi, 2026) evaluates this reading across the full history of resolved Kalshi markets; here we apply the same framework to the 2026 midterms.

2.2 Calibration

Let $Y \in \{0, 1\}$ be a binary outcome (here, "candidate wins the race"). 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 in which the forecast was p, the outcome occurs with frequency p:

$$\mathbb{P}\left(Y_i = 1 \mid p_i = p\right) = p \quad \text{for all } p \in [0, 1]. \tag{1}$$

In other words, a candidate priced at 70 cents should win about 70% of such races. Probability is only testable through repetition: by the Law of Large Numbers, pooling many races priced at p and comparing the realized win rate to p is the only way to give a probability forecast empirical content.

Because p is continuous, in practice calibration is assessed over bins $B_k = [p_k, p_k + \delta)$, comparing the observed win frequency in each bin with its mean forecast:

$$\frac{1}{|B_k|} \sum_{i \in B_k} Y_i \;\approx\; \bar{p}_k, \qquad \text{where } \bar{p}_k = \frac{1}{|B_k|} \sum_{i \in B_k} p_i. \tag{2}$$

Plotting observed frequency (vertical axis) against forecast probability (horizontal axis) gives the *reliability diagram* in Figure 1; perfect calibration is the 45° line. Calibration error is the distance of the observed points from that line.

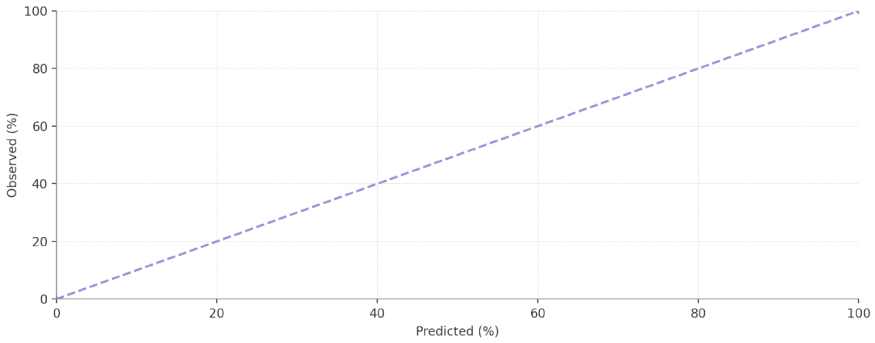


Figure 1: Reliability diagram. Perfect calibration is the 45° line: forecasts of p% come true p% of the time.

2.3 Accuracy

Calibration asks whether probabilities are right. Accuracy asks the coarser question of whether the favorite won. *Directional accuracy* takes the candidate with the highest price in each race as the predicted winner and records whether that candidate won; a tie for the highest price is scored

as incorrect. Over N races, with $\hat{w}_i$ the predicted winner of race i and w_i its actual winner,

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\hat{w}_i = w_i\right]. \tag{3}$$

This is the analogue of asking whether the outcomes priced as more likely than not actually happened. Directional accuracy is easy to interpret and to compare across forecasters, but it does not account for confidence: a race priced at 51 cents that resolves in the favorite's direction is scored identically to one priced at 95 cents. An improvement in accuracy can therefore reflect prices moving toward the extremes without that added confidence being well placed. For this reason we pair accuracy with the calibration and scoring measures of this section and the next.

2.4 The Brier Score and its Murphy Decomposition

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

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - Y_i)^2. \tag{4}$$

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. 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 $p = q$; Appendix A shows this for the Brier score. Alongside the Brier score we report each forecaster's bias, the mean of $p_i - Y_i$ for the Democratic candidate; a negative bias means the forecasts leaned Republican.

Following Murphy (1973), the Brier score splits into three components. Group the N forecasts into bins $k = 1, \ldots, K$ with n_k forecasts, mean forecast $\bar{p}_k$, and observed win rate $\bar{o}_k$, and let $\bar{o}$ be the overall win rate. Then

$$BS \;=\; \underbrace{\frac{1}{N}\sum_{k=1}^{K} n_k(\bar{p}_k - \bar{o}_k)^2}_{\text{Reliability}} \;-\; \underbrace{\frac{1}{N}\sum_{k=1}^{K} n_k(\bar{o}_k - \bar{o})^2}_{\text{Resolution}} \;+\; \underbrace{\bar{o}(1-\bar{o})}_{\text{Uncertainty}}. \tag{5}$$

Reliability is the calibration error: how far forecasts fall from observed win rates, which is the distance from the 45° line in the reliability diagram, weighted by bin size. Resolution measures how well forecasts separate the races a candidate wins from those the candidate loses. Uncertainty depends only on the outcomes, so every forecaster scored on the same races shares it. A lower reliability and a higher resolution both mean a better score, and because the decomposition is exact up to within-bin variation in forecasts, it lets us compare forecasters on calibration and discrimination separately.

2.5 Sample, Horizons, and Comparison Forecasts

Races. The sample consists of Kalshi markets on the Senate, House, and governor races of the 2026 U.S. midterm elections that were called by 6:26 a.m. ET on November 4, 2026. The calibration analysis also includes the local races on Kalshi called by the same time.

Horizons. We score forecasts three months, one month, one week, and one day before the first polls open in each race's state, when those polls open, and when they close. The companion study (Kagan and Baiocchi, 2026) shows that horizons measured from Kalshi's formal market-close timestamp can badly misstate election calibration, because election markets can remain open for days or weeks after the substantive uncertainty is resolved. We therefore anchor all horizons to real-world election events rather than to administrative close times.

Kalshi prices. For each race and horizon we use the last trade of each candidate contract before that time, so Kalshi never benefits from information that arrived later. For calibration, the prices of a race's candidates are scaled to sum to one. Polling averages are taken at the same horizons, at their latest value on or before each one. In the comparison with forecast models, every model snapshot and Kalshi price is taken at 9 a.m. ET on November 3, 2026. There we also report two other Kalshi prices: the midpoint of the bid and ask in the latest quote at or before that time, and the volume-weighted average price (VWAP) of trades in the preceding hour.

Comparison forecasts. We compare Kalshi with polling averages and with publicly available election models, including those of The Economist, Silver Bulletin, Race to the WH, and Decision Desk HQ. The polling averages are VoteHub's own average, and two simple averages that we compute from the polls listed by VoteHub and by Wikipedia: each pollster's latest poll, weighted by $0.5^{a/14}\sqrt{n}$, where a is the poll's age in days and n its sample size, using only polls public by each horizon. Polling averages are scored with the same directional rule as markets: the candidate leading the average is the predicted winner. A tie, whether two candidates share the highest Kalshi price or a polling average shows a margin of exactly zero, is scored as incorrect.

3 Results

3.1 Accuracy

We first turn to a naive interpretation of accuracy. Such a naive interpretation takes the following form: a race is scored as correct if the candidate priced above 50 cents at the relevant horizon is the one who wins, as in Equation (3). Figure 2 reports this directional accuracy as polls close approach. For races called by 6:26 a.m. ET on November 4, 2026, directional accuracy rises from 93.8% three months out to 98.2% at poll close. Given that most races weren't hotly contested (10-90% probability) we turn towards comparisons with polls and electoral forecasting models as proper benchmarks for accuracy.

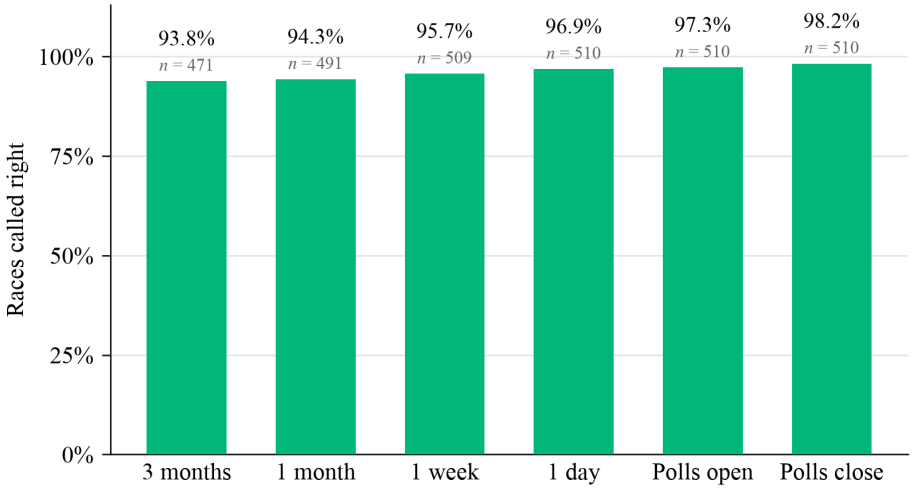


Figure 2: Directional accuracy of Kalshi race-level markets by time horizon, 2026 midterms.

3.2 Accuracy vs Polls

We next compare Kalshi directly with polling averages, scored at each horizon on the races for which a polling average exists. Directional accuracy is the natural common yardstick here, since a polling average gives a leader rather than a calibrated probability. Unlike the data-set we present above, polling averages are not available at poll closure time in aggregate. In line with the evidence that markets can match or beat polls in the long run (Berg, Nelson, and Rietz, 2008) and in Kalshi's early election markets (Crane and Strumpf, 2026), we ask whether Kalshi's advantage persists across the whole horizon rather than only close to the election.

Figure 3 reports the comparison. Note - not all races field polling, creating the difference in directional accuracy from that noted above. The gap comes from a handful of close house races, led by Iowa's 3rd, Virginia's 2nd, Ohio's 9th, Florida's 22nd and Colorado's 8th districts, which Democrats won and for which Kalshi favoured the Democratic candidate over the Republican ahead (or tied) in polling.

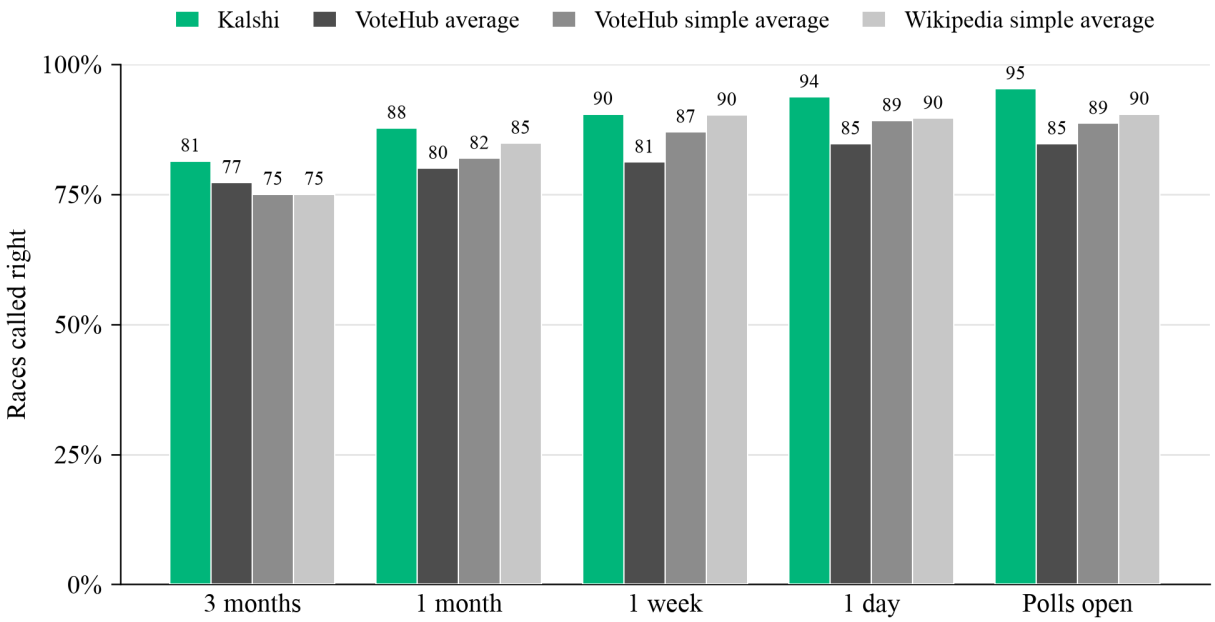
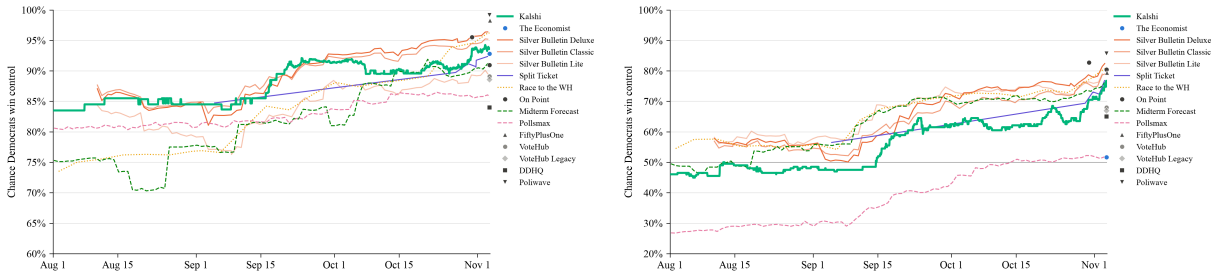


Figure 3: Directional accuracy of Kalshi versus polling averages by time horizon, races with an available polling average. Note: polling not available in all districts.

Per the above, Kalshi markets outperform polling across the assessed timeline. This outperformance occurs both in cases of comparison to a simple average (i.e. with no weighting by pollster credibility) and in the case of a comparison to a polling series weighed by pollster experience and past performance.

3.3 Accuracy vs Forecast Models

We repeat the comparison against the public forecast models listed in Section 2.5. Because models publish probabilities rather than only a leader, we score them on the Brier score. For the race-level scores we take every model snapshot and Kalshi price at 9 a.m. ET on Election Day, and on those races Kalshi's Brier score is below every model's (Table 1). Figure 4 tracks the chance that Democrats win control of each chamber. At 9 a.m. ET on Election Day, models converge at strictly above 50% for the Democrats to win the House and Senate (which they did).



(a) House (b) Senate

Figure 4: Chance that Democrats win control of each chamber, August 1 to 9 a.m. ET on November 3, 2026. Kalshi's hourly price, the last trade before each hour, is in bold green. Solid model lines show history as published, dashed lines recomputed history, and dotted lines sampled history. A model with no published history appears as a marker at each time we collected it.

Table 1 scores Kalshi and every forecasting model on exactly the same races. No matter which way Kalshi's forecast is taken (last price, midpoint or VWAP), Kalshi has the lowest Brier score in both panels: 0.0139 to 0.0140 with governor races, just ahead of Silver Bulletin Deluxe at 0.0141, and 0.0126 to 0.0128 without them, ahead of Split Ticket at 0.0130. Every forecast leaned Republican, Kalshi less than most models.

(a) Models with governor races

Forecaster	Brier	Bias
Kalshi VWAP	**0.0139**	**−0.015**
Kalshi last price	**0.0140**	**−0.014**
Kalshi midpoint	**0.0140**	**−0.020**
Silver Bulletin Deluxe	0.0141	−0.024
Silver Bulletin Classic	0.0163	−0.026
VoteHub	0.0175	−0.032
Model average	*0.0183*	*−0.031*
VoteHub Legacy	0.0185	−0.033
On Point	0.0202	−0.002
Race to the WH	0.0209	−0.036
Silver Bulletin Lite	0.0210	−0.028
DDHQ	0.0234	−0.031
Pollsmax	0.0281	−0.052
N		427

(b) Models without governor races

Forecaster	Brier	Bias
Kalshi VWAP	**0.0126**	**−0.014**
Kalshi last price	**0.0127**	**−0.013**
Kalshi midpoint	**0.0128**	**−0.019**
Split Ticket	0.0130	−0.028
FiftyPlusOne	0.0162	−0.028
Model average	*0.0167*	*−0.029*
The Economist	0.0233	−0.038
Poliwave	0.0263	−0.034
N		393

Table 1: Brier scores and bias of election forecasting models and Kalshi prices on the same races. Lower Brier scores indicate better accuracy. Bias is the mean of the forecast probability minus the outcome for the Democratic side; below zero means too Republican. Races resolved as of 6:26 a.m. ET on November 4 (33 Senate, 360 House, 34 governor). All model snapshots and Kalshi prices were taken at 9 AM on election day.

3.4 Calibration

Accuracy is blind to confidence, so we turn to calibration as defined in Equations (1) and (2). Figure 6 plots the Brier score of Kalshi markets. Future iterations of this study will present the observed win frequency of candidates against their Kalshi-implied probability, with the 45° line as perfect calibration when the quantity of observations allow. The Brier score appears to decline monotonically as a function of time, in slight aberration to what has been found in previous literature, and reflective of the information transmission mechanism present in election markets, where more data is revealed as Election Day approaches.

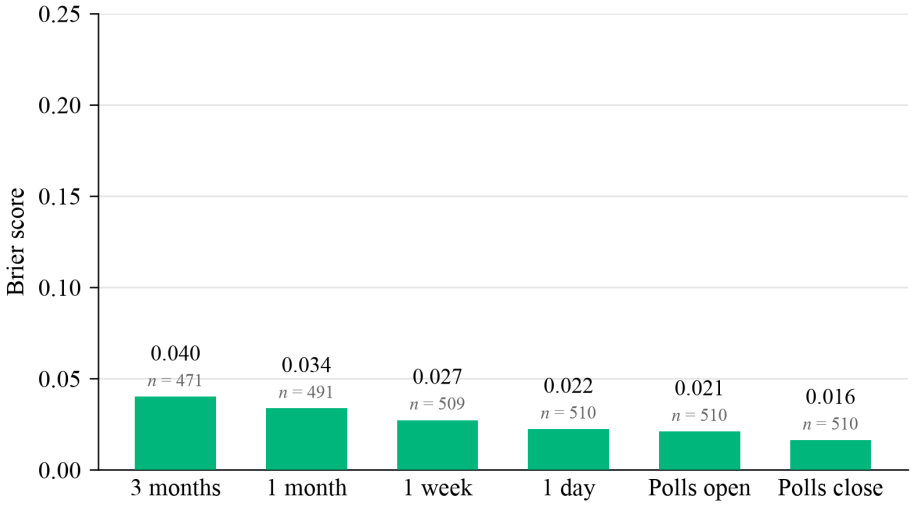


Figure 5: Brier scores of Kalshi race-level win probabilities for the 2026 midterms. Lower is better.

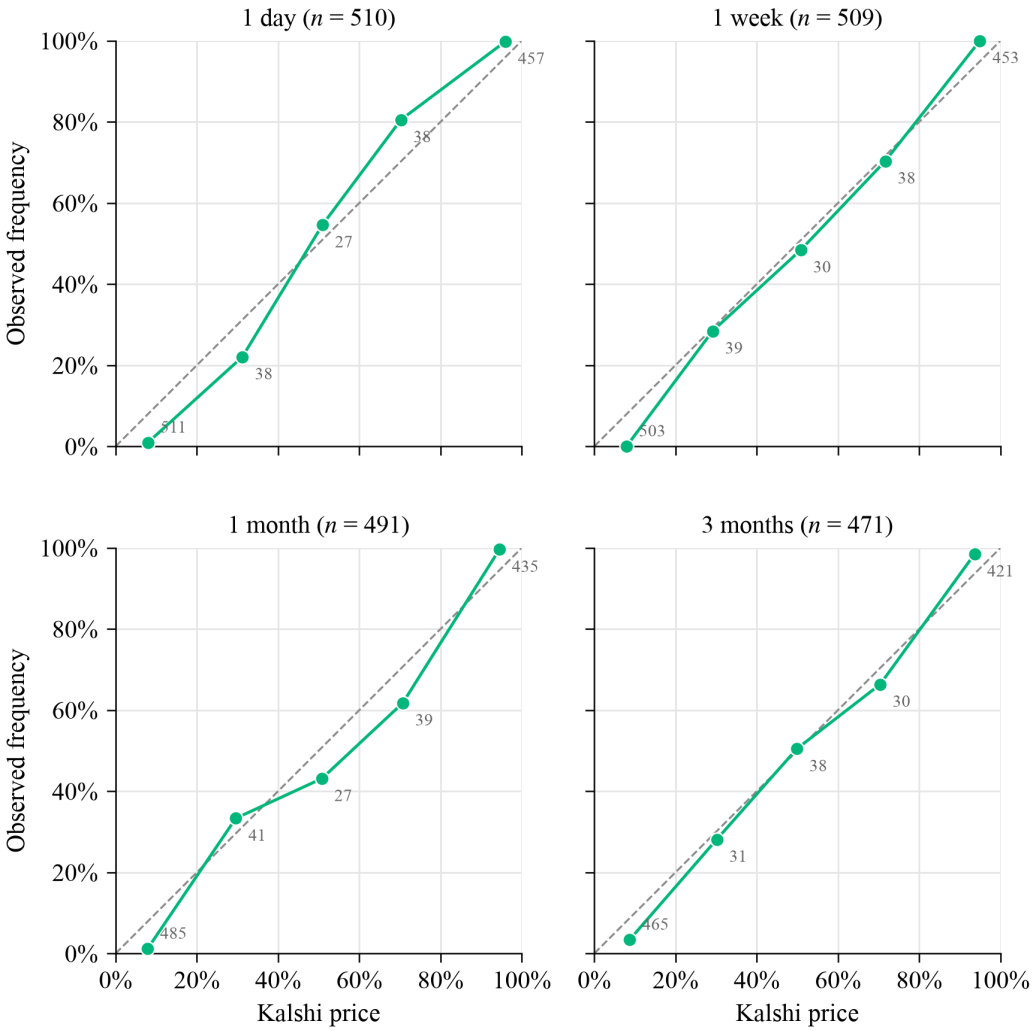


Figure 6: Calibration of Kalshi prices 1 day, 1 week, 1 month and 3 months before the first polls open in each race's state, 2026 midterms. Prices are scaled so that each race's candidates sum to one and grouped into five 20-point buckets; each point is a bucket's mean price against the share of its candidates who won, labelled with the number of candidate contracts. Perfect calibration is the dashed 45° line. n is the number of races priced at that horizon (at 1 day: 33 Senate, 360 House, 34 governor and 83 local races).

4 Conclusion

This study is expected to be augmented and amended as further results become known. That said, a preliminary reading of what unfolded on November 3, 2026, points in a clear direction - prediction markets offer both an accurate and well-calibrated sense of the electoral landscape that is capable of outperforming traditional measures. Naturally, the surface for potential research in this area is expansive, and future additions to the study shall include details of trader counts and market volumes, alongside angles that address the continuously updating nature of the markets and both their speed-to-call relative to news networks and services including the gold-standard Associated Press, and contagion effects shown from information derived from states with earlier poll closure time.

References

Berg, Joyce E., Forrest D. Nelson, and Thomas A. Rietz. 2008. "Prediction Market Accuracy in the Long Run." *International Journal of Forecasting* 24 (2): 285–300.

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

Crane, Harry, and Koleman Strumpf. 2026. "Diversity, Efficiency, and Incentives in Political Prediction Markets." Working Paper, October.

Erikson, Robert S., and Christopher Wlezien. 2008. "Are Political Markets Really Superior to Polls as Election Predictors?" *Public Opinion Quarterly* 72 (2): 190–215.

Gneiting, Tilmann, and Adrian E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477): 359–378.

Gruca, Thomas S., and Thomas A. Rietz. 2026. "Forecasting 2026 Congressional Control: A Comparison of IEM and Kalshi Prediction Markets." *PS: Political Science & Politics*, forthcoming.

Kagan, Nicole, and Rubens Baiocchi. 2026. "Calibration in Prediction Markets: Theory and Evidence." Kalshi Research Working Paper, August.

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

Page, Lionel, and Robert T. Clemen. 2013. "Do Prediction Markets Produce Well-Calibrated Probability Forecasts?" *Economic Journal* 123 (568): 491–513.

Rhode, Paul W., and Koleman S. Strumpf. 2004. "Historical Presidential Betting Markets." *Journal of Economic Perspectives* 18 (2): 127–142.

Rothschild, David. 2009. "Forecasting Elections: Comparing Prediction Markets, Polls, and Their Biases." *Public Opinion Quarterly* 73 (5): 895–916.

Snowberg, Erik, Justin Wolfers, and Eric Zitzewitz. 2013. "Prediction Markets for Economic Forecasting." In *Handbook of Economic Forecasting*, vol. 2A, edited by Graham Elliott and Allan Timmermann, 657–687. Amsterdam: Elsevier.

Wolfers, Justin, and Eric Zitzewitz. 2004. "Prediction Markets." *Journal of Economic Perspectives* 18 (2): 107–126.

Wolfers, Justin, and Eric Zitzewitz. 2006. "Interpreting Prediction Market Prices as Probabilities." NBER Working Paper 12200.

A Appendix

Strictly Proper Scoring Rules and the Brier Score

A scoring rule $S(p, Y)$ is strictly proper if a forecaster who genuinely believes $\mathbb{P}(Y = 1) = q$ minimizes expected score by reporting $p = q$ (Gneiting and Raftery, 2007):

$$q = \arg\min_{p} \; \mathbb{E}_{Y \sim \text{Bernoulli}(q)} \big[S(p, Y) \big]. \tag{6}$$

The Brier score is strictly proper:

$$\mathbb{E}\big[(p - Y)^2\big] = (p - q)^2 + q(1 - q), \tag{7}$$

which is uniquely minimized at $p = q$ (differentiate with respect to p and set to zero: $2(p - q) = 0 \Rightarrow p = q$). The same holds for the logarithmic score $S(p, Y) = -\big[Y \log p + (1 - Y) \log(1 - p)\big]$.

This is the formal reason prediction markets are theoretically well motivated as forecasting instruments. A market price is a payoff-linked forecast: if traders are risk-neutral expected-value maximizers, they 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, or whether frictions such as risk aversion, the favorite–longshot bias, or fee structure drive a wedge between price and true probability.