Kalshi Research

Polls, Models, and Markets in the 2026 Midterm Election

Nicole Kagan and Rubens Baiocchi

Working Paper

November 2026

Abstract

The 2026 U.S. Midterm elections represented the first full electoral cycle covered by prediction markets. Consequently, much interest has emerged in a comparison of performance to traditional instruments such as polls and election models.We evaluate Kalshi, a regulated real-money prediction market, against 13 publicly available election forecasting models across the 478 Senate, House, and gubernatorial races called by November 14. Kalshi's price at 9 a.m. on Election Day yields a Brier score of 0.0317, lower than that of every model (best model: 0.0364; model average: 0.0410). Kalshi also exhibits the highest resolution, the second-lowest bias, and the third-best calibration among the 13 models, and these results are robust to weighting each office equally. Kalshi's implied vote margins are likewise more accurate than polling averages: its median margin differs from the AP vote count by 4.6 points on average, compared with 6.2 points for both a simple average of polls and an independently computed average. Its advantage is concentrated in competitive races, House races, and redistricted seats, where past results and earlier polls map poorly onto new district boundaries. The favourite in Kalshi's market winning 95.8% of resolved races at the first poll closing, up from 91.8% three months earlier, and prices reached 99¢ before the AP call in a majority of races. Over the final eleven days of the campaign, Kalshi's forecast error was the lowest of any forecaster, whereas the leading models performed best in late September and early October. This timing is consistent with market prices incorporating early-voting information that models did not, although we do not directly test this. We interpret the results as evidence that markets and models are complementary rather than competing: polls and models supply the underlying evidence, and markets aggregate and weigh it, analogous to the relationship between financial markets and the fund managers who rely on them.

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

We thank Zachary Donnini and Luke Wines from VoteHub for discussion and access to its poll database.

1 Introduction

Markets on election outcomes are hardly new. The first 'prediction markets' on elections can be traced back to papal elections in the sixteenth century, and were a prominent "Wall Street" feature in the late nineteenth and early twentieth centuries. Of late, considerable academic interest has developed around comparisons between market-based forecasts and those produced by polls and election modelers.

Compared to individually designed polls and election models, prediction markets aggregate the views of many traders into a single, continuously updating forecast of a future event. In the best case, prices bounded between \$0 and \$1 admit a literal probabilistic interpretation, owing to participants' incentive to be correct (Wolfers and Zitzewitz, 2004, 2006; Snowberg, Wolfers, and Zitzewitz, 2013). This predictive potential is among their main promises (Arrow et al., 2008), and the record of their election forecasting performance stretches back more than a century. From 1884 to 1940, election market favourites 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 polls in 74% of 964 comparisons (Berg, Nelson, and Rietz, 2008). Other 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).

Yet large-scale, real-money prediction markets, which were authorized to list election events only in late 2024, remain largely unexamined. A recent study by Crane and Strumpf (2026) finds that, despite traders being unrepresentative of the voting population, Kalshi's 2024 election contracts were priced more accurately than every poll-based forecast in the national presidential race, the seven battleground states, and nineteen Senate races. Similarly, Zitzewitz (2026) find little political bias across a wide range of prediction markets.

The 2026 midterms allow us to extend this evidence considerably. Over \$2.5 billion was traded on Kalshi's midterm markets, which spanned not only the marquee contests but also a large number of low-level races, yielding a rich sample for analysis. Using this sample, we find that Kalshi exhibited less bias than all but one of the election forecasting models, as well as the lowest forecasting error of any. Moreover, against polling averages, Kalshi's vote margin forecasts were closer to the AP count and its average error was about 1.5 points lower. Despite frequent claims of bias or manipulation, these results are inconsistent with such hypotheses. More plausible explanations are momentum effects and the favourite–longshot bias, though these lie largely beyond the scope of this paper (Gruca and Rietz, 2026; Zitzewitz, 2026).

This paper is structured as follows. Section 2 describes the data, Section 3 the scoring methods, and Section 4 compares Kalshi with the models and the polls, first on election morning and then over the campaign. Section 5 discusses the results and Section 6 concludes. Appendix A estimates how the Senate outcome moved S&P 500 futures on election night; the other appendices give robustness checks.

2 Forecasting 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 priced from 0.1 to 99.9 cents, in tenth-of-a-cent steps on many markets, with trades matched between a "Maker," who posts the first offer, and a "Taker," who accepts it.

Election markets have been tradable on Kalshi since late 2024. Broken down by category in Figure 1, the 2024 presidential election saw over 1 billion contracts traded. Since then, election markets have seen a widening diversification into a variety of subgroups covering international elections, 2026 midterm primaries and general elections, as well as other local US elections. While in 2024 Kalshi listed about 200 election events (e.g., "Who will win Minnesota in the presidential election?"), in 2026 there were more than 2,000 active events in November for the Midterm elections. In addition to the actual winner for the most important races, these 2026 Midterm markets traded on all 436 House, 35 Senate, 39 Governor, and 177 local races along with their vote margins, vote share, and other classifications of outcomes. This is the first time in prediction market history that event contracts were available on such a wide spectrum for a U.S. election cycle.

The market's footprint grew with volume and traders, which rose sharply in the months before November 3 (Figure 1). Three months before polls opened, the median Senate and governor races of our sample had about 50,000 to 55,000 contracts traded and 185 to 200 traders, against about 6,500 contracts and 30 traders for the median House race. By the time polls opened on election day, those figures had reached 800,000 to 845,000 contracts and about 1,200 traders for Senate and governor races, and about 22,000 contracts and 90 traders for House races. What these median numbers hide is that volume and the number of traders were significantly concentrated in contested races trading between 10 and 90 cents. Across all these individual midterm races, more than 1 billion contracts had traded by the time polls closed, among about 570,000 traders.

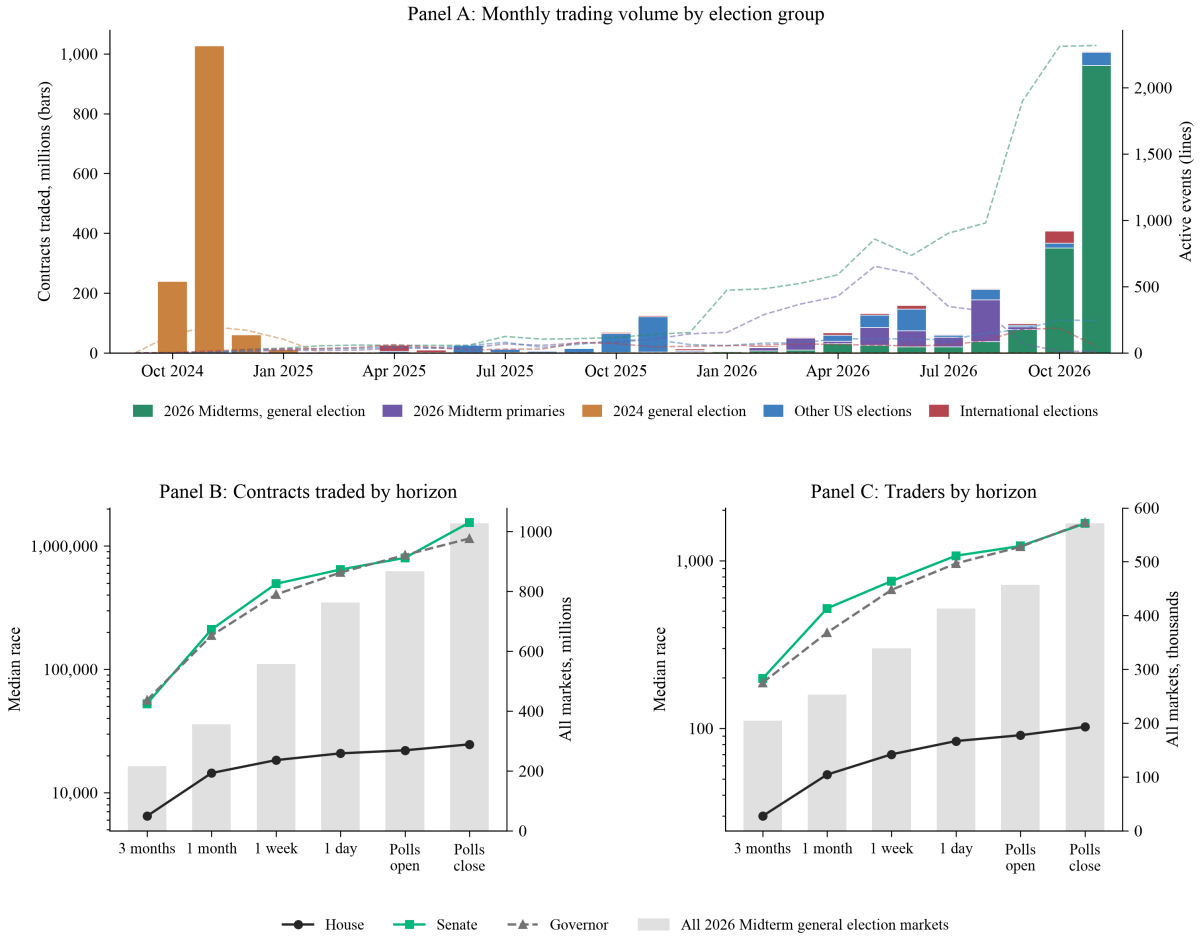


Figure 1: Trading volume. Panel A shows the number of contracts traded each month in each group of Kalshi election markets (bars) and the number of events with at least one trade that month (dashed lines). Panels B and C show the cumulative number of contracts traded and of distinct traders up to each horizon in the 2026 Midterms general elections group. They show the median House, Senate and governor race (lines, log scale) and all 2026 Midterm general election markets (bars). Horizons are measured from the first poll opening in each state.

2.2 Speed of Information Aggregation

What matters on Election Day isn't just the accuracy of directional positioning, but also the speed with which a correct, high-conviction, position is reached and held. And this doesn't just matter for those following the election, but also for public markets. Following Snowberg, Wolfers, and Zitzewitz (2011), the joint movement of Kalshi's Senate market and S&P 500 futures on election night implies that investors valued the index about 3.9% higher under a Republican Senate than under a Democratic one (Appendix A), a shock able to be accounted for only with real-time repositioning in response to continuously-updating signals.

Figure 2 compares when the AP called each race with when Kalshi's market on the eventual AP winner first traded at 99¢ or more and never again traded below 98¢. Kalshi reached 99¢ before the AP call in 22 of 23 Senate races, 26 of 27 governor races and 153 of 235 House races. The median lead, the AP call time minus the time Kalshi reached 99¢, was 58 minutes in the

Senate, 80 minutes for governor and 37 minutes in the House. The longest leads came from races the AP called days later - i.e. the races deemed by experts as the most difficult to call. Kalshi reached 99¢ in the Alaska governor race on November 4, ten days before the AP call; in the Kansas governor race on November 5, a week before; and in the Mississippi Senate race on November 6, five days before. These leads are consistent with the view that a continuously trading market gives traders, through a stated profit incentive, a reason to incorporate new information into prices as soon as it becomes available, so that prices approach certainty before mainstream outlets call the race.

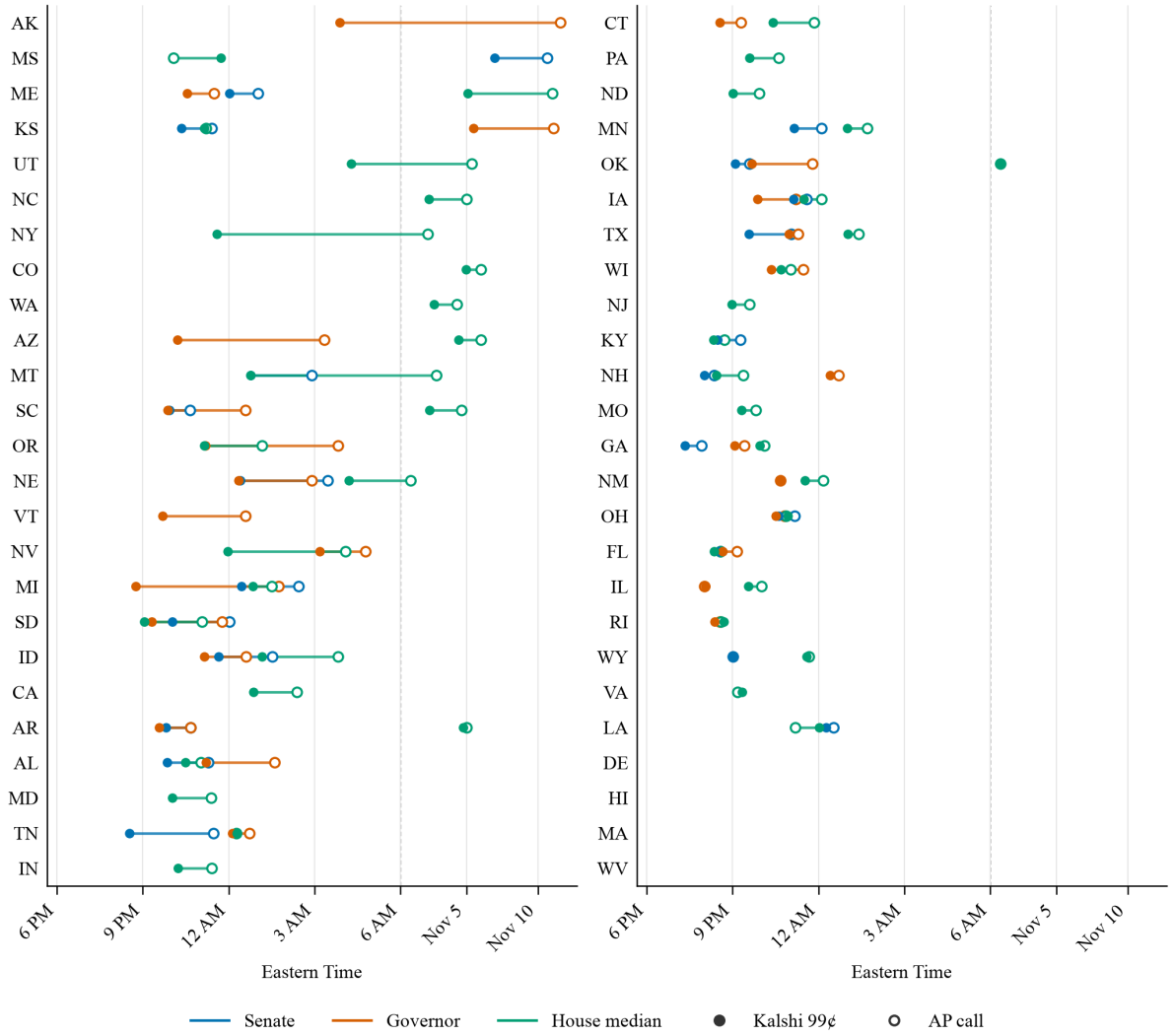


Figure 2: When Kalshi reached 99¢ against when the AP called the race, by state. Each row shows the state's Senate race (blue), governor race (orange) and House races (green) on one line. Filled dots mark when Kalshi's market on the AP winner reached 99¢ for good: its first trade at 99¢ or more after which it never traded below 98¢ again. Hollow dots mark the AP call. The House dumbbell links the median AP call time with the median time Kalshi reached 99¢ over the state's charted House seats. Races already at 99¢ before their state's first poll closing (165) and 48 thinly traded House races that never traded at 99¢ are left out, so Delaware, Hawaii, Massachusetts and West Virginia have no dumbbells. States are ordered by the mean lead of their chambers, AP call minus Kalshi time, widest first. Times are ET, linear from 6 p.m. November 3 to 6 a.m. November 4 (dashed line) and logarithmic after. AP calls as of November 17.

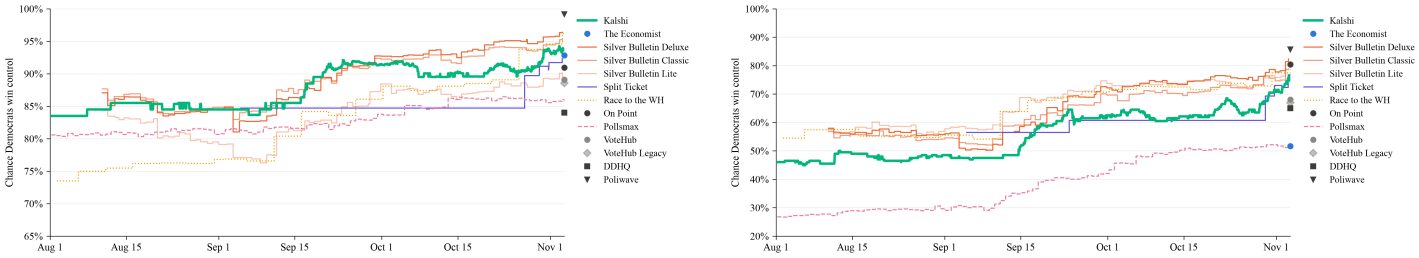
2.3 Election Forecasting Models

Alongside prediction markets, the 2026 midterms produced a wide range of publicly available forecasting models, built by individuals and teams of experts. We study ten of the most widely followed, together with their respective variants, as displayed on each model's website at 9 AM on Election Day - Tuesday, November 3rd. Not all models provide a daily history of their forecasts. The models included in our analysis are: Silver Bulletin, FiftyPlusOne, DDHQ, Race to the

White House, The Economist, VoteHub, OnPoint, Pollsmax, Poliwave, and Split Ticket. Two of these, VoteHub's headline model and DDHQ, incorporate prediction market prices as inputs.

Most of these models share a common methodological framework. For each race, polls are aggregated into a proprietary polling average, with weights reflecting each pollster's historical bias and error as well as its perceived partisan lean. Undecided voters are then allocated between the parties, and the resulting vote share is adjusted for fundamental factors such as fundraising, incumbency, past results, and demographics. For races without polling, models infer outcomes from correlations with comparable races that share similar political forces and for which polling is available. Expert ratings, such as those of Cook's Political Report or Sabato's Crystal Ball, are also incorporated into the estimated vote share. The adjusted vote share is converted into a probability of victory by simulating a large number of outcomes, drawing on estimated correlations across races. The output of each model is thus a probability of victory for each party in every House, Senate, and gubernatorial race, although some models omit gubernatorial forecasts. Update frequency also varies: most models update daily, while others do so less often.

The models then aggregate race-level probabilities into headline forecasts of chamber control and seat counts (and, in presidential years, of the presidency). We compare these forecasts with Kalshi's market-implied probabilities in Figure 3 for House and Senate control, and in Figure 4 for seat counts. Two observations stand out. First, in terms of the confidence attached to House and Senate control, Kalshi occupied a middle position among the models. The Silver Bulletin's Deluxe and Classic models assigned higher probabilities than Kalshi on election day and for most of their history (Figure 3). Split Ticket, by contrast, assigned lower probabilities than Kalshi for most of its history, in roughly 90% of hours for the House and 70% for the Senate. Second, Kalshi's series updated more frequently and extended considerably further back in time, whereas most models began their forecasts only in the summer.



(a) House (b) Senate

Figure 3: 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, each value held until the model's next update. A model with no published history appears as a marker at its 9 a.m. November 3 forecast.

While these headline chamber control outcomes are the most important predictions, they don't offer a wide cross section to judge competing forecasts. Moreover, since 2016, the Presidency, House, and Senate have all been won by the favourite with the exception of 2022, where Democrats unexpectedly held the Senate, and 2024, where the House was a toss-up. In 12 observations (2018–2026), only one was a true upset.

Digging one level deeper at the seat count forecast we find another important distinction. Kalshi's implied distribution for Senate and House seat counts has several important features. Compared to traditional election models that have a monotone distribution from thousands of simulated scenarios, Kalshi's curve is more distributionally rich. In the House, for instance, a week before the election a bump started to develop between 250 and 260 seats independent of the peak at 240. This shape was absent from election forecasting models. It is still unclear what the exact source of this bimodality is, but perceived correlations between scenarios are a likely explanation.

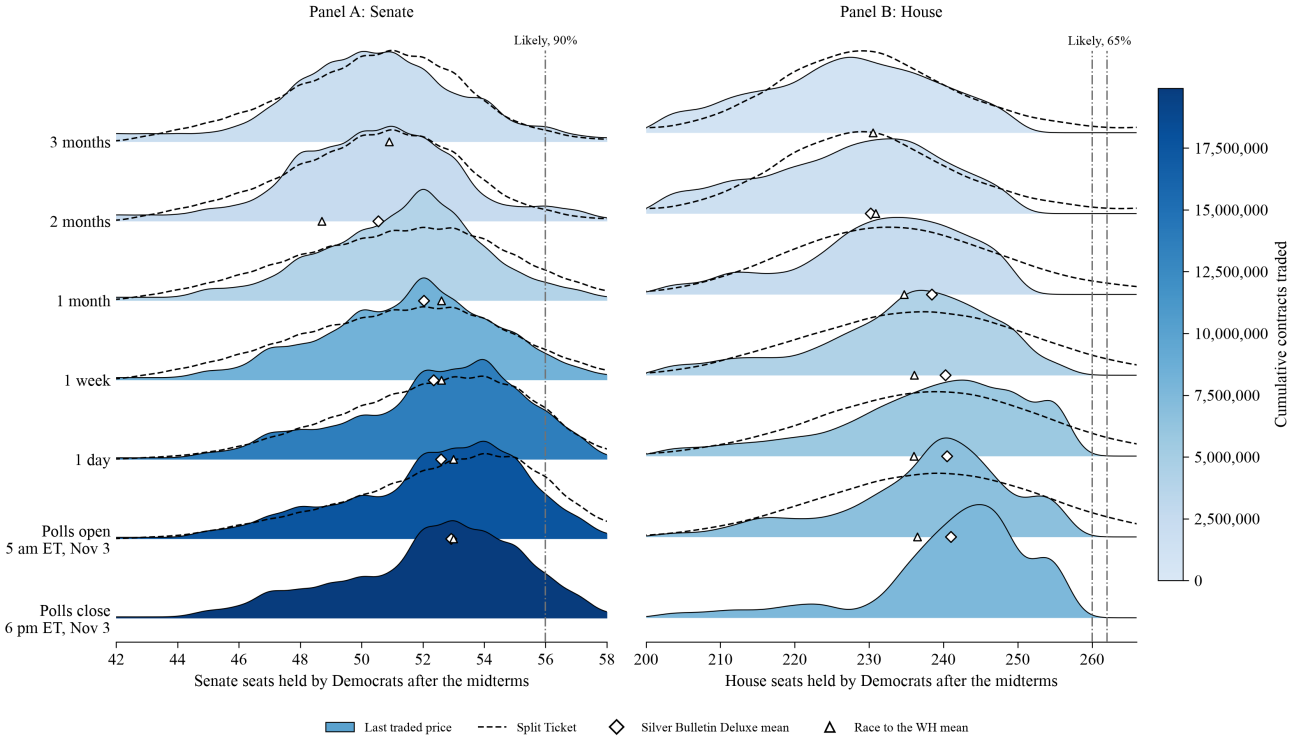


Figure 4: Market-implied distributions of Democratic seats for the House and Senate. Each ridge is the probability distribution over the number of seats held by Democrats after the midterms, implied by Kalshi's seat-count markets at a given time. Horizons are measured from the first poll opening (5 am ET, November 3): 3, 2 and 1 months, 1 week and 1 day before, at the opening, and at the first poll closing (6 pm ET, November 3). The filled ridge uses the last traded price of each market before that time, shaded by the contracts traded on the seat markets by then. For the Senate, two events (one with higher strikes) are combined, and an open-ended market is spread evenly over the 5 seats beyond its strike. For the House, the 4-seat bands are spread evenly across their seats, and the open-ended band above 249 seats is split at the strikes of the two other House seat events listed and traded by then. The dashed line is Split Ticket's seat distribution in its latest version published by then. The dash-dot lines mark the likely outcome, from trades on November 5.

Looking at every individual race in Figure 5, we find that Kalshi and the models had smaller differences when forecasting Senate seats. In the House, large differences emerge. In many House seats Kalshi was far more bullish than the model average. Interestingly, in states that redistricted (Alabama, California, Florida, Louisiana, North Carolina, Ohio, Tennessee, Texas and Utah) Kalshi was far more favorable to Democrats than the models in several key seats. On

the Governor side Nevada jumps out. Kalshi traders placed a much higher probability on the eventual Democratic winner than models. These differences are likely to have been driven by information from early voting which models didn't incorporate.

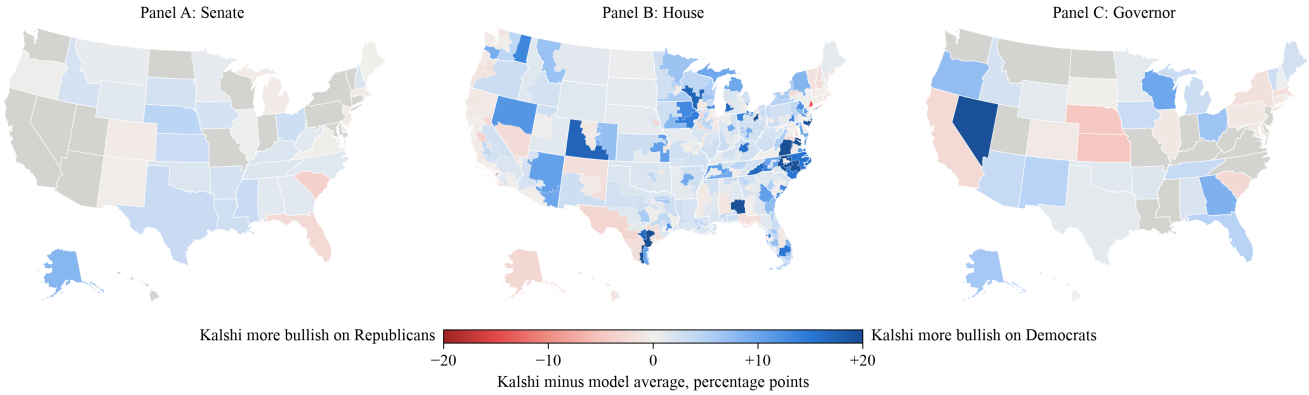


Figure 5: Kalshi minus the model average, by race. Each race is shaded by Kalshi's last traded price for the Democrat at 9:00 am ET on November 3, raw and clipped to 0–1 as in Table 1, minus the average Democratic win probability of the published election models covering that race, in percentage points. Blue: Kalshi was more bullish on the Democrat than the models; red: more bullish on the Republican. Grey: no race.

2.4 Polls and Polling Averages

Polls estimate each candidate's likely share of the vote. Experts aggregate these polls into a polling average, usually adjusting them for pollster bias, assigning undecided voters, and other factors. This aggregation is itself a very important input into election forecasting models as it inevitably determines their accuracy. We compare Kalshi with the independently built polling average from VoteHub (available for 122 races). We also use polls from Wikipedia which are available for a larger sample size of 172 and calculate our own simple average, which weights each pollster's latest poll in the race by its sample size and by its age, halving the weight every 14 days:

$$\bar{m} = \frac{\sum_j w_j m_j}{\sum_j w_j}, \qquad w_j = 0.5^{\,a_j/14}\sqrt{n_j}, \tag{1}$$

where m_j is poll j's Democratic minus Republican margin, n_j its sample size and a_j the days from the end of its fieldwork to November 3. We assume an even split of undecided voters.

Due to limited resources pollsters usually perform their work on the most important and hotly contested races, leaving most races that are thought to be a foregone conclusion unpolled. As of November 14, the underdog won three unpolled races (NC-06, NC-13, NJ-02). In Figure 6 we see how Kalshi compared to polls on election night for each race. State level effects seem to dominate.

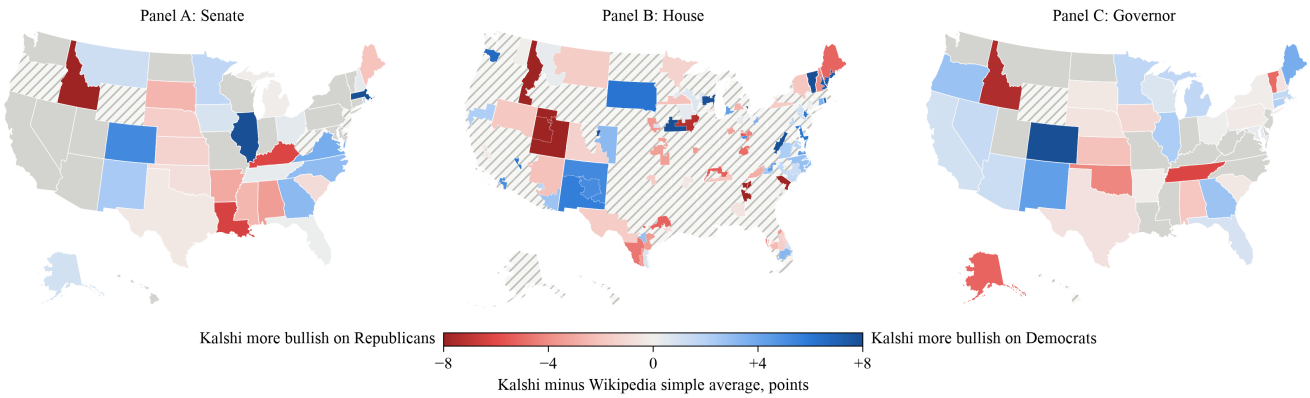


Figure 6: Kalshi minus the Wikipedia simple poll average, by race. Each race is shaded by the median of Kalshi's vote margin markets at 9:00 am ET on November 3, from last trades, minus the simple average of the race's polls on Wikipedia by then, both as the Democratic minus Republican margin, in points. Blue: Kalshi was more bullish on the Democrat than the polls; red: more bullish on the Republican. Hatched: no polls; grey: no race.

3 Methodology

3.1 Comparing Markets and Models

Races. The sample is the Senate, House and governor races of the 2026 U.S. midterm elections that have a Kalshi market and were called by 7:43 p.m. ET on November 14, 2026. We assign a race key to each final party or candidate winner event (whichever has the highest volume) on Kalshi and crosswalk it to model forecasts for the same races. We exclude two kinds of races. The first is the nine races in which both finalists were from the same party, eight California House races between two Democrats and one between two Republicans: the models give a party's chance of winning, not a candidate's, so they make no forecast there. The second is the nine uncontested races, with no Democrat or no Republican on the ballot, where every forecast is close to 0 or 1. The House races of the District of Columbia and the governor races of the territories have no model forecasts and are not in the sample. That leaves 478 races: 34 Senate, 408 House and 36 Governor contests. Where no Democrat is in contention and the models forecast an independent instead (the Nebraska, Idaho, Montana and South Dakota Senate races and Alaska's at-large House seat), the independent takes the Democratic side, for the models and for Kalshi. Elsewhere the Democratic side is the Democrat alone; an independent's chance is left out on both sides. We score markets and models on these races.

Prices. The comparison with forecast models uses one time for every model snapshot and Kalshi price: 9 a.m. ET on November 3, 2026, to avoid look-ahead bias from results. Our headline is the last traded price. We also report two other Kalshi forecasts for comparison: the midpoint of the bid and ask in the latest quote at or before 9 a.m., and the volume-weighted average price (VWAP) of trades in the hour before it, or, for a market with no trade in that hour, in the hour ending at its last trade before 9 a.m. Kalshi's Democratic probability is the price of the market on the candidate the models forecast. Prices are not rescaled within a race.

3.2 The Brier Score and the Murphy Decomposition

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

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

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. 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 E 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

$$BS \;=\; \underbrace{\frac{1}{N}\sum_{k=1}^{K} n_k(\bar{p}_k - \bar{o}_k)^2}_{\text{Reliability}} \;-\; \underbrace{\frac{1}{N}\sum_{k=1}^{K} n_k(\bar{o}_k - \bar{o})^2}_{\text{Resolution}} \;+\; \underbrace{\bar{o}(1-\bar{o})}_{\text{Uncertainty}}. \tag{3}$$

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.

3.3 Comparing Markets and Polls

Polls forecast a vote margin, not a probability, so we score Kalshi against them with the mean absolute error (MAE) and the root mean squared error (RMSE) rather than the Brier score:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|m_i - M_i|, \qquad RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(m_i - M_i)^2}, \tag{4}$$

where m_i is the forecast and M_i the outcome in race i, both the Democratic minus Republican margin in points. Kalshi's forecasts are the displayed median, mean and mode of each race's vote margin markets at 9 a.m. ET on November 3. The outcome is the margin in the Associated Press's unofficial vote count as of November 22, when at least 89% of the expected vote had been counted in every race, and 99% in the median race. In Alaska and Maine, which use ranked-choice voting, it is the first-round count. Kalshi and each poll average are scored on the same races. Certified counts could change the results slightly.

4 Results

Table 1 scores each forecaster's probability that the Democratic candidate wins, taken at 9 a.m. ET on November 3, 2026, on the 478 called races: 34 Senate, 408 House and 36 governor races. We score 13 models, their average and 3 Kalshi prices. The model average is the mean of the models covering a race. Kalshi's three prices have the three lowest Brier scores in the table. The last trade scores 0.0317, and VWAP and the midpoint both score 0.0319. The best model, Silver Bulletin Deluxe, scores 0.0364, and the model average 0.0410.

The Murphy decomposition shows where Kalshi's lead comes from. Reliability is the calibration error, so lower is better. Kalshi's reliability is 0.0096 for the last trade, 0.0097 for VWAP and 0.0098 for the midpoint. Its last price is better calibrated than every forecaster but On Point, at 0.0055, and Silver Bulletin Deluxe, at 0.0092; Silver Bulletin Classic, at 0.0097, ties Kalshi's VWAP. Resolution measures how well forecasts separate the races a candidate wins from those the candidate loses, so higher is better. Kalshi's resolution is the highest in the table, at 0.2184, 0.2183 and 0.2183 for its three prices. The closest model is VoteHub, at 0.2174; the model average reaches 0.2156, Silver Bulletin Deluxe 0.2125 and On Point 0.2022. Kalshi ordered races by outcome better than any model. Both terms are computed over ten equal-width bins of the forecast probability; with other binnings, rankings within about 0.001 of each other can change (Kagan and Baiocchi, 2026). Uncertainty depends only on the outcomes, so it only distinguishes forecasters who had Governor predictions from those that didn't.

Bias is the mean of $p - y$, where p is the forecast probability for the Democrat and y is 1 if the Democrat won and 0 otherwise. Every forecaster's bias is negative, so every forecaster underestimated Democrats. Kalshi's bias is -0.0281 for the last trade, -0.0289 for VWAP and -0.0335 for the midpoint: at the last trade, Kalshi's mean probability for the Democrat was 2.8 percentage points below the share of races Democrats won. Only On Point, at -0.0108, has a smaller bias. The model average's bias is almost double Kalshi's, at -0.0510.

Forecaster	Brier score	Bias	Reliability	Resolution	Uncertainty	N
Kalshi last price	**0.0317**	**−0.0281**	**0.0096**	**0.2184**	**0.2407**	**478**
Kalshi VWAP	**0.0319**	**−0.0289**	**0.0097**	**0.2183**	**0.2407**	**478**
Kalshi midpoint	**0.0319**	**−0.0335**	**0.0098**	**0.2183**	**0.2407**	**478**
Silver Bulletin Deluxe	0.0364	−0.0444	0.0092	0.2125	0.2407	478
Split Ticket	0.0380	−0.0481	0.0106	0.2132	0.2418	442
VoteHub	0.0388	−0.0523	0.0159	0.2174	0.2407	478
Silver Bulletin Classic	0.0390	−0.0449	0.0097	0.2107	0.2407	478
FiftyPlusOne	0.0397	−0.0453	0.0093	0.2111	0.2418	442
VoteHub Legacy	0.0402	−0.0534	0.0146	0.2140	0.2407	478
Model average	*0.0410*	*−0.0510*	*0.0161*	*0.2156*	*0.2407*	*478*
On Point	0.0440	−0.0108	0.0055	0.2022	0.2407	478
Silver Bulletin Lite	0.0447	−0.0482	0.0148	0.2102	0.2407	478
Race to the WH	0.0450	−0.0613	0.0149	0.2092	0.2407	478
DDHQ	0.0456	−0.0503	0.0146	0.2101	0.2407	478
The Economist	0.0515	−0.0637	0.0172	0.2069	0.2418	442
Pollsmax	0.0537	−0.0772	0.0228	0.2077	0.2407	478
Poliwave	0.0547	−0.0591	0.0195	0.2052	0.2418	442

Table 1: Brier scores of Kalshi and election models. Each forecaster's probability that the Democrat wins, at 9:00 am ET on November 3, scored on every called Senate, House and Governor race. Bias: mean of $p - y$; negative means the forecaster underestimated Democrats. Lower reliability means better calibration; higher resolution means sharper discrimination between races. Kalshi prices are the raw price of the market on the candidate the models forecast, clipped to 0–1: the last trade, the volume-weighted average price over the hour before, and the bid/ask midpoint. Model average: the mean of the models covering a race. N: races scored.

Kalshi's lead survives several important data cuts. When each office carries equal weight (Table 4), On Point ranks first at 0.0393. Kalshi follows, with VWAP at 0.0422, the last trade at 0.0424 and the midpoint at 0.0426, ahead of every other model; Silver Bulletin Deluxe scores 0.0435 and the model average 0.0487. That table leaves out The Economist, Split Ticket, FiftyPlusOne and Poliwave, which publish no governor forecasts. In the House (Table 6), Kalshi is ahead of every model, with the last trade at 0.0288 against 0.0344 for the best model, Silver Bulletin Deluxe. In governor races (Table 7), Kalshi ranks second to On Point, which scores 0.0268. Kalshi's last trade and VWAP both score 0.0386, ahead of Silver Bulletin Deluxe at 0.0393 and the model average at 0.0456.

The Senate (Table 5) is the exception. There Kalshi sits in the middle of the pack, with VWAP seventh and the last trade ninth of 17 rows, at 0.0589 and 0.0597. On Point is best at 0.0457, and the model average scores 0.0616. The Senate has only 34 races, 12 of them between 10% and 90% on Kalshi, so a handful of misses weigh heavily. The largest was Mississippi, where Kalshi gave the Democrat, who won, 14%, followed by South Carolina, Iowa, Kansas and Nebraska.

Kalshi also has the lowest Brier score both in races with at least one poll and in races without one (Table 8). In the 169 polled races, Kalshi's VWAP scores 0.0778 and its last trade 0.0780,

against 0.0874 for the best model, Silver Bulletin Deluxe, and 0.1008 for the model average. In the 309 unpolled races, Kalshi's last trade scores 0.0064, against 0.0073 for the best model, VoteHub, and 0.0084 for the model average.

Figure 7 shows, race by race, whether Kalshi or the model was closer to the outcome. Each race is shaded by the model's squared error minus Kalshi's, against the model average in the top row and against Silver Bulletin Deluxe, one of the top models, in the bottom row. Most races are near zero because they were safe. The large differences sit in a few competitive races, and they run both ways: green where Kalshi was closer and red where the model was.

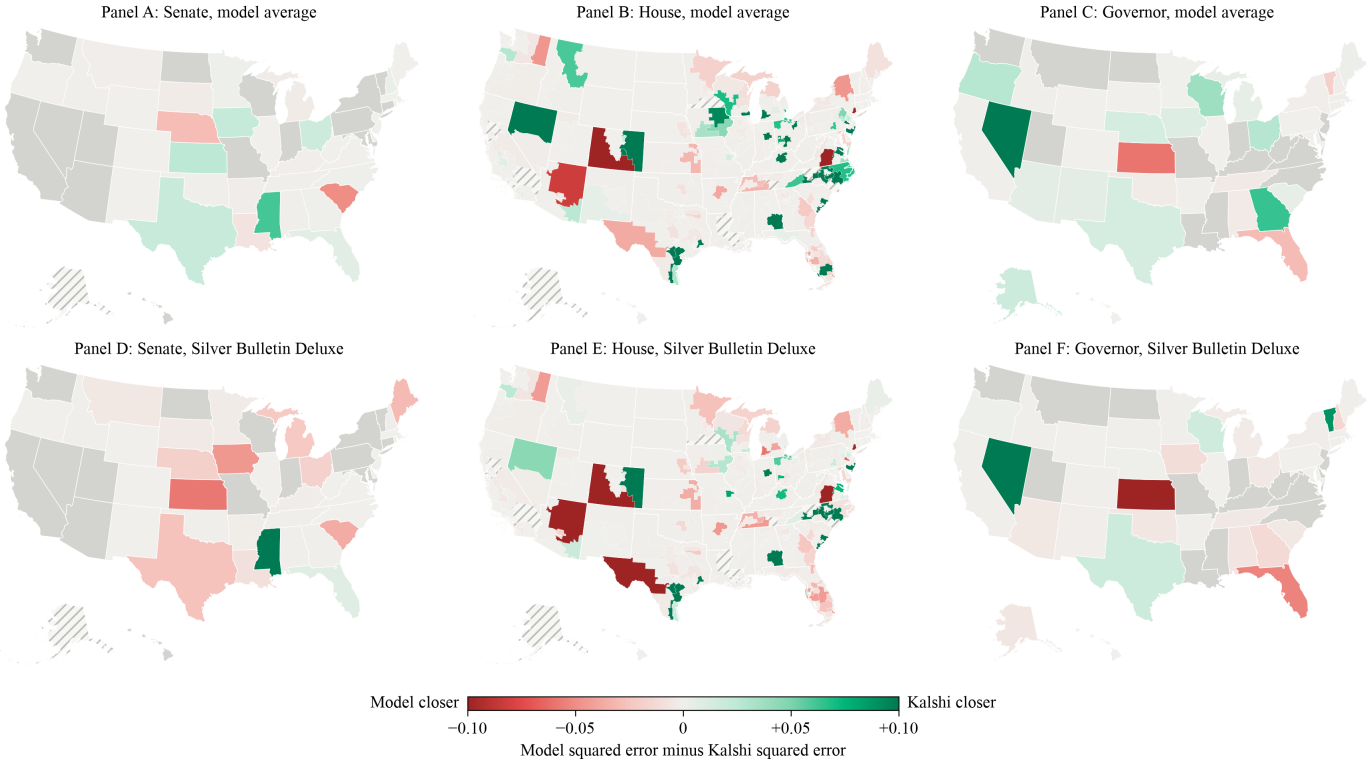


Figure 7: Who was closer to the outcome, Kalshi or the models, by race. Top row: Kalshi against the model average. Bottom row: Kalshi against Silver Bulletin Deluxe. Green: Kalshi closer; red: model closer. Hatched: no score; grey: no race.

Kalshi had a lower margin error than the polls. Kalshi's displayed median missed the margin by 4.55 points on average, against 6.17 for the VoteHub average in the 121 races it covers, and by 4.64 against 6.17 for the Wikipedia simple average in its 171 races. Kalshi's bias was smaller than the VoteHub average's and the same as the Wikipedia average's. These results are consistent with previous research that found prediction markets to be more accurate than polls and with the fact that prices incorporate additional information. This is true despite these markets having substantially lower volume compared to those trading directly on the candidate victory.

Forecaster	MAE	RMSE	Bias	N
Panel A: VoteHub average, 121 races				
Kalshi mean	**4.19**	**5.93**	**−3.36**	**121**
Kalshi displayed median	**4.55**	**6.34**	**−4.01**	**121**
Kalshi mode	**5.53**	**7.44**	**−1.78**	**121**
VoteHub average	6.17	8.57	−4.89	121
Panel B: Wikipedia simple average, 171 races				
Kalshi mean	**4.34**	**6.03**	**−3.40**	**171**
Kalshi displayed median	**4.64**	**6.29**	**−3.98**	**171**
Kalshi mode	**5.96**	**7.72**	**−1.71**	**171**
Wikipedia simple average	6.17	8.02	−3.98	171

Table 2: Vote margin errors of Kalshi and the poll averages, Democratic minus Republican, in points. Forecasts at 9:00 am ET on November 3: Kalshi's displayed median, mean and mode of its vote margin markets, from last trades, and the poll average. The outcome is the margin in the AP's unofficial vote count as of November 22. Error is the forecast minus the outcome; bias above zero means too Democratic. Each panel scores Kalshi and one poll average on the same races, those where both have a forecast. Rows are sorted by MAE.

Figure 8 maps, race by race, whether Kalshi or the poll average came closer to the vote margin. Kalshi was closer in 108 of the 171 races with a Wikipedia simple average (63%) and in 78 of the 121 races with a VoteHub average (64%). The House accounts for most of this: Kalshi was closer in 66 of 108 districts against Wikipedia and 51 of 71 against VoteHub. It was also closer in most governor races, 23 of 33 and 18 of 29. The Senate is the exception: Kalshi beat the Wikipedia average in 19 of 30 races but the VoteHub average in only 9 of 21, with the polls closer in several Southern states.

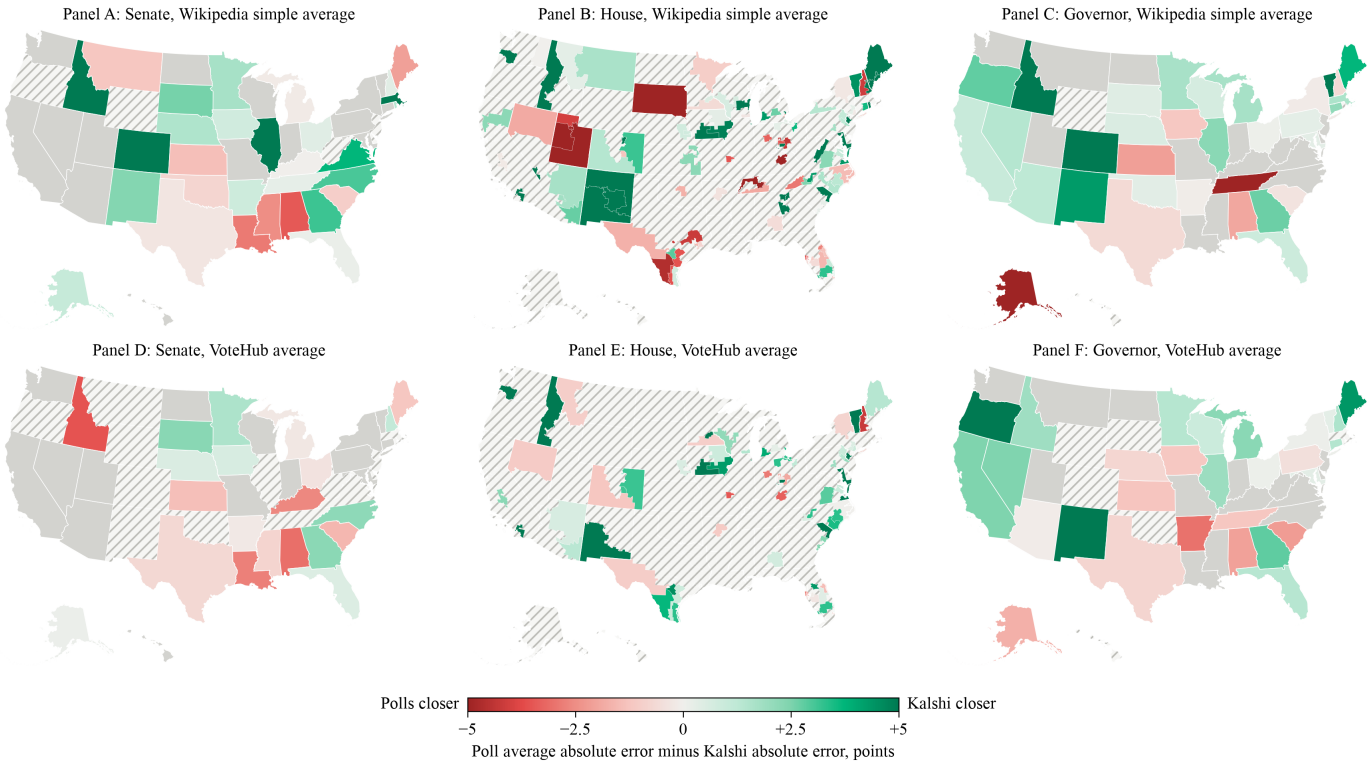
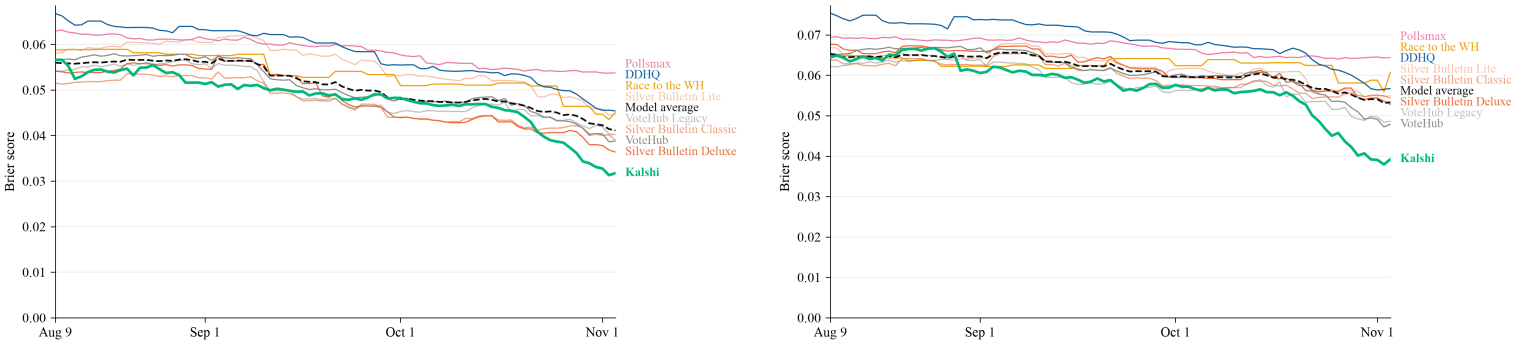


Figure 8: Who was closer to the vote margin, Kalshi or the polls, by race. Each race is shaded by the poll average's absolute error minus Kalshi's, in points of the Democratic minus Republican margin. Kalshi's forecast is the median of its vote margin markets at 9:00 am ET on November 3; the outcome is the margin in the AP's unofficial vote count as of November 22. Top row: Kalshi against the Wikipedia simple average. Bottom row: Kalshi against the VoteHub average. Green: Kalshi closer; red: poll average closer. Hatched: no poll average; grey: no race.

4.1 Forecast Error over Time

Figure 9a plots the Brier score of each day's forecasts from August 9 to November 3, scored on the races since called. Every forecaster's score fell as Election Day neared. From August through mid-October, Kalshi sat among the models, but below the model average, which stood at about 0.056 in early August, on every day except August 9–10 and September 30 to October 1. Silver Bulletin Deluxe had the lowest score of any model for most of October. Kalshi also had the lowest score of all on August 21 and from August 29 to September 10. From October 23, eleven days before the election, Kalshi's score was below every model's on every day, and on November 3 it was the lowest of all, as in Table 1.



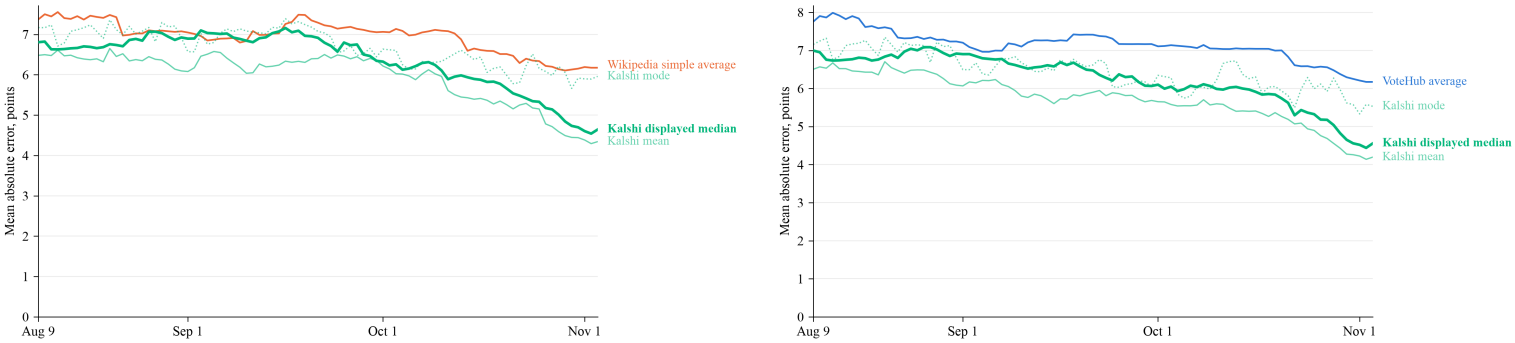
(a) All races

(b) Redistricted House races

Figure 9: Forecast error over the campaign. Brier score of each day's forecasts on the races since called, August 9 to 9 am ET on November 3. Panel (a): the 478 races of Table 1. Panel (b): House races in the nine states that redrew their congressional maps for 2026, namely Alabama, California, Florida, Louisiana, North Carolina, Ohio, Tennessee, Texas and Utah. Each day's forecasts are taken at midnight ET at the end of the day, and on November 3 at 9 a.m. ET. Models with a dated history only. The model average (dashed) is that of the models shown.

Figure 9b repeats the exercise for House races in the nine states that redrew their maps for 2026. District numbers there no longer match the 2024 territory, so past results and old polls map poorly onto the new districts. Kalshi's score was below the model average on every day except August 14 and August 18 to 28, and the gap widened sharply in late October. Kalshi ended near 0.039, against about 0.048 for VoteHub, the best model, and about 0.053 for the model average.

Figure 10 tracks the vote margin error from August 9 to Election Day. For virtually the whole period, Kalshi's displayed median had a lower error than both poll averages. The average error decreased from about 7 points in early August to about 4.5 for Kalshi and 6 for the polls in early November. The gap between Kalshi and the polls widened in the final month: on the morning of November 3, Kalshi's error was 4.64 points against 6.17 for the Wikipedia simple average and 4.55 against 6.17 for the VoteHub average. This is likely the effect from early voting data. Kalshi's mean and mode are shown for robustness.



(a) Wikipedia simple average

(b) VoteHub average

Figure 10: Vote margin error over the campaign, Kalshi against the poll averages. Mean absolute error, in points of the Democratic minus Republican margin, of each day's forecasts, August 9 to 9 am ET on November 3: Kalshi's displayed median, mean and mode of its vote margin markets, from last trades, and one poll average per panel. Each day's forecasts are taken at midnight ET at the end of the day, and on November 3 at 9 a.m. ET. The outcome is the margin in the AP's unofficial vote count as of November 22. Each day scores Kalshi and the poll average on the same races, those where both have a forecast.

4.2 Forecast Error by Call Order

Figure 11 adds races in the order they were called, on election night and after, and plots the cumulative Brier score of the November 3 forecasts. That is the score over every race called up to that point. Vertical lines mark the 8 p.m. poll closings, the North Carolina and Texas Senate calls, the calls of Senate and House control, the Maine and Michigan Senate calls, and the calls made the next day.

The first races called were the least competitive. These were safe seats called as polls closed, and every forecaster scores near zero on them. Through roughly the first 200 calls, Kalshi sits in the middle of the pack, behind Split Ticket, Silver Bulletin Deluxe and several other models. A possible explanation for this is a small longshot bias that acts as an overlay to the naturally lower calibration error achieved through 100/0 pricing. From the Texas Senate call onward, as competitive races were called, Kalshi moves near the top, with only Split Ticket and Silver Bulletin Deluxe lower. Every score jumps in the last races, which were the closest and were called late on election night or the next day. Kalshi's score rises least, and it finishes lowest at 0.0317. Kalshi's advantage therefore comes from the competitive, hard-to-call races.

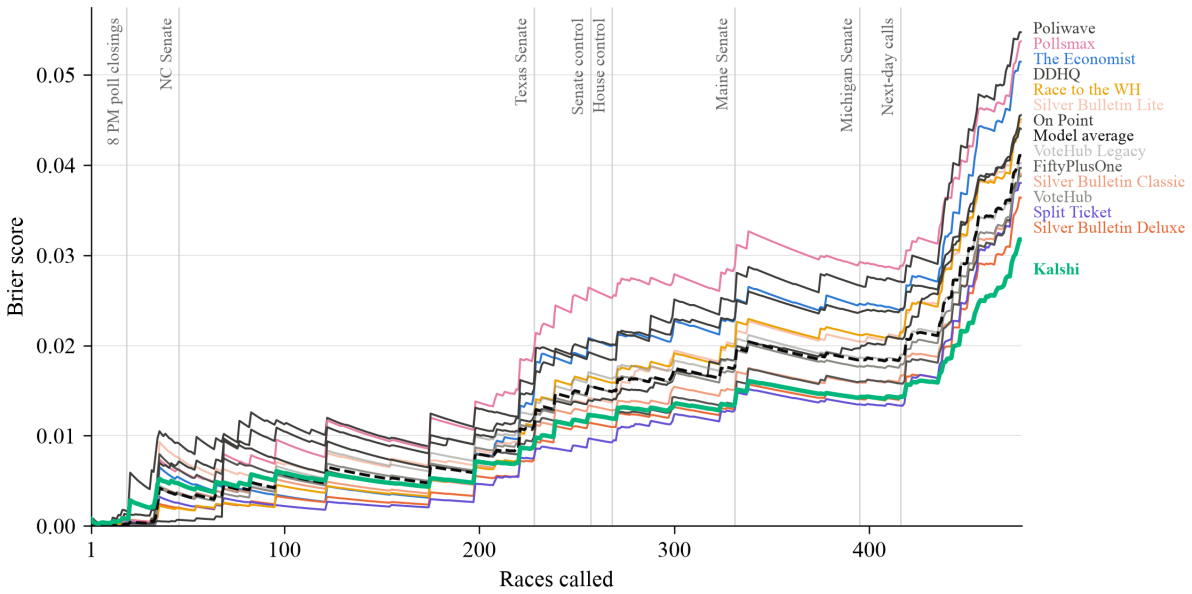


Figure 11: Forecast error as races were called. Cumulative Brier score of the 9 am ET November 3 forecasts, races in the order they were called. Vertical lines mark milestones of election night; the model average is dashed.

Table 3 sorts races into four ratings, taken from the model average alone, so that Kalshi's price has no part in choosing its own rating. Of the 478 races, 350 are Solid, with a chance under 5%; 32 are Likely, at 5 to 10%; 45 are Lean, at 10 to 25%; and 51 are Toss-ups, at 25 to 50%. The closeness row averages that chance within each rating: 0.9%, 7.4%, 16.4% and 35.0%.

In Solid races, the model average beats Kalshi: Kalshi's Brier score is 0.0010 against 0.0002. The difference, 0.0008, has a t-statistic of 8.32, but the errors involved are tiny. In Likely races Kalshi scores 0.0526 against 0.0595 for the model average, a difference of -0.0069 with a t-statistic of -0.69. The best models in these ratings, FiftyPlusOne in Solid races and Silver Bulletin Lite in Likely races, score 0.0000 and 0.0561.

In Lean and Toss-up races, Kalshi beats the model average by 0.0462 and 0.0478, with t-statistics of -2.60 and -4.35. Kalshi scores 0.1080 in Lean races against 0.1541 for the model average, and 0.1621 in Toss-ups against 0.2099. It also beats the best model in each of these ratings: VoteHub at 0.1434 in Lean races and On Point at 0.1702 in Toss-ups. The t-statistics treat races as independent; since races are correlated through the national environment, they overstate the precision of these differences. The Lean and Toss-up races therefore account for Kalshi's lower overall Brier score in Table 1.

	Solid	Likely	Lean	Toss-up
Closeness, %	0.9	7.4	16.4	35.0
Kalshi	**0.0010**	**0.0526**	**0.1080**	**0.1621**
Model average	*0.0002*	*0.0595*	*0.1541*	*0.2099*
Kalshi − model average	0.0008	−0.0069	−0.0462	−0.0478
	(8.32)	(−0.69)	(−2.60)	(−4.35)
Best model	0.0000	0.0561	0.1434	0.1702
	FiftyPlusOne	Silver Bulletin Lite	VoteHub	On Point

Table 3: Brier scores of Kalshi, the model average and the best model, by race rating. Ratings follow the chance of the less likely side in the model average at 9 am ET on November 3: Solid under 5% (350 races), Likely 5–10% (32), Lean 10–25% (45) and Toss-up 25–50% (51). t-statistics of the paired difference in squared errors in parentheses, treating races as independent.

5 Discussion

Prediction markets and the top election forecasting models performed very well in the 2026 Midterms. In the week leading up to November 3rd, Kalshi had a smaller average forecasting error than any model across all races, and in the House. Kalshi's predicted vote margins were also closer to the outcome than the polling averages, both on election morning and for almost the entire campaign. Kalshi's lead was especially pronounced in competitive races, redistricted states, and the House of Representatives.

The results have several limits. The largest is that they come from one election. The 478 races are correlated and were decided in the same national environment, so a forecaster that read that environment well was right in many races at once. In effect we observe one draw of the national error, not 478 independent ones. Second, ten of the races in scope were still uncalled when we froze the results on November 14 (the District of Columbia and the territories are outside the sample), among them the Alaska Senate race, where Kalshi gave Mary Peltola a 70% chance against 62% for the model average. A Peltola win would widen Kalshi's lead. The results of this study are not intended to suggest prediction markets as a strict replacement for election models or polls. Without election models, as with valuation models in asset prices, markets wouldn't be fair estimates of these real-world outcomes, and without polling, the underlying data powering such estimates would be unavailable. Where markets show their true potential is in the ability to discriminate between received articles of evidence and decide which to include, rather than to create from nothing.

It is interesting to note that though many House markets were thinly traded, with about 100 traders in the median race by poll close, Kalshi still outperformed in races even when not covered by pollsters. More research ought be done on the trader and participant dynamics that created such an outcome.

6 Conclusion

For most of the history of election forecasting, the question has been which method predicts best. The 2026 midterms suggest a more useful question: what each method knows that the

other does not. Models are strong where evidence is structured and plentiful, converting polls, fundamentals and expert ratings into coherent, correlated probabilities across hundreds of races. Markets are strong where evidence is scarce, fragmentary or new. The races in which Kalshi gained most, the competitive contests, the redrawn districts and the late campaign, are those in which the available evidence was least tidy, and Kalshi also scored best in the 309 races that had no polls at all.

The 2026 cycle is best read as a baseline rather than a verdict. Markets at this scale are new, the comparison rests on a single election, and the mechanisms behind their late-campaign advantage remain to be tested. The presidential election of 2028, with deeper liquidity and a larger trader base, will show whether the pattern documented here persists as markets become a more established part of the forecasting landscape.

References

Arrow, Kenneth J., Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O. Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D. Nelson, George R. Neumann, Marco Ottaviani, Thomas C. Schelling, Robert J. Shiller, Vernon L. Smith, Erik Snowberg, Cass R. Sunstein, Paul C. Tetlock, Philip E. Tetlock, Hal R. Varian, Justin Wolfers, and Eric Zitzewitz. 2008. "The Promise of Prediction Markets." *Science* 320 (5878): 877–878.

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.

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. 2011. "How Prediction Markets Can Save Event Studies." NBER Working Paper 16949.

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.

Zitzewitz, Eric. 2026. "Are Prediction Markets Politically Biased?" NBER Working Paper 35846.

A Appendix

Following Snowberg, Wolfers, and Zitzewitz (2011), the full effect E of an event on an asset price s over a window from t to t', given the event's prior probability π, is

$$E = \frac{s_{t'} - s_t}{1 - \pi}, \tag{5}$$

with π taken from a prediction market. From 4 p.m. ET to the market's close at 11:47 p.m. (Figure 12), Kalshi's probability of a Republican Senate fell from 27.0% to 0.1% and S&P 500 futures fell 1.04%, implying that a Republican rather than a Democratic Senate was worth 3.9% of the index. Snowberg, Wolfers, and Zitzewitz (2011) find a 3.0% rise (standard error 1.2%) on Republican presidential victories from 1928 to 1996.

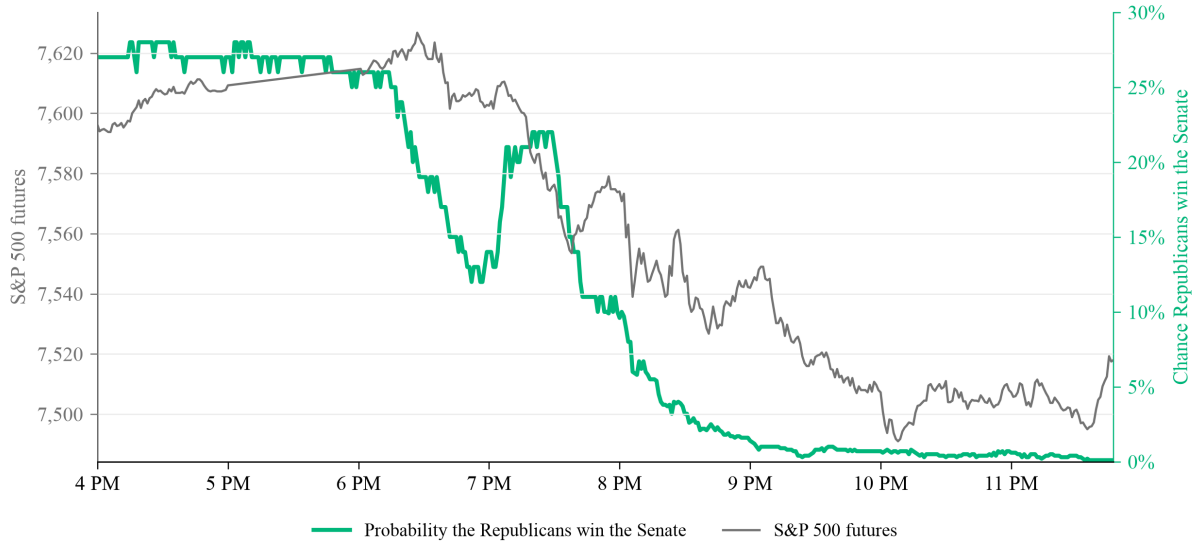


Figure 12: Kalshi's probability that Republicans win the Senate (right axis) and E-mini S&P 500 futures (left axis), 1-minute data, 4 p.m. to 11:47 p.m. ET on November 3, 2026. Futures do not trade from 5 to 6 p.m. ET.

B Forecast Error by Chamber

This appendix reports the robustness checks behind Section 4. Every table uses the measures of Table 1, scored on the 9 a.m. ET November 3 forecasts, and sorts forecasters by Brier score.

Table 4 weights each office equally, so the 34 Senate and 36 governor races count as much as the 408 House races. Only the forecasters covering all three offices appear. On Point ranks first, at 0.0393, and Kalshi's three prices follow, between 0.0422 and 0.0426, ahead of every other model.

Forecaster	Brier score	Bias	Reliability	Resolution	Uncertainty	N
On Point	0.0393	−0.0196	0.0156	0.2116	0.2346	478
Kalshi VWAP	**0.0422**	**−0.0502**	**0.0130**	**0.2049**	**0.2346**	**478**
Kalshi last price	**0.0424**	**−0.0498**	**0.0131**	**0.2050**	**0.2346**	**478**
Kalshi midpoint	**0.0426**	**−0.0540**	**0.0134**	**0.2050**	**0.2346**	**478**
Silver Bulletin Deluxe	0.0435	−0.0560	0.0138	0.2038	0.2346	478
Silver Bulletin Classic	0.0484	−0.0585	0.0165	0.2020	0.2346	478
Model average	*0.0487*	*−0.0685*	*0.0289*	*0.2160*	*0.2346*	*478*
Silver Bulletin Lite	0.0489	−0.0547	0.0157	0.1997	0.2346	478
VoteHub	0.0510	−0.0754	0.0221	0.2040	0.2346	478
Race to the WH	0.0510	−0.0719	0.0206	0.2037	0.2346	478
VoteHub Legacy	0.0538	−0.0789	0.0271	0.2069	0.2346	478
DDHQ	0.0592	−0.0795	0.0288	0.2054	0.2346	478
Pollsmax	0.0713	−0.1068	0.0441	0.2062	0.2346	478

Table 4: Brier scores of Kalshi and the election models, each office weighted equally. Same measures as Table 1, but each race is weighted by one over three times the number of called races in its office, so Senate, House and governor races each count for a third. Bias, the Murphy bins and the base rate are weighted the same way. Only forecasters covering all three offices are shown: The Economist, Split Ticket, FiftyPlusOne and Poliwave publish no governor forecasts. N: races scored (34 Senate, 408 House, 36 governor). Rows are sorted by Brier score.

Tables 5, 6 and 7 score each office on its own. In the Senate, Kalshi sits in the middle of the pack, and On Point is best. In the House, which makes up most of the sample, Kalshi's three prices are the three best. In governor races, Kalshi is second to On Point.

Forecaster	Brier score	Bias	Reliability	Resolution	Uncertainty	*N*
On Point	0.0457	−0.0348	0.0287	0.2215	0.2362	34
Poliwave	0.0490	−0.0760	0.0211	0.2067	0.2362	34
Race to the WH	0.0522	−0.0700	0.0303	0.2166	0.2362	34
Silver Bulletin Lite	0.0548	−0.0654	0.0172	0.1950	0.2362	34
Silver Bulletin Deluxe	0.0567	−0.0782	0.0173	0.1947	0.2362	34
Split Ticket	0.0583	−0.0862	0.0200	0.1943	0.2362	34
Kalshi VWAP	**0.0589**	**−0.0795**	**0.0222**	**0.1994**	**0.2362**	**34**
Kalshi midpoint	**0.0597**	**−0.0833**	**0.0246**	**0.2018**	**0.2362**	**34**
Kalshi last price	**0.0597**	**−0.0804**	**0.0243**	**0.2018**	**0.2362**	**34**
Silver Bulletin Classic	0.0605	−0.0835	0.0209	0.1947	0.2362	34
Model average	*0.0616*	*−0.0911*	*0.0402*	*0.2166*	*0.2362*	*34*
FiftyPlusOne	0.0624	−0.0909	0.0367	0.2097	0.2362	34
VoteHub	0.0683	−0.1032	0.0298	0.1945	0.2362	34
VoteHub Legacy	0.0719	−0.1052	0.0324	0.1945	0.2362	34
DDHQ	0.0748	−0.1050	0.0586	0.2215	0.2362	34
The Economist	0.0906	−0.1406	0.0652	0.2090	0.2362	34
Pollsmax	0.1002	−0.1459	0.0744	0.2092	0.2362	34

Table 5: Brier scores of Kalshi and the election models, Senate. Same measures as Table 1, on the 34 called Senate races. Rows are sorted by Brier score.

Forecaster	Brier score	Bias	Reliability	Resolution	Uncertainty	*N*
Kalshi last price	**0.0288**	**−0.0221**	**0.0092**	**0.2225**	**0.2422**	**408**
Kalshi midpoint	**0.0290**	**−0.0279**	**0.0094**	**0.2223**	**0.2422**	**408**
Kalshi VWAP	**0.0290**	**−0.0230**	**0.0093**	**0.2224**	**0.2422**	**408**
Silver Bulletin Deluxe	0.0344	−0.0413	0.0087	0.2156	0.2422	408
VoteHub	0.0354	−0.0459	0.0159	0.2225	0.2422	408
Split Ticket	0.0363	−0.0450	0.0100	0.2149	0.2422	408
Silver Bulletin Classic	0.0364	−0.0412	0.0101	0.2149	0.2422	408
VoteHub Legacy	0.0365	−0.0463	0.0135	0.2182	0.2422	408
FiftyPlusOne	0.0378	−0.0415	0.0090	0.2130	0.2422	408
Model average	*0.0389*	*−0.0462*	*0.0141*	*0.2165*	*0.2422*	*408*
DDHQ	0.0418	−0.0422	0.0125	0.2130	0.2422	408
Race to the WH	0.0433	−0.0583	0.0142	0.2115	0.2422	408
Silver Bulletin Lite	0.0436	−0.0463	0.0156	0.2139	0.2422	408
On Point	0.0454	−0.0084	0.0042	0.2011	0.2422	408
The Economist	0.0482	−0.0573	0.0141	0.2076	0.2422	408
Pollsmax	0.0489	−0.0690	0.0183	0.2093	0.2422	408
Poliwave	0.0552	−0.0577	0.0201	0.2057	0.2422	408

Table 6: Brier scores of Kalshi and the election models, House. Same measures as Table 1, on the 408 called House races. Rows are sorted by Brier score.

Forecaster	Brier score	Bias	Reliability	Resolution	Uncertainty	N
On Point	0.0268	−0.0155	0.0264	0.2222	0.2222	36
Kalshi last price	**0.0386**	**−0.0469**	**0.0123**	**0.1944**	**0.2222**	**36**
Kalshi VWAP	**0.0386**	**−0.0480**	**0.0123**	**0.1944**	**0.2222**	**36**
Kalshi midpoint	**0.0391**	**−0.0509**	**0.0126**	**0.1944**	**0.2222**	**36**
Silver Bulletin Deluxe	0.0393	−0.0487	0.0255	0.2083	0.2222	36
Model average	*0.0456*	*−0.0682*	*0.0449*	*0.2222*	*0.2222*	*36*
Silver Bulletin Lite	0.0483	−0.0525	0.0266	0.2014	0.2222	36
Silver Bulletin Classic	0.0485	−0.0509	0.0480	0.2222	0.2222	36
VoteHub	0.0494	−0.0771	0.0321	0.2037	0.2222	36
VoteHub Legacy	0.0530	−0.0852	0.0522	0.2222	0.2222	36
Race to the WH	0.0576	−0.0875	0.0399	0.2037	0.2222	36
DDHQ	0.0609	−0.0914	0.0309	0.1944	0.2222	36
Pollsmax	0.0648	−0.1054	0.0642	0.2222	0.2222	36

Table 7: Brier scores of Kalshi and the election models, governor. Same measures as Table 1, on the 36 called governor races. The Economist, Split Ticket, FiftyPlusOne and Poliwave publish no governor forecasts; the model average is the mean of the models shown. Rows are sorted by Brier score.

C Polled and Unpolled Races

Table 8 splits the races by whether they had a poll by November 3. Kalshi's three prices have the three lowest Brier scores in both groups: in the 169 polled races, where every forecaster underestimated Democrats by the most, and in the 309 unpolled races, which are mostly safe seats and where every forecaster's error is small.

Panel A: races with polls

Forecaster	Brier score	Bias	*N*
Kalshi VWAP	**0.0778**	**−0.0894**	**169**
Kalshi last price	**0.0780**	**−0.0883**	**169**
Kalshi midpoint	**0.0784**	**−0.0933**	**169**
Silver Bulletin Deluxe	0.0874	−0.1154	169
Silver Bulletin Classic	0.0944	−0.1208	169
On Point	0.0944	−0.0727	169
VoteHub	0.0963	−0.1365	169
VoteHub Legacy	0.0997	−0.1399	169
Model average	*0.1008*	*−0.1378*	*169*
Split Ticket	0.1045	−0.1397	136
Silver Bulletin Lite	0.1088	−0.1368	169
FiftyPlusOne	0.1097	−0.1347	136
Race to the WH	0.1112	−0.1523	169
DDHQ	0.1142	−0.1470	169
Pollsmax	0.1363	−0.1935	169
The Economist	0.1456	−0.1918	136
Poliwave	0.1566	−0.1791	136

Panel B: races without polls

Forecaster	Brier score	Bias	*N*
Kalshi last price	**0.0064**	**0.0048**	**309**
Kalshi midpoint	**0.0065**	**−0.0008**	**309**
Kalshi VWAP	**0.0067**	**0.0041**	**309**
VoteHub	0.0073	−0.0063	309
VoteHub Legacy	0.0077	−0.0061	309
DDHQ	0.0080	0.0026	309
Model average	*0.0084*	*−0.0036*	*309*
Split Ticket	0.0085	−0.0074	306
Silver Bulletin Deluxe	0.0085	−0.0056	309
FiftyPlusOne	0.0086	−0.0056	306
Pollsmax	0.0086	−0.0136	309
Silver Bulletin Classic	0.0087	−0.0034	309
Race to the WH	0.0089	−0.0115	309
Poliwave	0.0094	−0.0058	306
The Economist	0.0096	−0.0068	306
Silver Bulletin Lite	0.0097	0.0003	309
On Point	0.0164	0.0230	309

Table 8: Brier scores of Kalshi and the election models in races with and without polls. A race has polls if, by 9 am ET on November 3, it had VoteHub's polling average, or at least one poll with a reported sample size of its Democratic (or non-Republican) nominee against its Republican nominee in VoteHub's poll database or Wikipedia's poll tables, or a poll we added by hand. Rows are sorted by Brier score.

D Races Where the favourite Won by Horizon

Figures 14, 15 and 16 show, at each horizon, the races in which Kalshi's favourite won. The share called right rises as Election Day nears: in the Senate from 29 of 34 races three months out to 32 at the first poll closing, in the House from 389 of 417 to 401, and in governor races from 32 of 37 to 35. These counts cover every resolved race with a Kalshi market, so they include the nine uncontested House races and the Virgin Islands governor race, which are not in the 478-race sample.

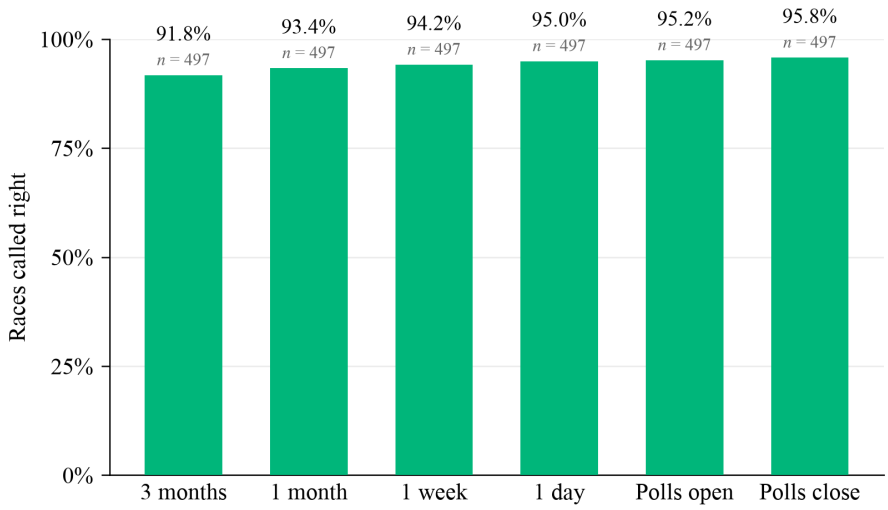


Figure 13: Share of races in which Kalshi's favourite won, by horizon: three months, one month, one week and one day before polls opened in the state, when they opened, and at the first poll closing. The favourite is the race's highest-priced market at its last trade before each time. Every resolved race with a Kalshi market is counted, n is the number of races priced at that horizon.

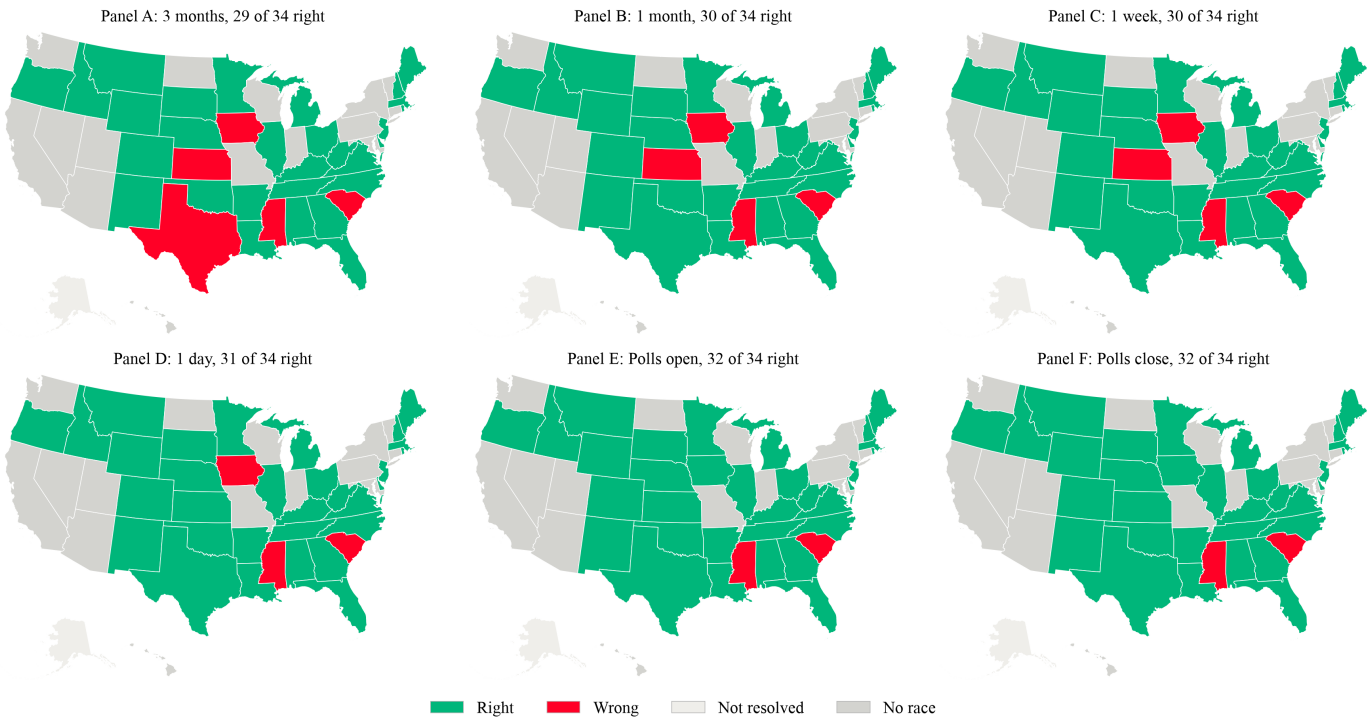


Figure 14: Senate races Kalshi called right, from three months before polls opened to the first poll closing in each state.

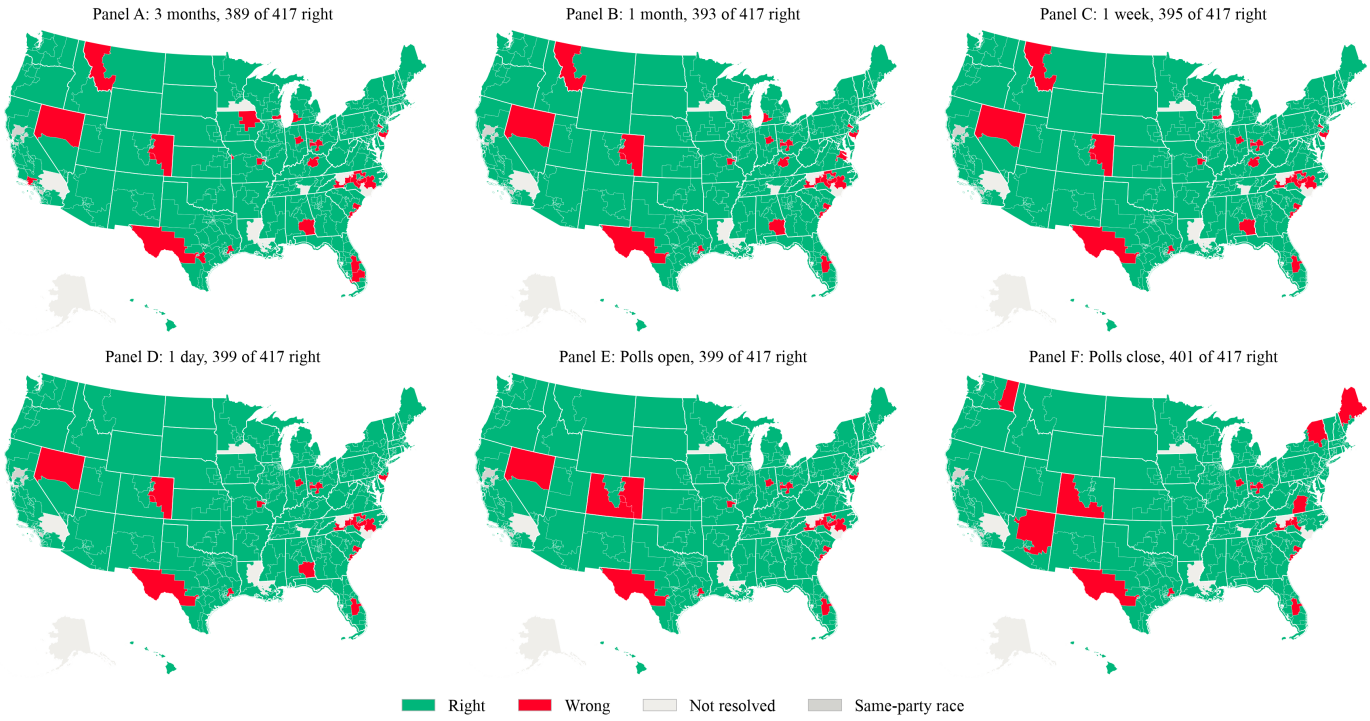


Figure 15: House races Kalshi called right, from three months before polls opened to the first poll closing in each state. Grey: same-party race.

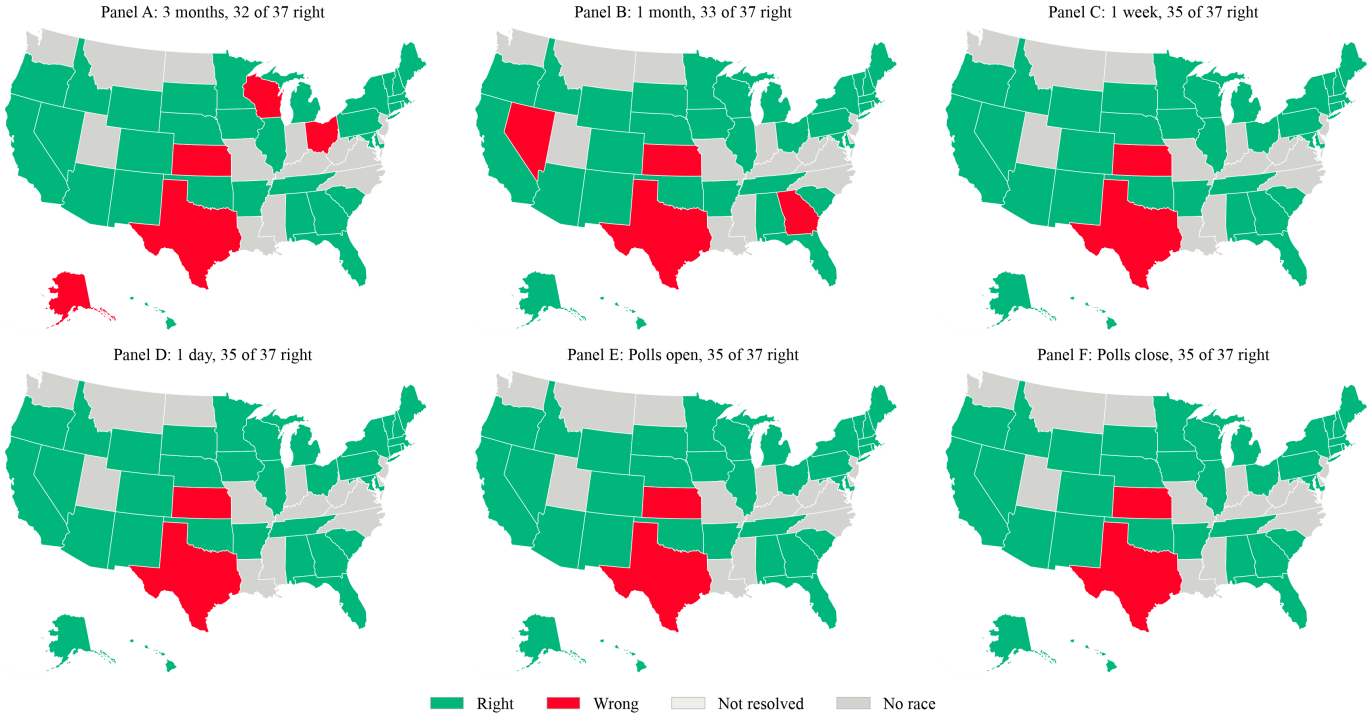


Figure 16: Governor races Kalshi called right, from three months before polls opened to the first poll closing in each state.

E 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 check consistent with the market functioning as an efficient information-aggregation device, or whether frictions such as risk aversion, the favourite–longshot bias, or fee structure drive a wedge between price and true probability.