

Do Learned Combination Weights Help Sparse-Polling Congressional Forecasts?

A chronological development study, with separate prospective extensions

Intuitionist-Fin research working paper · Lampbotics · revised 4 October 2026

AI contributors: Codex — implementation, analysis and manuscript drafting. Claude Code — audit and independent recomputation of the preceding results. **Human stewardship:** Lampbotics.

Status: Revised development working paper, not peer reviewed. AI contributors are named by their roles, not presented as journal-approved human authors. Human authorship and approval must be established before journal submission. Claude Code's 3 October audit independently reproduced the original model, scores and comparisons, and identified defects and methodological weaknesses. The new calibrated-reference comparisons and operational revisions below are builder-tested; they have not received another independent audit.

Version note (4 October): Sections 1–7 retain the original six-cycle, 2012–2022 development study and its audit-revised comparisons. Appendix C documents the subsequently implemented Compass/Tide comparison, poll-weighting study and whole-chamber scenarios. The chamber diagnostic uses a different eligible sample through 2024; it must not be pooled with the original study. No 2026 election outcomes have been observed. Claude Code has now reproduced the v2 chamber calculations and identified input defects and directional bias. The v3 repairs described in Appendix D await follow-up review; other new extensions are not evidence of improvement over established forecasters.

Abstract

Congressional forecasting faces uneven polling coverage, changing district boundaries and shared election-wide errors. We examine whether a small forecasting workflow can use limited local polls more effectively without treating additional complexity as an achievement in itself. Separate Senate and House models combine local polling, geographic partisanship measured on the relevant district map, and a conservative indicator of prior federal elected experience. A regularized logistic function selects the balance between polling and background predictive distributions using earlier chronological forecasts. We evaluate six election cycles, 2012–2022, at 28-, 14- and 7-day horizons; all fitted coefficients, uncertainty adjustments and combination weights use earlier cycles. The learned combination improves average Brier scores over ordinary regression in all six chamber-horizon groups, but the corresponding exploratory

election-cycle uncertainty intervals include zero. Stronger comparisons substantially qualify the result. A fixed blend assigning 75% to the polling model has lower average Brier scores in four of six groups, and in all six groups restricted to the three cycles in which learned weights are active. The original interval-score comparison favored learned weights, but the fixed blends had not received their own scale calibration and their ranges were overwide. Separately identified, earlier-cycle recalibration weakens that advantage. The learned weights also vary little within fitted folds. These qualifications favor a simpler primary research arm. These are development findings, not evidence of superiority over established forecasters. The contribution is an auditable comparison workflow and a specific negative result: sparse election histories do not yet justify learning flexible combination weights. We implement separately registered prospective statistical comparisons that separate prediction, source correction and AI interpretation. An October 4 extension adds a simple national-swing reference, a poll-weighting experiment and conditional chamber seat scenarios; these do not change the original empirical findings.

Keywords: congressional elections; sparse polling; forecast evaluation; predictive distributions; redistricting; AI-assisted research.

1 Introduction

A small election forecasting project can acquire polls and publish probabilities quickly. Producing evidence that those probabilities deserve trust is harder. A district with one campaign-sponsored poll may appear data-rich in a dashboard while providing little independent information about the electorate. More elaborate software does not repair that limitation.

This study begins with a practical question: can a small, reproducible workflow improve predictions for a selected set of congressional races when local polling is scarce? Its theoretical premise is modest. Past geographic voting patterns and candidates' electoral histories can provide useful background, while current polls can reveal departures from that background. Neither is a sufficient description of a campaign. The question is how much to rely on each, and whether estimating that balance improves forecasts enough to justify its complexity.

We develop and compare separate Senate and House models rather than treating their election environments as interchangeable. We also preserve the geographic meaning of House predictors. A district number does not identify the same electorate after redistricting. Prior presidential voting is therefore calculated on the map associated with the election being predicted.

The central result is cautionary. Learned combination weights improve on a relatively weak regression benchmark, but do not establish an advantage over a simple fixed blend. That distinction matters for political methodology: a favorable comparison with a model that has fewer predictors is not evidence that a more flexible combination function adds information. We treat the fixed-blend comparison as a revision of the evaluation, not an excuse to select a new winner from the same historical elections.

This paper is a development working paper. We inspected historical outcomes while designing the workflow. Chronological fitting reduces direct leakage of held-out outcomes into parameters, but does not undo researcher knowledge or retrospective choices. No prospective 2026 outcome evidence is available, and no journal-readiness or commercial-performance claim follows from the results.

2 Political and statistical background

2.1 Why combine polls and background information

Political forecasting has long distinguished short-term polling variation from relatively stable conditions that help predict votes. Gelman and King’s presidential study offers an influential account of this distinction; it does not establish that the same relationships or error rates hold in congressional races. Here it motivates a question, not a transferred empirical coefficient. [Gelman and King 1993](#).

A much closer precedent is Chen, Garnett and Montgomery’s Senate forecast. Their hierarchical Dirichlet model combines fundamentals and time-varying polls with Gaussian-process priors, and evaluates both retrospective and prospective forecasts. They already demonstrate that few covariates can support serious election prediction. Our regression-and-mixture implementation is simpler and is not a new invention of sparse-input forecasting. [Chen, Garnett and Montgomery 2023](#).

The political interpretation of our predictors is also limited. Relative presidential voting is a proxy for local partisan alignment, not a direct measure of congressional preferences. A record of prior federal victory may capture name recognition, political experience, selection into candidacy or other advantages. Its coefficient is predictive; it is not an estimate of the causal effect of experience or incumbency.

2.2 Why agreement is not necessarily confirmation

Polls can share sampling difficulties, turnout assumptions and weighting practices. AAPOR’s review of 2024 describes remaining problems in representing voter groups and changes in turnout, alongside improved overall polling performance. It also documents variation in directional errors across election cycles. We therefore do not assume that several firms, or repeated apparent agreement, eliminate systematic uncertainty. [AAPOR 2025](#).

The same caution applies to combinations of polls, news and prediction-market prices. They may respond to the same information. In this implementation, geographic history also appears inside the polling expert, so mixing it with a background expert does not combine two independent measurements. Our mixture describes alternative predictive distributions; it is not a multiplication of independent likelihoods.

2.3 What model combination adds

Stacking predictive distributions has an established statistical literature. Yao and colleagues develop distributional combination using predictive performance, and subsequent hierarchical stacking allows weights to vary with inputs. Our logistic weighting function is related in spirit but uses penalized point estimates and chronological validation. It is neither full Bayesian hierarchical stacking nor an original theory of model averaging. [Yao et al. 2018](#), [Yao et al. 2022](#).

Our distinguishing practical choice is to make the availability and character of polling visible to the combination function. Number of firms, poll age and disagreement are candidate signals of when polling deserves more weight. This produces a testable expectation: learned weights should improve on a fixed blend with the same component forecasts. The present results do not support that expectation consistently.

3 Data and forecast target

3.1 Outcome definition

The margin target is the Democratic nominee's vote share minus the Republican nominee's vote share on the specified general-election ballot, in percentage points. Shares use the reported total vote denominator, including other candidates. Fusion ballot lines for the identified nominee are combined. The binary event is that this margin is strictly positive. A tie is coded as not Democratic-ahead.

This target is not automatically equivalent to winning a seat. Runoffs, special elections, ranked-choice rounds, independent candidates and same-party final ballots can break that equivalence. We require one distinct named nominee from each major party in the historical result source, exclude special elections and unsupported ballot-round cases, and do not substitute eventual seat victory for the margin target.

The resulting models describe individual races. They do not simulate correlated seat outcomes or estimate which party will control either chamber.

3.2 Historical results and geography

The source inventory contains 9,229 normalized House contests, 775 Senate contests and 663 state-presidential records before modeling exclusions. These counts describe the normalized archives, not the evaluation sample. Federal results come from MIT Election Lab sources. Geographic presidential results come from The Downballot's map-specific exports, including earlier Daily Kos Elections work. [MIT Election Lab](#), [The Downballot](#).

For each modeled election, we use the most recent strictly earlier presidential election. Local presidential margin is centered on the national margin from that same presidential year. The national calculation requires complete records for all 50 states and the District of Columbia. House geography is matched to the declared map vintage; we never carry a previous House outcome across changed boundaries solely because its district number matches.

The usable modeling data from 2006–2024 contain 3,650 House cases and 295 Senate cases at each horizon before selecting polled test cases. Missing map features are excluded explicitly. Some historical district percentages are rounded or omit small write-in totals. The exports were retrieved in 2026 and may contain retrospective corrections. “Strictly prior election” is therefore not the same as “verified source publication available at the time.” Census map versions provide an additional check, but do not resolve every district’s legal status. [U.S. Census Bureau](#).

Prior elected experience is a deliberately narrow indicator. We match normalized candidate names to earlier federal major-party winners in the same state during the previous eight years. A past winner must have beaten the other major nominee and the maximum possible individual third-party vote total. An unmatched name is absence of affirmative evidence, not proof of no experience. State office, appointed service, name changes and some long careers are not captured. Contemporary incumbent columns in district spreadsheets are ignored.

3.3 Polling and its availability

The frozen polling archive covers 2006–2022. We retain regular Democratic–Republican general-election pairs, use MIT margins for scoring, and retain disagreements with archive outcome fields as diagnostics rather than dropping cases because outcomes differ. The archive defines candidates retrospectively using the final top two finishers and records a median survey date, not a publication timestamp. These features limit any claim to reconstructing information available before an election. [FiveThirtyEight data documentation](#).

The main replay assumes a poll becomes available three days after its median survey date. A separate replay uses six days. Available polls are retained within 21 days of the forecast cutoff. Their weights halve every 14 days. Questions from the same survey are collapsed, each firm’s polls are averaged, and the firm’s total contribution is capped at the weight of its most recent poll. Counts of firms are not claims of statistical independence.

The present numerical model does not estimate sponsor effects, sample-design effects or respondent-level uncertainty. These omissions are especially consequential when only one sponsored poll is available. Publication of public chances of winning remains withheld; richer numerical comparisons are saved as research forecasts.

3.4 Evaluation population

Polled forecasts are evaluated in 2012, 2014, 2016, 2018, 2020 and 2022 at 28, 14 and 7 days before election day. The evaluation includes 158 distinct Senate election-races and 421 distinct House election-races across any of those dates. A race may appear at several horizons; summing the table’s counts would count some outcomes repeatedly.

Within each chamber and horizon, all methods use exactly the same test cases. The samples differ across horizons and availability assumptions. Our primary estimates apply to the selected, matched, polled contests in this archive. They do not estimate performance over all congressional contests, all close races or the current watchlist.

We define a competitive subgroup before consulting its current outcome: absolute centered prior presidential margin at most ten points. This definition is imperfect, but avoids selecting close races by the result being predicted. A separate background-only evaluation extends through 2024; it is not pooled with these polling comparisons. Absence of 2024 polls in the archive is not evidence that those elections had no polls.

4 Methods

4.1 Forecast components

Models are fitted separately by chamber and forecast horizon. Each earlier election year contributes equal total weight, preventing a poll-rich cycle from dominating solely through its number of cases. This does not remove correlations within or across cycles.

We compare eight methods:

Method	Inputs and purpose
Polls alone	Recent polling average with a fixed normal six-point standard deviation
Ordinary regression	Geographic partisanship and polling average
Experience regression	Geographic partisanship, polls and prior federal elected-experience difference
Robust regression	The same three predictors, fitted with Huber residual weights
Background only	Geographic partisanship and prior elected-experience difference, scored on the same polled cases
Fixed 50/50 blend	Equal mixture of background and experience-regression predictive distributions
Fixed 75/25 blend	75% experience-regression distribution and 25% background distribution
Learned weighting	The same two distributions with a weight estimated from earlier chronological forecasts

The fixed blends are added after inspecting the initial development results. They are untuned reference settings, not preregistered confirmatory tests. The experience component is called a “polling expert” below because it incorporates polls; it also includes the background predictors.

4.2 Regression and uncertainty

Ordinary regression uses no ridge penalty and rejects a rank-deficient design. Other regressions use a fixed small ridge penalty of 0.01, with an unpenalized intercept. Vote-margin predictors are divided by ten; the experience difference remains on its integer scale. Robust regression uses Huber’s 1.345 cutoff and at most 60 iterations.

Regression predictive distributions use Student-t tails with five degrees of freedom. Their scale combines a weighted training residual variance, a three-point minimum residual standard deviation, and an approximate coefficient-uncertainty term. The covariance approximation is not a full posterior and is not adjusted for election-level clustering. The same coefficient-uncertainty construction for robust regression is also approximate. Gate uncertainty and uncertainty in fitted scale factors are not fully integrated.

Once at least three earlier chronological forecast cycles and thirty cases are available, scales are adjusted using an equal-cycle-weighted quantile of absolute standardized forecast errors. Before that, training-error uncertainty remains an explicitly labeled fallback. The adjustment targets an 80% central margin range; it does not guarantee future coverage, especially under changes in polling methods or electoral behavior.

4.3 Learned combination

Let F and P denote the background and polling-expert distributions. The learned forecast is a linear mixture:

$$G(y \mid x) = [1 - w(x)] F(y \mid x) + w(x) P(y \mid x)$$

$$w(x) = \text{logistic}(b_0 + b_1 \log(1 + \text{number of firms}) \\ + b_2 \text{average poll age} / 14 \\ + b_3 \text{disagreement across firms} / 6)$$

The four coefficients maximize an equal-election-weighted predictive density objective, with a fixed L2 penalty of 2 around a prior centered at a 75% polling weight. Coefficients are bounded between -8 and 8 . The signs are unrestricted; the fit is not guaranteed to give newer or more numerous polls greater weight.

The training records are forecasts of earlier cycles made without those cycles' outcomes in the component fits. Their uncertainty adjustments likewise use still earlier forecast errors. The combination function requires three such cycles and thirty cases. As a result, learned weighting is inactive in the 2012–2016 test folds and equals the experience-regression component there. It is active in 2018, 2020 and 2022. We report those three active cycles separately.

A mixture preserves component disagreement as possible outcomes rather than replacing it with only an average estimate. Its 10th and 90th percentiles are obtained from the mixture distribution. This can produce wider ranges than a single component, but disagreement does not guarantee that every mixture interval is wider in every case.

4.4 Chronological evaluation and scores

Every test year uses only earlier years for coefficient fitting, uncertainty adjustment and gate fitting. This follows the general predict-then-evaluate logic of prequential analysis, while remaining a retrospective development exercise. [Dawid 1984](#).

The primary binary-event metric is the Brier score, the squared difference between the predicted Democratic-ahead probability and the observed zero-or-one outcome. We also report binary log loss. For margins, we report absolute error, central-80 coverage, interval width and the central-80 interval score. The latter penalizes both unnecessary width and missed outcomes. [Gneiting and Raftery 2007](#).

```
Brier(p, z) = (p - z)^2

IntervalScore80(l, u, y) = (u - l)
                        + 10 max(l - y, 0)
                        + 10 max(y - u, 0)
```

Scores are first averaged within election year and then equally across years. Coverage is a calibration diagnostic, not sufficient evidence that a range is useful. A very wide range can achieve high coverage while earning a poor interval score. Binary probability accuracy also need not rank methods in the same order as margin-range usefulness.

We calculate paired differences within each year on matched cases. Exploratory uncertainty intervals resample the six whole election-year differences 10,000 times. Election-wide error makes individual-race resampling inappropriate for the main comparison. Six cycles are still few, and consecutive elections may be dependent; these bootstrap intervals are descriptive sensitivity summaries, not reliable confirmatory confidence bounds. We make no multiplicity-adjusted superiority test.

5 Results

5.1 More background helps relative to the initial comparison

Table 1 gives equal-election-year average Brier scores. Lower is better. The learned combination improves on ordinary regression in all six groups, but that comparison changes both the set of inputs and the way distributions are combined. Experience regression or robust regression already obtains much of the gain.

Table 1. Matched historical Brier scores with the three-day availability assumption

Chamber	Days	Cases	Polls alone	Ordinary	Experience	Robust	Fixed 75/25	Learned
Senate	28	134	0.0791	0.0724	0.0687	0.0684	0.0686	0.0685
Senate	14	130	0.0793	0.0672	0.0659	0.0669	0.0664	0.0657
Senate	7	149	0.0670	0.0622	0.0613	0.0599	0.0607	0.0613
House	28	260	0.1603	0.1333	0.1310	0.1304	0.1255	0.1282
House	14	288	0.1622	0.1359	0.1296	0.1297	0.1206	0.1278
House	7	319	0.1532	0.1288	0.1244	0.1252	0.1199	0.1244

Cases are election-races available at each date, summed across six cycles. The averages give each cycle equal weight; they are not a pooled average across all cases. Exact values and the additional background-only and 50/50 comparisons are in the accompanying CSV tables.

All six exploratory uncertainty intervals for the learned-minus-ordinary Brier differences cross zero. Figure 1 also compares learned weighting with the stronger fixed 75/25 reference. The fixed blend does better on average in four groups. Its largest advantage is in House forecasts two weeks before election day: learned weighting has a Brier score about 0.0072 higher. That difference's descriptive bootstrap interval is approximately 0.0015 to 0.0140. Because the comparison was added after development, with few cycles and several comparisons, this is not a confirmatory finding of fixed-blend superiority.

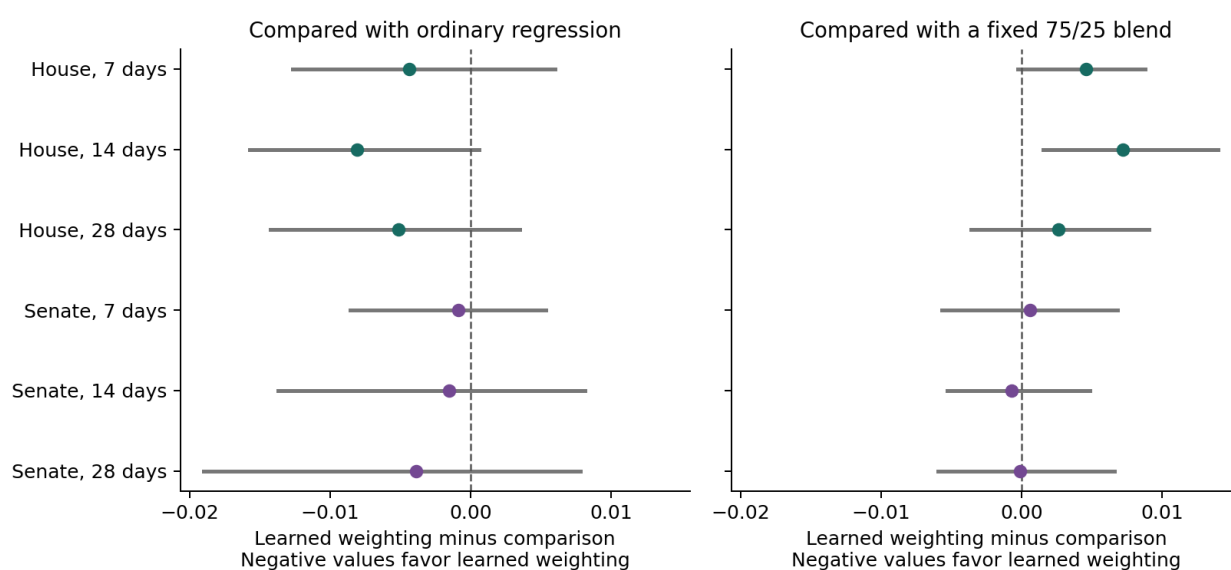


Figure 1. Paired election-year differences in Brier score. Negative values favor learned weighting. Lines show exploratory cycle-bootstrap intervals, not confirmatory tests.

5.2 The learned weights have not earned their complexity

Table 2 isolates 2018–2022, the cycles with enough earlier forecast records to estimate the gate. On these matched active-year cases, the fixed 75/25 blend has a lower mean Brier score in all six groups. We cannot count identical early-year component forecasts as evidence that the learned gate itself worked.

Table 2. Brier scores when learned weighting is active

Chamber	Days	Experience regression	Fixed 75/25	Learned weighting
Senate	28	0.0639	0.0619	0.0635
Senate	14	0.0609	0.0597	0.0604
Senate	7	0.0599	0.0582	0.0599
House	28	0.1331	0.1232	0.1274

Chamber	Days	Experience regression	Fixed 75/25	Learned weighting
House	14	0.1356	0.1292	0.1321
House	7	0.1325	0.1318	0.1326

Each value averages three election years. This is a diagnostic subgroup, not a new held-out validation sample.

The fixed 50/50 blend has an even lower House two-week Brier score, 0.1185, than the 75/25 blend's 0.1206. Conversely, its Senate scores are generally worse. Selecting a chamber-specific weight from these same outcomes would be another development decision. The appropriate next step is to preserve both references and test them prospectively, rather than advertise whichever historical setting wins.

5.3 The original interval comparison needs a calibration qualification

Learned weighting has narrower average ranges than the fixed 75/25 blend. It also has a lower central-80 interval score than that blend in all six main comparisons, despite losing four Brier comparisons. Binary log loss favors learned weighting in all three Senate groups and the fixed blend in all three House groups. This is not evidence that adaptive weighting itself produced better ranges: the fixed mixtures were not separately scale-calibrated and their central-80 coverage was approximately 88–93%. Table 3 preserves the original result; the revision below addresses this comparison imbalance.

Table 3. Central-80 interval score with the three-day assumption

Chamber	Days	Ordinary	Experience	Robust	Fixed 75/25	Learned
Senate	28	25.36	24.04	23.97	24.29	23.26
Senate	14	23.81	23.63	23.47	24.63	24.02
Senate	7	20.21	19.85	19.44	22.65	19.79
House	28	28.40	27.95	27.91	28.55	27.63
House	14	26.65	25.84	25.86	27.54	26.39
House	7	27.45	26.89	26.81	28.34	27.62

Scores are in percentage-point units and penalize width plus outside misses. Lower is better. “Learned” is not best in every group.

The learned method's central-80 coverage ranges from 74.7% to 82.8% in Senate tests and 77.7% to 81.5% in House tests. Its Senate four-week range remains too narrow relative to its stated target. No automatic widening was selected after observing that held-out coverage. A future scale revision needs earlier-only estimation and evaluation under a separately identified model version.

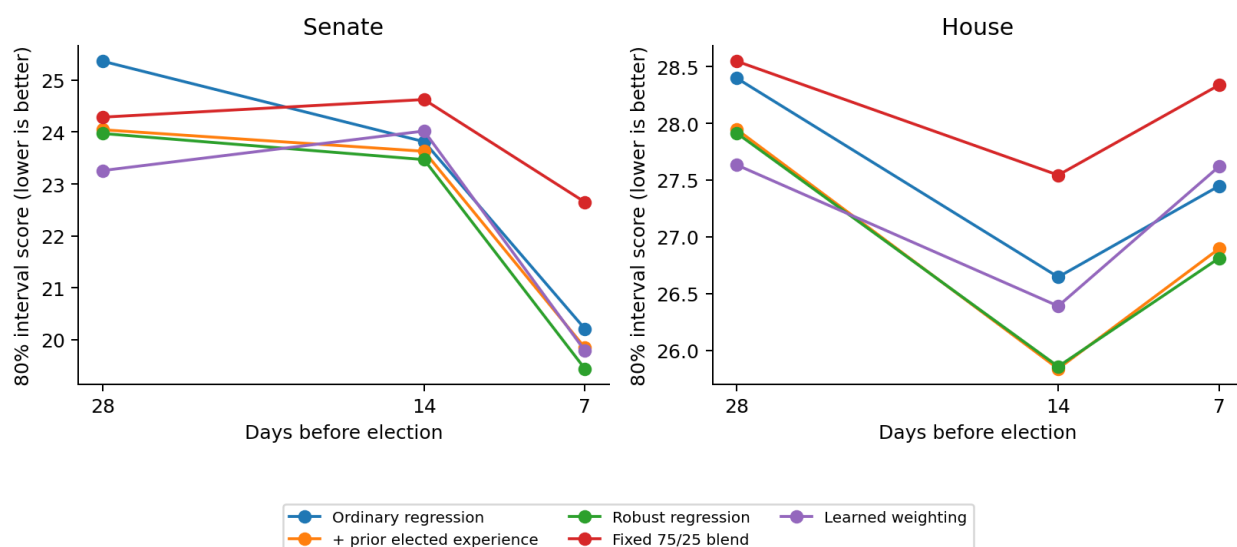


Figure 2. Original interval scores before calibrating fixed blends separately. These comparisons are affected by overly wide fixed-blend ranges; see the audit revision below.

5.4 Sensitivity and close races

The six-day availability replay preserves the learned model’s average advantage over ordinary regression in all six groups, while the fixed 75/25 blend has lower Brier scores in four. The two buffers select different cases, so these findings do not establish the causal effect of waiting longer for information.

Table 4. Brier scores under the six-day assumption

Chamber	Days	Cases	Ordinary	Fixed 75/25	Learned
Senate	28	129	0.0814	0.0761	0.0778
Senate	14	125	0.0711	0.0674	0.0669
Senate	7	135	0.0668	0.0666	0.0662
House	28	234	0.1260	0.1227	0.1235
House	14	253	0.1476	0.1301	0.1392
House	7	290	0.1305	0.1218	0.1280

In the competitive subgroup, learned weighting’s Senate two-week Brier score is 0.1117, compared with ordinary regression’s 0.1116: essentially no gain. House results favor learned weighting over ordinary regression, but the fixed 75/25 blend is better than learned weighting at all three horizons in the main replay. The close-race findings are therefore weaker than the favorable all-case comparison initially suggested.

Background-only forecasts are markedly worse on these same polled cases than when evaluated across a broad population dominated by safe seats. For example, the main House four-week background-only Brier score is 0.1561 on polled cases. A previously

reported 0.0371 across a much broader background-only population cannot be treated as a competing score for the same task. Forecast coverage and forecast difficulty must be reported together.

5.5 Audit revision: simpler primary arm and better references

Claude Code independently reproduced the original artifacts and scores on 3 October. Its review found that fitted polling weights across active folds ranged from about 0.75 to 0.98 and varied only modestly within a fold. This suggests a mostly constant polling-heavy blend rather than strong adaptation to polling conditions. This characterization supports simplifying the primary arm; it is not a causal account of every fitted coefficient.

We add two separately labelled references without altering earlier forecasts. The polls-only mean is the local polling margin; its normal standard deviation is the root mean squared out-of-sample polling error from earlier matched cycles, with equal cycle weights and a three-point floor. It requires three earlier cycles and thirty cases; otherwise the original six-point assumption remains an explicitly uncalibrated fallback. This is a fitted error reference, not a fitted polling bias correction.

For each fixed mixture, calculate each earlier forecast's absolute margin error divided by the distance from its mean to the relevant lower or upper central-80 endpoint. The cycle-weighted 80th percentile of these ratios supplies a scale factor, floored at 0.5, after three earlier cycles and thirty cases. Apply it as an affine transformation of the full mixture distribution around its original mean: both component means move relative to the mixture mean, and component scales multiply by the factor. Recompute event probabilities and quantiles from that transformed mixture. Do not merely shrink displayed endpoints while keeping an inconsistent probability. Component coefficients and original predictions remain unchanged.

Only earlier held-out forecasts from the same chamber and horizon enter reference calibration. Insufficient-history folds retain their disclosed fallback. These additional choices follow inspection of historical results and remain development choices. They do not guarantee future coverage.

Table 5. Audit-revised references, three-day assumption

Chamber	Days	Experience	Fixed 75/25 recalibrated	Polls only, fitted scale	Learned (unchanged)
Senate	28	0.0687 / 72.1% / 24.04	0.0654 / 86.7% / 23.84	0.0800 / 76.6% / 26.75	0.0685 / 74.7% / 23.26
Senate	14	0.0659 / 79.3% / 23.63	0.0627 / 88.3% / 22.81	0.0810 / 74.3% / 26.95	0.0657 / 82.8% / 24.02
Senate	7	0.0613 / 80.0% / 19.85	0.0585 / 88.1% / 21.10	0.0674 / 74.2% / 22.15	0.0613 / 80.0% / 19.79
House	28	0.1310 / 78.2% / 27.95	0.1237 / 87.8% / 26.98	0.1567 / 70.9% / 32.71	0.1282 / 81.5% / 27.63

Chamber	Days	Experience	Fixed 75/25 recalibrated	Polls only, fitted scale	Learned (unchanged)
House	14	0.1296 / 76.3% / 25.84	0.1195 / 86.7% / 26.68	0.1627 / 68.9% / 30.25	0.1278 / 77.7% / 26.39
House	7	0.1244 / 77.4% / 26.89	0.1191 / 87.4% / 27.36	0.1540 / 72.0% / 31.26	0.1244 / 78.2% / 27.62

Columns give Brier score / central-80 coverage / central-80 interval score. Lower Brier and interval scores are better. Coverage is a diagnostic, not an optimization objective. The calibrated fixed blend is still overwide in the full-period average, partly because early folds use the unchanged fallback. The calibrated polls-only reference still undercovers, especially in the House. We report these deficiencies rather than force the held-out results to 80%.

The calibrated fixed 75/25 blend beats adaptive weighting on Brier in all six groups, and on interval score in three of six groups. Thus the original six-of-six interval-score advantage is not robust to a more comparable scale treatment. These are post-development, dependent comparisons; no superiority claim follows. Results and calibration receipts are exported separately as `audit_revision_results.csv` and `audit_revision_receipts.json`.

6 Revised workflow and prospective design

The review changes the evaluation and the daily research workflow. It does not replace the pinned statistical fits. Daily records include fixed 50/50 and 75/25 blends, their separately calibrated revisions, a calibrated polls-only reference, and background-only predictions on the same eligible polled races, alongside the existing methods. Future margin scoring also records interval width and interval score. These are CPU-only additions; they require no paid model call.

The race-level watchlist contains 26 races; 24 are structurally supported, but fresh evidence and other checks can reduce the number with a published estimate on any date. Maine and Ohio Senate remain withheld pending support for their election formats. Missouri uses a polling-only fallback while its district map is reviewed. Candidate mappings remain research assumptions pending official ballot verification. These limits are recorded with the forecasts and should be treated as coverage failures or exclusions in reporting, not silently omitted successes.

Prospective statistical evaluation uses forecasts recorded before fixed 14:00 UTC deadlines at 28, 14 and 7 days before the 2026 general election. For each method and race, the latest eligible record within the preceding 24 hours is selected. Models are distinguished by method and frozen artifact hash. Certified, matching general-ballot margins are required to score the Democratic-ahead event. An eventual nominee victory cannot replace a missing margin.

For the amended prospective continuation, the primary statistical contrast is the experience regression against the separately calibrated fixed 75/25 blend, with Brier score as the primary metric. Learned weighting is secondary. The original fixed blend

remains a separate secondary reference. An append-only policy record identifies this amendment, its race list, target, source-artifact hash and reference version before the first deadline. Other methods, binary log loss and interval score remain secondary comparisons. This is an internal prospective design amendment, not externally registered research. The method choices have already used historical outcomes; only future results are genuinely prospective.

We will publish matched-case scores, total registered races, attempted forecasts, abstentions, exclusions and missing outcomes. Forecasts revised later will not replace earlier records. Comparisons with established forecasters require probabilities for the same target and timestamp; seat-winner probabilities are not automatically equivalent. Historical aggregator probabilities have not been acquired or evaluated in this study.

AI contextual interpretation remains separate from numerical models. The matched statistical-versus-agent evaluation described here is a proposed comparison, not a completed test. Existing workers can read a pinned set of polling facts and background information without seeing statistical probabilities. They can abstain and must name the same outcome target. If they identify a source error, the correction should be given to all arms: correcting bad input is a useful research service but does not demonstrate superior interpretation when the comparison model is left with incorrect facts.

A second agent arm using additional reporting would test the value of the entire research workflow, not reasoning alone, because its information differs. Repeated model calls, source freshness, cost and unavailable replies must be counted. No live AI forecast trial was run for this manuscript. Retrieval-assisted language-model forecasting has prior empirical research, but its findings do not validate these workers or congressional tasks. [Halawi et al. 2024](#).

The election workers currently use isolated sessions with native learning disabled. This study estimates statistical coefficients and weights; it does not train a foundation model or demonstrate agents developing intuition. Persistent memory and skill learning would require a distinct prospective design that measures retrieval of earlier lessons, preserves failures and controls for information already contained in model training.

7 Limitations and political science contribution

The largest threat to validity is the historical information set. The polling archive uses final candidates, publication times are assumed, and geographic exports can contain corrections made after the modeled elections. Strict chronological fitting cannot certify real-time availability. Future data collection should preserve both original publication time and collection time, including source corrections and ballot eligibility.

The second threat is selection. Polling is not allocated randomly across races; campaigns can choose which internal results to publish. Requiring a conventional Democratic-Republican pair also narrows the political universe. Success on this selected sample cannot establish accuracy on all House seats or provide a chamber-control forecast.

Third, the number of political environments is small. Hundreds of district observations do not create hundreds of independent national elections. Equal-year weighting and cycle resampling recognize this problem but do not solve it. Coefficient and gate uncertainty may be understated. Student-t margin distributions also allow values outside the mathematically possible -100 to $+100$ range; they are continuous approximations, not a bounded vote-share model. A richer national-environment model might help, but would require historical input vintages and stronger evaluation; it is not automatically the next improvement.

Fourth, the background variables are incomplete. Geographic presidential preference is several years old. The experience proxy omits state office, candidate fundraising and appointed service, and exact-name matching can fail. Predictive gains do not identify which political mechanisms generated them.

The current intellectual contribution is therefore an application and evaluation argument, not a new estimator. It links geography, information availability, matched comparisons and model restraint in a small congressional forecasting workflow. Its most informative finding is the failure of a learned gate to justify its complexity against a fixed mixture when genuine gate-active cycles are isolated. That result supports reporting stronger comparisons and multiple forecast objectives rather than equating sophistication with progress.

A stronger political science contribution could come from a prospective explanation of when contextual evidence helps: candidate replacement, turnout information, poll sponsorship or local issues, evaluated through recorded cases and matched targets. That would require observations not yet present in this manuscript. It is a research direction, not a claimed result.

8 Conclusion

Polls combined with a few relevant background variables provide a reasonable experimental starting point for congressional forecasting. In our chronological development evaluation, adding prior elected experience and simple distributional combination is more defensible than claiming a breakthrough from learned weights. Fixed blends remain strong references; interval comparisons also require comparable calibration.

The next evidence must come from forecasts preserved before outcomes, original-source verification and independent review. Until then, the workflow is useful for learning what fails and how to compare alternatives fairly. It does not establish that a small forecasting project can outperform established services.

Appendix A Implementation details

A regression expert predicts a margin from an intercept and its listed features. Each training observation in election year t receives weight $N/(T n_t)$, where N is the number of training observations, T the number of earlier cycles and n_t that cycle's sample size. The intercept is unpenalized. The residual variance is the weighted mean squared residual, bounded below by 9; it is not an unbiased degrees-of-freedom estimate.

Approximate coefficient covariance equals that residual variance times the inverse penalized weighted cross-product matrix. Predictive variance adds the residual variance and the quadratic-form coefficient term. Multiplication by $\sqrt{3/5}$ converts the predictive standard deviation to the five-degree-of-freedom Student-t scale before the chronological adjustment is applied.

For calibration, each earlier forecast error is divided by its unadjusted predictive scale. The equal-cycle-weighted 80th percentile of the resulting absolute errors is divided by the t distribution's 90th-percentile quantile. The fitted scale factor has a lower bound of 0.5. Calibration uses unscaled errors to avoid applying a correction twice.

The gate optimizes mixture density for the continuous margin, not Brier score directly. This difference in objectives can contribute to the binary-score and interval-score rankings. The density objective and binary event scores do not provide causal attribution for the estimated gate coefficients.

The fixed blends combine the two full distributions. Their probabilities equal the weighted average of component probabilities, but their interval endpoints are mixture quantiles, not averages of component endpoints. Background-only predictions in the matched comparison use exactly the earlier fitted expert; no additional parameter fit uses the test outcomes.

Appendix B Reproduction and data availability

The deployed model remains `bop-evidence-aware-v1`. Its main artifact SHA256 is `2eca72999b2d88919349059c8602a63b6c0ac9459d78b09f90b91cb59f1596b8`. The six-day model artifact is `dc37364b51208674925947a6772814b3c997be154835d4f93eeefe8f82a6f557`.

The revised evaluation is `bop-matched-ablation-v1`. The main evaluation artifact is `e642b0dec915c1a155d98502bf2f755abd26bf4e170a23e318b5eb4576088a7`; the sensitivity evaluation is `8466195040f8cd2ec3f3ed42f3c3a2b9d2ec066f1ca09510517ce8a5b76cd9ea`. The evaluations add comparisons without refitting components or modifying existing forecasts. Full local artifacts contain per-case predictions, fitted objects and exclusions. Aggregate tables and figures accompany this manuscript.

The original full model rebuild reproduced its artifact hash. The revised evaluation checks source-artifact integrity, earlier-only fitting metadata, identical comparison samples and output hashing. The original builder run passed seventy-nine runtime tests, including four new tests of interval-score behavior, immutable matched comparisons, tampering and future-calibration refusal, and equal-cycle aggregation. Software tests cannot establish scientific validity or prospective accuracy.

Raw source exports and original poll text are not redistributed with this paper because their respective reuse permissions require verification. Source URLs, retrieval receipts, hashes and exclusion logs are available in the project’s local artifacts. This is not yet a fully public replication package. Instructions for reproducing the tables, figures and document exports are in the accompanying README.

Research disclosures

This is an in-house experimental project at Lampbotics. No external funding is asserted. The project has an interest in developing an election-analysis product; that interest does not supply evidence of forecast quality. The election work remains separate from stock decisions.

Codex wrote the code and draft. Human stewards must verify its facts, citations, authorship and conclusions before circulation as a submission. Claude Code’s audit of the preceding system informed the comparison design, but no independent audit of these new results has been completed. No human-subject survey was conducted for this study; it uses existing aggregate election and poll records. No new paid or live language-model call was made for this review.

Audit revision receipts

The new comparison is `bop-matched-ablation-v1-audit-revision-v1`, with source-model hash unchanged. Main revised evaluation:

58bd267e7cca8feb8fa723f61340aa794c2397b533427c58b0900e7d4d443426. Calibration is reconstructed from earlier fold predictions only. Six-day revised evaluation:

79af101263468c466c9776856f1811a46dc4b717365e003161f0c7986a2977a6. Both full replays completed; receipts accompany the aggregate tables. These revisions have not received Claude Code’s follow-up review.

References

AAPOR. 2025. Task Force on 2024 Pre-Election Polling Report. Executive summary. [Report](#).

Chen, Yehu, Roman Garnett, and Jacob M. Montgomery. 2023. “Polls, Context, and Time: A Dynamic Hierarchical Bayesian Forecasting Model for US Senate Elections.” *Political Analysis* 31(1): 113–133. Published online 2022. [doi:10.1017/pan.2021.42](https://doi.org/10.1017/pan.2021.42).

Dawid, A. P. 1984. “Statistical Theory: The Prequential Approach.” *Journal of the Royal Statistical Society, Series A* 147(2): 278–292. [doi:10.2307/2981683](https://doi.org/10.2307/2981683).

FiveThirtyEight. Pollster Ratings Data and Codebook. Archived polling source and documentation, accessed 3 October 2026. [Documentation](#).

Gelman, Andrew, and Gary King. 1993. “Why Are American Presidential Election Campaign Polls So Variable When Votes Are So Predictable?” *British Journal of Political Science* 23: 409–451. [Author manuscript](#).

Gneiting, Tilmann, and Adrian E. Raftery. 2007. “Strictly Proper Scoring Rules, Prediction, and Estimation.” *Journal of the American Statistical Association* 102(477): 359–378. [doi:10.1198/016214506000001437](https://doi.org/10.1198/016214506000001437).

Halawi, Danny, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. 2024. “Approaching Human-Level Forecasting with Language Models.” arXiv:2402.18563. [Preprint](#).

MIT Election Lab. Election Results Data. House, Senate and state-presidential archives. [Data directory](#). Specific files and versions are identified by the project receipts.

The Downballot and Daily Kos Elections. Presidential Election Results by Congressional District. Map-specific exports. [Directory](#). Retrieved source versions are identified by the project receipts.

U.S. Census Bureau. Congressional Districts. [Map-vintage directory](#). Accessed 3 October 2026.

Yao, Yuling, Aki Vehtari, Daniel Simpson, and Andrew Gelman. 2018. “Using Stacking to Average Bayesian Predictive Distributions.” *Bayesian Analysis* 13(3): 917–1007, including discussion. [doi:10.1214/17-BA1091](https://doi.org/10.1214/17-BA1091).

Yao, Yuling, Gregor Pirš, Aki Vehtari, and Andrew Gelman. 2022. “Bayesian Hierarchical Stacking: Some Models Are (Somewhere) Useful.” *Bayesian Analysis* 17(4). First released online 2021. [doi:10.1214/21-BA1287](https://doi.org/10.1214/21-BA1287).

Appendix C. Implemented extensions as of 4 October 2026

C.1 Scope and separation of experiments

The deployed race regression is called **Compass**. It uses local polls when available, geographically matched past voting and earlier federal elected experience. The learned weighting method remains a secondary research arm. The 26-race watchlist contains six Senate and twenty House races. A structurally eligible race is not necessarily a currently available forecast: missing evidence, unusual ballots, source conflicts and stale runs can all require withholding it.

The following extensions preserve the earlier fitted artifact and historical results. They have separate policies and outcome targets. They should not be merged into a single claim that the system has learned to forecast better. Daily AI source reading is operational, but that is not a completed, controlled test of AI forecasting skill. AI prose, prediction-market prices and search attention do not adjust the published numerical models.

C.2 Tide: a deliberately simple national-swing reference

Tide’s Democratic-minus-Republican margin is:

local 2024 presidential margin – national 2024 presidential margin
+ current national congressional polling margin

Both coefficients equal one. No coefficient is fitted and no model judgment changes the result. House geography must match the reviewed 2026 district input. Unsupported ballots and unresolved map cases are withheld. National polls come from the retained VoteHub feed, with source review and first-observed timestamps. Use one latest eligible survey per reviewed firm within 21 days; prefer likely voters to registered voters within a survey, and exclude adults-only surveys from the main estimate. Average eligible firms equally and require at least three. Source conflicts and ambiguous duplicates are excluded rather than repaired with guessed numbers. Firm labels do not prove independent respondents.

The initial October 4 reviewed snapshot had one eligible firm, so Tide did not publish race estimates. That is a dated launch observation, not a permanent state or a zero margin. A current run can publish estimates when the fixed eligibility rules are satisfied. Tide has no win probabilities or uncertainty ranges. Presidential partisan geography may not transfer to congressional contests; local candidates and turnout can invalidate uniform swing.

Compass and Tide are compared on the same races and forecast deadlines using absolute margin error. This prospective comparison is distinct from the Brier-score evaluation of probabilistic models. No prospective scores are available before certified results.

C.3 Poll detail inspection and four weighting arms

The source reader locates passages about sample size, population, field dates, recruitment, mode, weighting, voter screens, wording, sponsorship and related features. These passages are inspection aids, not verified structured methodology. The initial October 4 snapshot contained 64 current poll rows, five readable original-source profiles and zero fully verified methodology audits. Missing fields remain unknown. Source profiles do not silently change numerical weights.

A separate polls-only experiment compares (1) recency weighting, (2) recency plus sample size, (3) those weights plus historical peer consistency, and (4) those weights plus peer-relative house-effect adjustments. All use a 21-day window and a 14-day recency half-life. The sample multiplier is $\sqrt{n/800}$, limited to 0.5–1.5; missing sample sizes receive neutral weight. This is not an effective-sample-size calculation. Within-firm polls are averaged; a firm's total influence is capped at its largest individual weight. Multiple releases do not multiply its influence.

Historical firm deviations are measured against other firms in the same race, using at least two peers. Earlier election cycles alone fit chamber- and horizon-specific adjustments. Deviations are averaged equally across cycles, shrunk by $n/(n+20)$ after at least two cycles, and capped at plus or minus three points before subtraction from

current margins. Peer-consistency weights are also shrunk and bounded. They measure agreement with peers, not truth: a correctly dissenting pollster can be penalized. No automatic adjustment for partisan sponsorship or inferred demographics is applied.

Chronological tests cover 2012–2022, with earlier training observations back to 2006. Three- and six-day assumed publication delays address sensitivity, not verified historical availability. Table C1 reports equal-cycle average absolute margin error for the three-day assumption. It is a polls-only comparison, not a test against Compass.

Chamber	Days	Recency	+ sample	+ peer consistency	+ house effect
Senate	28	5.620	5.634	5.624	5.621
Senate	14	5.747	5.795	5.776	5.783
Senate	7	5.164	5.191	5.192	5.334
House	28	7.098	7.073	7.074	7.020
House	14	6.809	6.808	6.812	6.775
House	7	6.686	6.678	6.680	6.607

Small House gains and Senate regressions do not justify replacing Compass. The same races at several horizons are dependent observations. Live weighting estimates remain a separate experiment, with abstentions recorded. Full per-cycle results are in the accompanying POLL_WEIGHTING_RESULTS.json.

C.4 Conditional chamber seat scenarios

The initial registered version described in this subsection was `bop-chamber-scenarios-v2`; the audited repair is `v3`, described in Appendix D. Its October 4 inventory accounts for all 435 House and 100 Senate seats. House accounting is 405 modeled races, nine certified same-party ballots and 21 unresolved races. Senate accounting is 27 modeled races, eight unresolved races and 65 continuing seats. At registration, 14 modeled House seats and five Senate seats had matched local polls; other modeled seats used background inputs. This is an inventory and evidence snapshot, not comprehensive polling coverage.

Continuing Senate seats are 32 Democratic, 31 Republican and two independent. Independents remain separate from Democratic party totals, despite their current caucus alignment. California’s nine same-party general ballots contribute eight Democratic and one Republican seat, conditional on the certified slate remaining in force. They are not safe ratings converted to certainty. Unusual ballot formats, independent candidates, incomplete candidate slates and unresolved district boundaries can prevent modeling. Provisional aggregator candidate lists are not certified ballots everywhere.

Each modeled race retains its fitted Student-t marginal distribution. A Gaussian copula introduces common errors without adding variance a second time. For race i in state s :

$$Z_i = \sqrt{a} G + \sqrt{b} H_s + \sqrt{1 - a - b} E_i$$

$$M_i = \mu_i + \text{shift} + \text{scale}_i * T_{df_i}^{-1}(\Phi(Z_i))$$

G , each H_s and each E_i are independent standard normal draws; Φ is the standard normal cumulative distribution and $T_{df_i}^{-1}$ is the Student-t quantile. National and state latent variance shares (a, b) are (0,0), (0.20,0.10) or (0.40,0.15). Thus same-state latent correlation is $a+b$ and different-state latent correlation is a . These are assumptions, not fitted electoral correlations; nonlinear transformation need not preserve those Pearson correlations in margins. The common shock is drawn separately for each chamber run: this is not a validated joint House-Senate control model.

The model uses 8,192 simulation draws per displayed setting, seed 20261004, and margin shifts of -4 , 0 or $+4$ percentage points. A shift changes Democratic-minus-Republican margin, not either party's vote share by four points. The version does not blend national polling into these chamber marginals. AI judgments, betting prices and outside ratings do not set them.

Count positive simulated margins as Democratic wins among modeled races, with the complement Republican. Add continuing and certified-ballot seats. Unresolved races do not get invented probabilities. Instead, each party's expected-total bounds range from its known base plus expected modeled wins to that quantity plus every unresolved seat. A wider envelope adds the same unresolved allocation bounds to the modeled 10th and 90th percentiles. These are separate per-party bounds: their endpoints do not form one joint allocation. The envelope is not a calibrated 80% interval for the whole chamber. No majority or chamber-control probabilities are published; Senate ties are not automatically allocated.

C.5 Historical seat-count diagnostic: a different sample

The chamber check reuses earlier-cycle fits for 2012–2024, selecting eligible 28-day polling predictions where available and background-only predictions otherwise. It excludes unsupported ballots and counts eligible subsets, not complete chambers. House subsets contain 326–381 seats; Senate subsets contain 25–31. This diagnostic differs from the original six-cycle matched-poll study in years, eligible races, prediction availability and outcome unit. Its results cannot extend Tables 1–5 by pooling observations.

Chamber	Error assumption	Years covered by the range	Mean absolute seat-count error
House	independent	1 / 7	10.36
House	shared	7 / 7	10.34
House	stronger shared	7 / 7	10.34
Senate	independent	6 / 7	1.54
Senate	shared	6 / 7	1.53
Senate	stronger shared	7 / 7	1.52

Shared shocks mostly change uncertainty, not expected seat counts. Seven election years provide little evidence about tail calibration. The settings were not fitted or chosen by optimizing these results, and apparent subset coverage does not validate current full-chamber envelopes. Independent errors are particularly implausible as a complete account of national election uncertainty. This is a diagnostic reason to retain dependence sensitivity, not proof of any one dependence setting.

C.6 Prospective rules, data gaps and publication

Registered comparisons use deadlines at 14:00 UTC on October 6, 20 and 27, 2026. Select the latest eligible run within the preceding 24 hours, including failures and abstentions; do not rescue a failed latest run using an older successful forecast or backfill after a deadline. Race-margin models require matching certified margins; chamber scoring needs complete attested party outcomes. The chamber, Tide and weighted-poll registrations are separate from the original probabilistic comparison. Review and version policy changes before adoption.

The current operational schedule collects races and runs forecasts twice daily, at 07:45 and 19:45 New York time. AI election research runs daily at 08:15; outside comparison checks run at 08:30; source/model exception reminders are checked daily at 08:45. A scheduled attempt is not a guarantee of a successful source retrieval or model response. Model/source changes require review, not silent automatic tuning.

Outside ratings cover the selected watchlist, not all 535 seats with matched numerical predictions. Poll averages and model-projected final margins have different targets. RealClearPolling automatic access remains blocked; its dated snapshots need review. Missing comparisons are not agreement. Fixed issue-level Google Trends measurements remain unavailable; the trending-topics feed is not an adequate substitute and is excluded from forecasting. No empirical benefit from search attention has been demonstrated.

The website distinguishes the whole-chamber experiment from race forecasts. Its state map shows evidence coverage, not a statewide House winner, and its tile grid preserves unresolved, continuing and same-party-ballot categories. Forecasts are withheld when publication checks fail. The public working paper and aggregate results do not redistribute private original report text or constitute a complete replication archive.

C.7 Contribution and next evaluation

These additions make the workflow more inspectable and provide competing explanations; they do not make the estimator novel. Uniform swing, shrinkage, recency weighting, copulas and shared election errors are established ideas. The defensible contribution remains disciplined comparison under sparse and uneven evidence, including negative findings and explicit abstentions. A stronger research contribution requires prospective results, credible information timestamps, stronger external benchmarks and independent review. No claim of journal readiness or outperformance is supported by the present extensions.

Reproduction pointers: `src/balance_of_power/tide.py`, `weighted_polls.py`, `chamber.py` and `chamber_sources.py` in the runtime repository. The October 4 handoffs describe source decisions, receipts, policy versions and audit questions. The original historical tables and figures remain unchanged. Codex implemented and wrote this revision under Lampbotics stewardship. Claude Code audited earlier work and has since reproduced the v2 chamber calculations; Appendix D records defects and repairs. The v3 repairs await follow-up review.

C.8 Earlier editorial interpretation — now under review

Audit status: The following opinion was made before the chamber audit, is retained for accountability, and is no longer presented as a current call. The input corrections and remaining bias require reassessment.

At the owner's request, the October 4 public page originally published Codex's low-confidence judgment: a narrow Democratic House majority and a Republican Senate majority. This is a provisional interpretation, not a calibrated chamber-control probability, exact seat forecast or registered evaluation arm. Its evidence record is `EDITORIAL_CHAMBER_CALL.json`.

One disclosed illustrative starting point assigns unresolved seats to their listed incumbent party: 13 Democratic and eight Republican House seats, and one Democratic and seven Republican Senate seats. Added to no-shift modeled averages and fixed/continuing seats, this gives about 219.5 D / 215.5 R in the House and 46.2 D / 51.8 R / 2 I in the Senate. This assumption is not adopted by the statistical model. Roster labels need not match current candidate affiliation; California 6 specifically includes an independent candidate. The illustration cannot settle an unsupported race or justify treating it as certain. The published numerical ranges continue to leave these seats unresolved.

The Senate judgment is contestable: the [270toWin Senate consensus dated October 2, checked October 4](#) leaves five toss-ups, and its prediction-market comparison points toward Democrats. Such disagreement is preserved rather than described as confirmation. A small shared shift can reverse the narrow House interpretation, while additional Republican losses in unresolved Senate races can reverse the Senate call. The opinion is dated and manually revisable; it must not be passed off as a daily recalculation, validated model output or evidence of scientific contribution. Preserve earlier calls if revised.

Appendix D. Chamber audit and version 3 repair, 4 October 2026

Claude Code reproduced all 18 v2 chamber scenarios and all 42 historical-check rows. This verifies calculation, not predictive validity. The audit identified live prior-service under-credit, candidate-name mismatches that prevent polling use, inconsistent Ohio Senate eligibility, and unreported directional errors. The detailed report accompanies the public paper downloads. These findings qualify, rather than overturn, the original study's conclusion that complexity has not established superiority.

D.1 Identity correction without a new fitted coefficient

Version `bop-chamber-scenarios-v3` uses an explicit, seat-scoped, hash-pinned identity review for 2026 live inputs. Candidate spellings in the retained roster are linked to the federal-results archive; the existing prior-winner function then applies its unchanged eight-year, same-state eligibility conditions. This is an input repair, not a new coefficient or a blanket award of experience to incumbents. Historical inputs, fitted models and earlier predictions remain frozen. All 42 earlier count-check values reproduce unchanged, with signed-error and interval-width fields added. This confirms preservation of earlier results; it is not a new historical validation of broader name matching.

First-initial/surname matching is not used. It would confuse Trevor Nehls with Troy Nehls and can conflate Shontel Brown with Sherrod Brown. Explicit aliases apply to challengers as well as incumbents: both Jen Kiggans and Elaine Luria have qualifying earlier federal wins. Thus VA-02 does not flip simply by repairing the incumbent's name; both candidates must receive the same treatment. CA-22 and PA-01 do change to Republican-leading central margins in the repaired run. Remaining listed incumbents without archive matches are reported as unknown, not treated as proof of no experience.

Case-insensitive comparisons and reviewed aliases reconcile the CA-48, FL-27 and NM-02 watchlist names. A saved poll record that fails identity matching is now flagged in the run and public view instead of silently looking like no polling exists. Freshness and fieldwork windows still apply. Source-read passages and inferred demographics are not substituted for poll values.

These aliases are deliberately scoped to the new chamber policy. They do not silently rewrite the registered Compass, Tide or weighted-poll arms. Their broader live feature coverage requires a separately versioned review before changing those comparisons.

D.2 Coverage is narrowed where official evidence conflicts

Ohio Senate is withheld consistently with Compass pending tested special-election provenance. Official [Louisiana election guidance](#), read October 4, distinguishes Senate closed-party primaries from the 2026 House's November open primary and possible December runoff. All Louisiana House seats are therefore unresolved in v3; the old two-party general-election approximation cannot simply be assumed valid. Louisiana Senate remains withheld pending a reviewed extension.

The official [Massachusetts candidate list](#) includes non-major-party candidates absent from the retained aggregator slates for MA-03, MA-08 and MA-SEN. Those seats are now unresolved. It also lists single candidates in MA-02/05/07, but absent verification of write-in eligibility this repair does not turn them into guaranteed party seats. Source flags for unopposed candidates in Florida, Massachusetts and Wisconsin are preserved instead of discarded; official resolution of those contests remains incomplete.

V3 accounting is 398 modeled House seats, nine certified same-party ballots and 28 unresolved seats; Senate accounting is 25 modeled, ten unresolved and 65 continuing. In the first repaired run, 15 House and four Senate races have matched recent local

polls. These are dated run counts, not guarantees of future coverage. The new no-shift expected bounds are D203.74–231.74 and R203.26–231.26 in the House; D43.92–53.92 and R44.08–54.08 in the Senate, plus two continuing independents and unresolved other-party possibilities. They remain separate conditional bounds, not control probabilities.

D.3 Directional error and interval width

The shared-error House check overpredicted Democratic wins in six of seven eligible-subset elections. Mean signed error is about +4.9 seats, including a roughly –19.2-seat miss in the 2018 Democratic wave. Senate signed error averages about +0.9 seats. Reported range coverage alone concealed this pattern. The shared House ranges average about 36 seats wide, compared with about ten under independent errors. Coverage improvement therefore cannot be read as improved centering.

Chamber	Year	Eligible seats	Actual D wins	Expected D wins	Signed error	Range width
House	2012	326	143	152.5	+9.5	50
House	2014	337	142	156.8	+14.8	35
House	2016	327	146	154.7	+8.7	33
House	2018	350	183	163.8	-19.2	38
House	2020	381	194	206.1	+12.1	38
House	2022	372	191	198.4	+7.4	34
House	2024	341	167	167.7	+0.7	24
Senate	2012	31	23	20.7	-2.3	7
Senate	2014	27	11	13.6	+2.6	3
Senate	2016	29	11	12.3	+1.3	5
Senate	2018	29	19	19.7	+0.7	5
Senate	2020	26	12	13.7	+1.7	5
Senate	2022	25	10	10.6	+0.6	5
Senate	2024	28	15	16.5	+1.5	5

These retrospective findings motivate a separately tested national-environment term, preferably where local polls are absent, with fitting restricted to earlier cycles. They do not justify subtracting the observed average error from 2026 forecasts, choosing dependence settings by the best observed coverage, or relaxing Tide’s three-firm threshold. No new national adjustment was activated in this repair.

D.4 Accountability and remaining work

The earlier editorial House/Senate call in Appendix C is marked under review on the website. It remains in the record rather than being silently rewritten. A renewed judgment should consider the repaired features, unresolved ballots and directional bias; the present data do not justify a precise control probability. Certified-ballot and geometry checks, independent audit of identity links, and a separately registered national-environment comparison remain priorities. Code tests and a successful restricted-service run address implementation, not these scientific limitations.