Accurate Local-Level Estimation of U.S. Public Opinion:
Methodology and Out-of-Sample Validation of MOSAIC, a Synthetic MRP Model Calibrated to Various Political and Census Demographic Targets

G. Elliott Morris
Strength In Numbers

July 23, 2026

Abstract

Before trusting a model that claims to measure public opinion in every state, congressional district, and community, one should ask whether its geographic engine can recover something true and known. We report an out-of-sample election backtest of MOSAIC (**M**ultilevel **O**pinion **S**moothing **A**nd **I**nference across **C**ommunities), a small-area estimation system for survey data. MOSAIC post-stratifies survey-based multilevel models against a *synthetic* population frame — a Census-derived joint distribution of demographics, geography, turnout, past vote, and party identification, assembled from public administrative and probability-sample data and calibrated to multiple political targets: certified election returns, Pew's validated-voter demographic benchmarks, and party benchmarks from the National Public Opinion Reference Survey (NPORS). Including party identification as a dimension of the frame, measured for voters and non-voters alike, allows the model to represent the political composition of the non-voting and independent population that election calibration alone cannot reach. To test the engine, we rebuild the frame-construction and vote-reconstruction pipeline as it would have stood before the November 2024 election — trained on data available at the time and calibrated to the 2020 presidential result — and compare its predictions to the certified 2024 returns, which are held out of the model. The reconstruction recovers the presidential result to a state-level root-mean-square error (RMSE) of 1.60 percentage points (pp) across all 50 states, calling all 50 state winners correctly (Pearson $r = 0.99$); U.S. House district results to an RMSE of 3.4 pp across 393 contested districts, calling 386 correctly ($r = 0.98$); and the U.S. Senate result to an RMSE of 2.0 pp across 33 races, calling 32 correctly ($r = 0.98$). State-level RMSE is lower in competitive races, while district-level RMSE is slightly higher. A second held-out cycle, the 2022 midterm, replicates the exercise on a structurally different election: the same 2020-calibrated engine recovers the U.S. Senate to 3.8 pp (33 of 33 winners) and the U.S. House to 3.9 pp (304 of 320 contested districts, after excluding six states redistricted between 2022 and 2024). We document the pipeline under test, present the results with interval-calibration diagnostics, and state the limitations plainly. We then develop a dynamic MRP model on top of this calibrated benchmark to track opinion at a fine geographic level monthly, using contemporaneous survey data from the polling firm Verasight. This approach is validated against 230 independent state polls ($r = 0.88$; poll-level RMSE 3.9 pp) measuring President Donald Trump's approval rating in 2025 and 2026.

1 Introduction

Most published opinion data describe the national mood. Yet the questions that matter to journalists, researchers, advocates, and public-affairs professionals are usually local: How does a particular congressional district view a bill? How has approval changed in a specific state? Where does opinion on an issue

differ from the national average? Answering those questions from a national survey of roughly 1,500 respondents is not a matter of filtering to a subgroup — the per-area samples are far too small — but of modeling.

The standard tool is multilevel regression and post-stratification (MRP), introduced for opinion estimation by Gelman and Little (1997) and Park, Gelman, and Bafumi (2004), shown to dominate simple disaggregation for state-level opinion by Lax and Phillips (2009), and refined for small subgroups by Ghitza and Gelman (2013). MRP fits a multilevel model relating the outcome to demographic and geographic predictors, then re-weights the model's cell-level predictions to a population frame that describes how many people occupy each demographic-geographic cell. Its accuracy depends on two things that receive uneven attention in practice: the richness of the model, and the fidelity of the frame.

The frame is the more difficult component to validate. A modeling error shows up as a poor fit; a frame error — the wrong picture of who lives where, who votes, and how — silently shifts every downstream estimate while leaving fit diagnostics untouched. The strongest available test of a frame is an election it did not see: rebuild the frame from information available before the vote, predict the result at fine geographic resolution, and compare against the certified count. This report is that test, applied to MOSAIC (Multilevel Opinion Smoothing And Inference across Communities), the small-area estimation system behind Strength In Numbers Pro, for the 2024 U.S. elections for President, House, and Senate. The report then documents the dynamic opinion-tracking model built on the validated frame and validates its estimates against independent state polling.

The exercise reflects three commitments. First, **transparency**: the pipeline is documented end to end. Second, **validation**: the system is tested against a held-out election, and the results — including where it is weakest — are reported. Third, **nonpartisanship**: the pipeline is calibrated to certified election returns and public benchmarks, not to any political objective, and is maintained independently of any campaign or party.

Section 2 describes the pipeline under test, Section 3 the backtest design, and Section 4 the backtest results, including interval-calibration diagnostics and a second held-out cycle (the 2022 midterm). Section 5 specifies the dynamic opinion layer and validates it against independent state polling. Section 6 states limitations, and Section 7 discusses what the two validations jointly suggest.

2 The pipeline under test

2.1 Estimands and definitions

The post-stratification frame is restricted to **citizen adults**, so every post-stratified estimate in this report targets the U.S. citizen-adult population (or, for vote quantities, the citizen-adult electorate); the underlying survey universe is U.S. adults aged 18 and over, and the implications of that mismatch are discussed in Section 6. **Registered-voter** estimates re-weight frame cells by each cell's modeled registration probability. **Two-party share** denotes the Democratic share of the Democratic-plus-Republican vote (presidential and House results). **Two-way share** denotes the analogous two-category share where one side aggregates a non-major-party candidate or response: for the Senate, the Democratic or Democratic-aligned candidate's share of the two-way vote (independents in Vermont, Maine, and Nebraska are counted on the Democratic side); for approval, approve ÷ (approve + disapprove). A race is **competitive** when the actual two-party (or two-way) share fell between 42.5% and 57.5%. A **winner call** assigns each race to the side of 50% on which the posterior median falls.

2.2 Data

Construction of the post-stratification frame draws on public administrative and survey sources: the American Community Survey (ACS) 5-year Public Use Microdata Sample for the demographic cell structure (U.S. Census Bureau, 2024); harmonized 2020 precinct returns from the Voting and Election Science Team (VEST, 2021) for PUMA-level presidential vote, benchmarked against MIT Election Data and Science Lab (2021) county totals; the Current Population Survey (CPS) Voting and Registration Supplement and the Cooperative Election Study (CES) Common Content (Schaffner & Ansolabehere, 2025) — the latter supplying validated registration, turnout, and vote choice via TargetSmart's voter-file match — for the turnout and vote-choice reconstruction; Pew Research Center's National Public Opinion Reference Survey (NPORS, 2020–2025) and 2023–24 Religious Landscape Study — address-recruited probability samples that measure party identification for the full adult population, voters and non-voters alike — for the frame's party dimension; and Pew's validated-voter benchmarks for demographic calibration. The backtest targets are the certified 2024 returns: state-level presidential results, district-level U.S. House results for the district map in force in 2024 (the 118th-Congress boundaries), and state-level U.S. Senate results.

2.3 The post-stratification frame

Cell structure. The frame is built from the ACS 5-year PUMS, restricted to citizen adults, with one cell per unique combination of state, PUMA, sex, age (four categories), race (five categories), education (six levels), and income (four brackets). The cell population count is the sum of ACS person-weights. Missing demographics are imputed using predictive mean matching implemented in the `mice` package (van Buuren & Groothuis-Oudshoorn, 2011) before collapsing.

Vote at the PUMA level. VEST harmonized precinct polygons are reduced to planar centroids and spatially joined to 2020 Census PUMA boundaries; presidential votes are summed within each (state, PUMA) pair to produce candidate totals and a two-party Democratic share. The roll-up is benchmarked against county totals to detect alignment errors.

Turnout and vote reconstruction. Onto the CES+CPS file — harmonized to common demographic schemas and stacked, with missing CPS validated outcomes imputed by random-forest `mice` — a sequence of four weighted Bayesian multilevel models (`brms`/`cmdstanr`; Bürkner 2017, Gabry et al. 2024, Stan Development Team 2024) is fit: (A) validated registration; (B) validated turnout given registration; (C) major-party versus other vote; and (D) Democratic versus Republican vote given a major-party vote. Each stage is a multilevel logistic regression with an extensive set of interactions following Ghitza and Gelman (2013); stage (D), for example, models the probability that validated voter i chose the Democratic candidate, given a major-party vote, as

$$\Pr(y_i = 1) \;=\; \text{logit}^{-1}\Big(\beta_0 + \mathbf{x}_{p[i]}^{\top}\gamma + \alpha^{\text{race}}_{r[i]} + \alpha^{\text{age}}_{a[i]} + \alpha^{\text{edu}}_{e[i]} + \alpha^{\text{inc}}_{k[i]} + \alpha^{\text{sex}}_{s[i]} + \textstyle\sum_{m\in\mathscr{M}} \alpha^{(m)}_{j_m[i]} + \delta_{\text{region}[i]} + \delta_{\text{state}[i]}\Big),$$

where $r[i], a[i], e[i], k[i], s[i]$ index respondent i's race, age, education, income, and sex; each varying-intercept family is partially pooled, $\alpha^{(m)}_j \sim \mathcal{N}(0, \sigma^2_m)$ with the scale σ_m estimated from the data; $\mathscr{M}$ collects the two- and three-way demographic interaction families; the δ terms form the nested region/state hierarchy; and $\mathbf{x}_{p[i]}$ holds the PUMA-level contextual covariates (partisanship, demographic shares, rural–urban continuum, and, for vote choice, a battery of civic and religious covariates). The CES enters this

reconstruction through its native validated-vote survey weights. Because the CES supplies validated behavior through the voter-file match, the reconstruction trains on validated rather than self-reported registration, turnout, and vote wherever possible.

Weighting the survey inputs. The frame reconstruction above weights the CES by its native validated-vote survey weights. The vote-choice models fit for the election backtest instead carry an inverse-probability (IPW) pseudo-weight that reaches the within-cell political skew a demographic weight cannot, following the synthetic "reference-sample" approach of Mercer et al. (2017). The survey is stacked with a large probability-anchor sample drawn with replacement, population-weighted, from the calibrated frame — each anchor row carrying a party identification stochastically drawn from that cell's imputed $P(\text{D/R/I})$ — and a model for source membership (survey versus anchor) is fit on demographics, region, a battery of PUMA-level contextual covariates, state presidential lean, and leaned party identification, with no vote outcome among the predictors. A cross-validated elastic net (`glmnet`) is used for the membership model in preference to a random forest, which over-fit the panel-versus-frame boundary on the continuous covariates and inflated the design effect in backtesting (state-level presidential RMSE 2.0 vs. 2.5 pp). Each respondent's out-of-fold membership probability $\hat{\pi}_i$ yields a pseudo-weight $(1 - \hat{\pi}_i)/\hat{\pi}_i$, trimmed at $\hat{\pi}_i \in [0.02, 0.98]$ and normalized to mean one, entering the model's weighted pseudo-likelihood. The production opinion models instead run on Verasight's own survey weights (Section 5.1).

Calibration to certified results. The reconstructed cell probabilities are calibrated to official election totals via logit shifts. The workhorse operation, analyzed by Rosenman, McCartan, and Olivella (2023), adjusts a set of cell probabilities p_c to hit an aggregate target P^* by a uniform shift on the log-odds scale: find the scalar δ solving

$$\frac{\sum_c n_c \,\text{logit}^{-1}(\text{logit}(p_c) + \delta)}{\sum_c n_c} \;=\; P^*, \qquad p_c^{\text{cal}} \;=\; \text{logit}^{-1}(\text{logit}(p_c) + \delta),$$

which matches the target exactly while preserving the cell-level ordering and heterogeneity of the model's predictions. Calibration proceeds in three stages: (a) a national and then state-specific turnout shift so each state's cell-weighted turnout equals its certified total votes; (b) a nested within-state shift so each state's two-party-and-other vote shares match the certified margin exactly; and (c) a *multi-way demographic calibration* — a constrained `optim()` solver over 11 logit deltas $\delta = (\delta^{\text{race}}_{1..5}, \delta^{\text{age}}_{1..4}, \delta^{\text{sex}}_{1..2})$ chosen to minimize

$$\hat{\delta} \;=\; \arg\min_{\delta} \sum_{b \in \mathcal{B}} \left(m_b(\delta) - m_b^{\text{Pew}}\right)^2 \quad \text{subject to the stage-(b) state vote constraints,}$$

where $\mathcal{B}$ indexes the benchmark moments (Democratic share of validated voters by race, age, sex, sex × race, and sex × age), $m_b(\delta)$ is the frame-implied moment after applying the deltas, and the hard state constraints are re-imposed inside the objective on every evaluation, so demographic calibration can never move a state off its certified result. This use of the logit shift to reconcile survey-based small-area estimates with known election results and demographic benchmarks simultaneously follows the survey-calibration approach of Kuriwaki, Ansolabehere, Dagonel, and Yamauchi (2024). In production, the calibration target is the 2024 result; in the backtest it is the 2020 result (Section 3). The output is a long-format frame with one row per (state × PUMA × sex × age × race × education × income × past vote) cell and a calibrated population count.

Party identification in the frame. A distinguishing feature of the frame is that three-category party identification (Democrat / Republican / Independent, with leaners assigned to the party toward which they lean) is imputed onto every cell, making party a post-stratification *axis* rather than a weighting lever. The imputation is trained exclusively on stacked address-recruited probability samples — Pew's NPORS (2020–2025) and the 2023–24 Religious Landscape Study, roughly 26,000 complete cases — never on the survey being modeled or backtested, and it is built in three stages: a demographic base, a geographic logit shift, and a past-vote-conditional multilevel model, pinned to the party benchmark by iterative proportional fitting. A random forest (`ranger`; Wright & Ziegler, 2017) first supplies a demographic base $P(\text{party} \mid \text{demographics})$ from the full stacked donor. Geographic variation is then introduced through empirical logit shifts on each PUMA's recent presidential Democratic two-party share d_p (the 2024 result in production, the 2020 result in the firewalled backtest), in the recalibration form of Rosenman, McCartan, and Olivella (2023), with slopes estimated from CES microdata against the demographic base as an offset: cell c in PUMA $p(c)$ receives

$$\text{shift}_c^{DR} = \beta_{\text{dem}} \cdot \text{logit}(d_{p(c)}), \qquad \text{shift}_c^{I} = \beta_{\text{ind}}^{g} \cdot \text{logit}(d_{p(c)}) + \beta_{\text{ind}}^{s} \cdot |d_{p(c)} - 0.5|,$$

applied on the logit scale to the Democratic-versus-Republican contrast and to the Independent share respectively, each population-weighted demeaned so the shifts carry geographic variation but no national-level signal.

The operative model conditions the imputation on *past vote* — the axis the frame is already long over — through a pair of partial-pooled multilevel logistic regressions fit in `lme4` (Bates et al., 2015), one for the Independent share and one for the Democratic share among major-party identifiers:

$$\text{logit Pr}(y_i) = \beta_0 + \beta_{s[i]}^{\text{sex}} + \beta_{v[i]}^{\text{vote}} + \alpha_{a[i]}^{\text{age}} + \alpha_{r[i]}^{\text{race}} + \alpha_{e[i]}^{\text{edu}} + \alpha_{k[i]}^{\text{inc}} + \alpha_{g[i]}^{\text{region}} + \alpha_{v[i],r[i]}^{\text{vote×race}} + \alpha_{r[i],g[i]}^{\text{race×region}},$$

with sex and past vote entered as fixed effects and every remaining term partial-pooled ($\alpha \sim \mathcal{N}(0, \sigma^2)$, scales estimated from the data). The *vote* × *race* interaction — party's dependence on past vote differs sharply by race — and the *race* × *region* interaction — the geography of party-by-race — are the terms that earn their place: this multilevel imputer improves held-out subgroup calibration and roughly halves competitive-state bias in the backtest relative to a joint random forest over the same predictors. The PUMA-level d_p shift above is then layered on top of this multilevel base on the logit scale, so geography enters twice by design — coarsely through the pooled region effect, finely through the d_p shift.

Two iterative-proportional-fitting (IPF) steps pin the result to the party benchmark. A national rake first sets the imputed D/R/I marginal to a blended NPORS reference; then a two-constraint IPF on the (cell × past-vote) × party table places party mass onto the vote-calibrated frame while preserving the election-anchored geography exactly — state-by-vote, district-by-vote, and state-by-race totals are untouched — with the national party marginal matched to the benchmark. The party benchmark therefore enters the system exactly once, as the calibrated marginal of the frame, where the outcome models can partially pool away from it — a buffered post-stratification input rather than a hard rake on the likelihood. The result is a synthetic electorate: a joint distribution of demographics, geography, turnout, past vote, and party, calibrated to certified vote totals, validated-voter demographics, and probability-sample party benchmarks simultaneously.

Why party belongs in the frame. Calibrating to certified returns pins down the politics of *voters*, but voters are only part of the population: roughly a third of adults did not vote, and demographic-geographic post-stratification says nothing about their politics beyond what demographics imply. The failure mode is concrete: matching a survey to the population on demographics and geography cannot correct *within-cell*

political skew — in the CES, a residual Democratic lean of roughly 2.4 points survives full demographic-geographic alignment. A probability-sample benchmark measured on the full adult population is the missing piece: NPORS supplies what party identification looks like inside each demographic cell, so the frame's non-voter and independent mass is politically balanced rather than inherited from the realized survey respondent pool. Ablations on the backtest support the design. Assigning leaners to the party toward which they lean — rather than carrying a 39% "pure Independent" group whose party signal is weak — is what lets state-level party variation translate into vote spread (Alaska's presidential error falls from +7.5 pp to +1.6 pp when leaners are allocated to a major party), and replacing a joint random forest with the partial-pooled multilevel imputer roughly halves competitive-state bias (+1.45 to +0.74 pp).

2.4 Producing small-area estimates

Post-stratification preserves uncertainty by aggregating *within* each posterior draw before summarizing. For posterior draw s and geography g, the estimate is the population-weighted mean over the frame cells belonging to g,

$$\hat{\theta}_g^{(s)} \;=\; \frac{\sum_{c \in g} N_c \, w_c \, p_c^{(s)}}{\sum_{c \in g} N_c \, w_c},$$

where N_c is the calibrated cell count, $p_c^{(s)}$ the cell's predicted probability under draw s, and w_c an optional propensity weight (the modeled registration or turnout probability, set to 1 for all-adult estimates). For vote estimates the turnout probability is the operative weight; Appendix A backtests this past-vote-calibrated turnout weight against an intent-based likely-voter alternative and finds the past-vote weight the more accurate choice. The point estimate is the median of $\hat{\theta}_g^{(s)}$ across draws and the interval its central quantiles; aggregating within each draw before taking quantiles means the interval reflects covariance between cells rather than assuming independence.

Congressional districts are not fit directly. Because PUMAs predate current district boundaries, a single PUMA can fall across multiple districts; district estimates are therefore aggregated from PUMA cells through a **population-weighted PUMA→CD crosswalk**: every 2020 Census block is assigned to its (PUMA, district) pair by an internal point — 2020 blocks nest cleanly into both 2020 PUMAs and the district maps — and block populations are summed, so each PUMA contributes to each overlapping district in proportion to the share of its 2020 population living there. Population weighting matters because population is far from uniform over land; an area-based split misallocates population across PUMA–district intersections wherever a PUMA straddles an urban boundary. The production 119th-Congress crosswalk is validated against The Downballot's hand-curated presidential retabulation (correlation > 0.99 on the 2020 two-party share); the backtest uses the same construction for the district map in force in 2024.

3 Backtest design

We rebuild the frame-construction and vote-reconstruction pipeline (Section 2.3) using data available before the November 2024 election and calibrated to the 2020 presidential result rather than to 2024. We then use it to predict the 2024 presidential result at the state level, the 2024 U.S. House result at the district level, and the 2024 U.S. Senate result at the state level, and compare those predictions to the certified outcome, which is not used to fit or calibrate the reconstruction.

Two qualifications keep this claim honest. First, several specification choices — the leaned party coding and the choice of party-imputation model — were adopted after comparing candidate configurations on this same backtest; the certified outcomes never enter fitting or calibration, but the exercise is best read

as retrospective validation rather than a pre-registered out-of-sample test, and the 2026 cycle offers a genuinely untouched confirmation. Second, the party-imputation donor pool stacks NPORS waves through 2025 for its demographic base; the vote-conditional donor is restricted to the pre-election 2021 wave by design, but a strictly pre-November-2024 reconstruction would restrict the demographic donor pool as well. We treat the demographic structure of party identification as slowly varying and catalogue this as a deviation from a strict cutoff.

Results are evaluated on RMSE, mean signed error (bias), Pearson correlation with the certified result, and winners called correctly (posterior median relative to 50%; Section 2.1), reported for all races and separately for competitive races. All error metrics weight geographic units equally. All figures are reported for the production configuration: an optional nonresponse-bias sensitivity adjustment (Andridge et al., 2019) is disabled, because on the 2024 target it shifts estimates by hundredths of a point and changes no state and no material number of districts.

4 Backtest results

4.1 Geographic resolution by office

Each office is estimated at the resolution of its *estimand*. The House result is a set of district outcomes, so it is modeled at the congressional-district level, aggregating post-stratification cells to districts. The presidential and Senate results are sets of *statewide* outcomes, so they are modeled at the state level, with state-level predictors only and the frame collapsed to (demographic × state × vote × party) cells.

Reducing the presidential and Senate models to state geography is not a shortcut but the correct choice, for two reasons that a controlled ablation confirms. First, a statewide total is an average over the state; where within a state its support sits is orthogonal to that average, so sub-state variation cannot improve a statewide estimate. Second, in a frame already conditioned on calibrated individual past vote and party (Section 2.3), residence is a redundant encoding of partisanship — "which district you live in" is largely a proxy for "how you voted," which the model already observes directly and at higher fidelity. Toggling only geography on a shared model backbone bears this out: adding full sub-state (PUMA) geography changes state-level RMSE by 0.01 pp for President (1.73 vs 1.72) and 0.02 pp for Senate (2.11 vs 2.13), and the size of that change is uncorrelated ($r = 0.01$) with how internally heterogeneous a state's districts are — the opposite of what one would see if sub-state detail were correcting geographic sampling imbalance. The heavy lifting is done by the frame's vote and party axes; the geographic covariate is left with nothing to add. (Strip past vote and party from the frame and geography would matter again, as it does in a demographics-only MRP — but that is not this design.) The House keeps its district resolution because there the district *is* the estimand.

4.2 Presidential (state-level, Harris two-party share)

The presidential contest is a set of statewide races, so it is estimated at the state level, with state-level predictors only (no sub-state covariates); Section 4.1 gives the reasoning. Across all 50 states, the reconstruction achieves an RMSE of **1.60 pp** (state-resampled bootstrap 95% interval 1.3–1.9) and a mean signed error of +0.66 pp, with a Pearson correlation of **0.99** against the certified two-party result and **all 50** state winners called correctly (Table 1, Figure 1). The median absolute state error is 1.1 pp, the 90th percentile 2.5 pp, and the maximum 4.2 pp. Across the 25 competitive states — the subset in which winner classification is most sensitive to estimation error — RMSE is **1.38 pp** with bias +0.40 pp, and all 25 are called correctly.

Table 1: Presidential backtest: state-level Harris two-party share (state-level model).

	States	RMSE	Bias	r	Winners
All states	50	1.60 pp	+0.66 pp	0.99	50 / 50
Competitive (42.5–57.5%)	25	1.38 pp	+0.40 pp	0.97	25 / 25

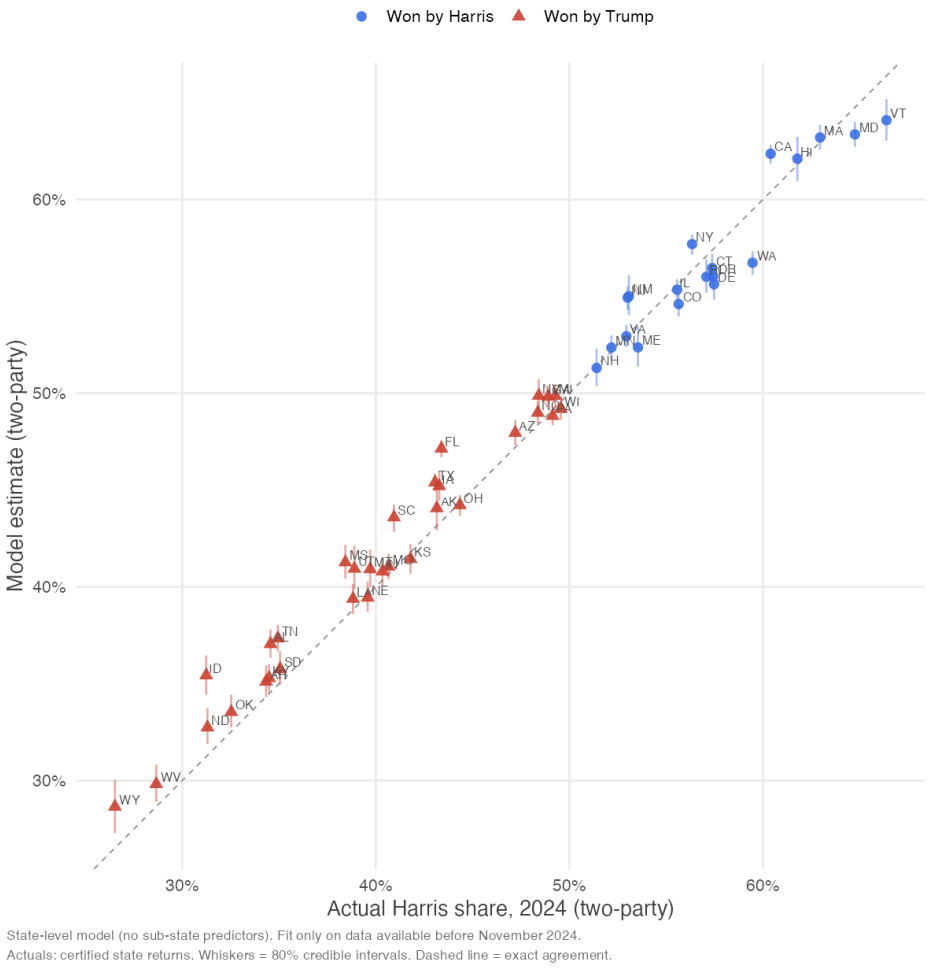

Figure 1: Presidential backtest: predicted versus actual state-level Harris two-party share, with 80% credible intervals. The dashed line marks exact agreement. Color and point shape both encode the actual winner.

4.3 U.S. House (district-level, Democratic two-party share)

Restricting to the 393 districts with a contested major-party result, the reconstruction achieves an RMSE of **3.4 pp** (state-clustered bootstrap 95% interval 2.8–3.8) and a mean signed error of +1.56 pp, with a correlation of **0.98** and **386 of 393** district winners called correctly (Table 2, Figure 2). The median absolute district error is 1.7 pp, the 90th percentile 5.5 pp, and the maximum 16.0 pp. Across the 105 competitive districts, RMSE is 3.6 pp and 98 of 105 winners are called correctly. The share estimates carry a small positive (Democratic) level shift of about a point and a half, concentrated in the seven misclassified districts, all near the 50% line.

Table 2: House backtest: district-level Democratic two-party share.

	Districts	RMSE	Bias	r	Winners
All contested	393	3.4 pp	+1.56 pp	0.98	386 / 393
Competitive (42.5–57.5%)	105	3.6 pp	+1.45 pp	0.84	98 / 105

4.4 U.S. Senate (state-level, Democratic-aligned two-way share)

Across the 33 Senate races, the reconstruction achieves an RMSE of **2.0 pp** (bootstrap 95% interval 1.6–2.4) and a mean signed error of +0.81 pp, with a correlation of **0.98** and **32 of 33** race winners called correctly (Table 3, Figure 3). The median absolute error is 1.4 pp, the 90th percentile 3.6 pp, and the maximum 3.9 pp. Across the 15 competitive races, RMSE is 2.0 pp and 14 of 15 winners are called correctly. The single miss is Pennsylvania, where Democratic incumbent Bob Casey narrowly lost; Montana, where Jon Tester also lost a close race, is called correctly (Appendix B traces the earlier Montana miss to survey weighting). Independent candidates in Vermont, Maine, and Nebraska are counted on the Democratic side of the two-way share (Section 2.1).

Table 3: Senate backtest: state-level Democratic-aligned two-way share.

	Races	RMSE	Bias	r	Winners
All races	33	2.0 pp	+0.81 pp	0.98	32 / 33
Competitive (42.5–57.5%)	15	2.0 pp	+1.04 pp	0.91	14 / 15

4.5 Interpretation

Two points bear emphasis. The district-level test is the most demanding of the three — congressional districts are numerous, some are close, and district predictions inherit both modeling and crosswalk error — and it is the closest available test of the geographic resolution at which the system operates. And the observed error pattern is generally favorable: state-level errors are lower rather than higher in competitive races (district-level RMSE is slightly higher in the competitive subset), the level shift is small, and the winner misclassifications show no partisan direction.

4.6 Interval calibration and probabilistic accuracy

Point accuracy is only half of the story; the intervals and probabilities should also be checked against outcomes. Empirical coverage of the posterior credible intervals falls short of nominal levels: the 80%

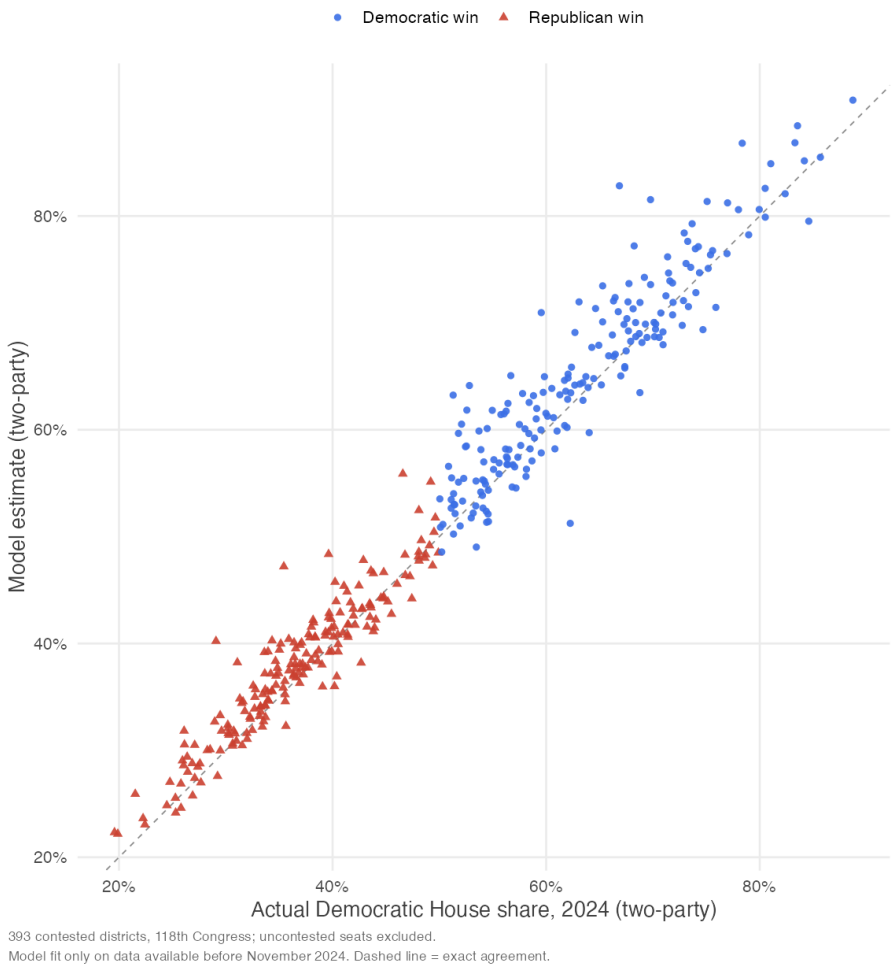

Figure 2: House backtest: predicted versus actual Democratic two-party share across 393 contested districts. The dashed line marks exact agreement. Color and point shape both encode the actual winner.

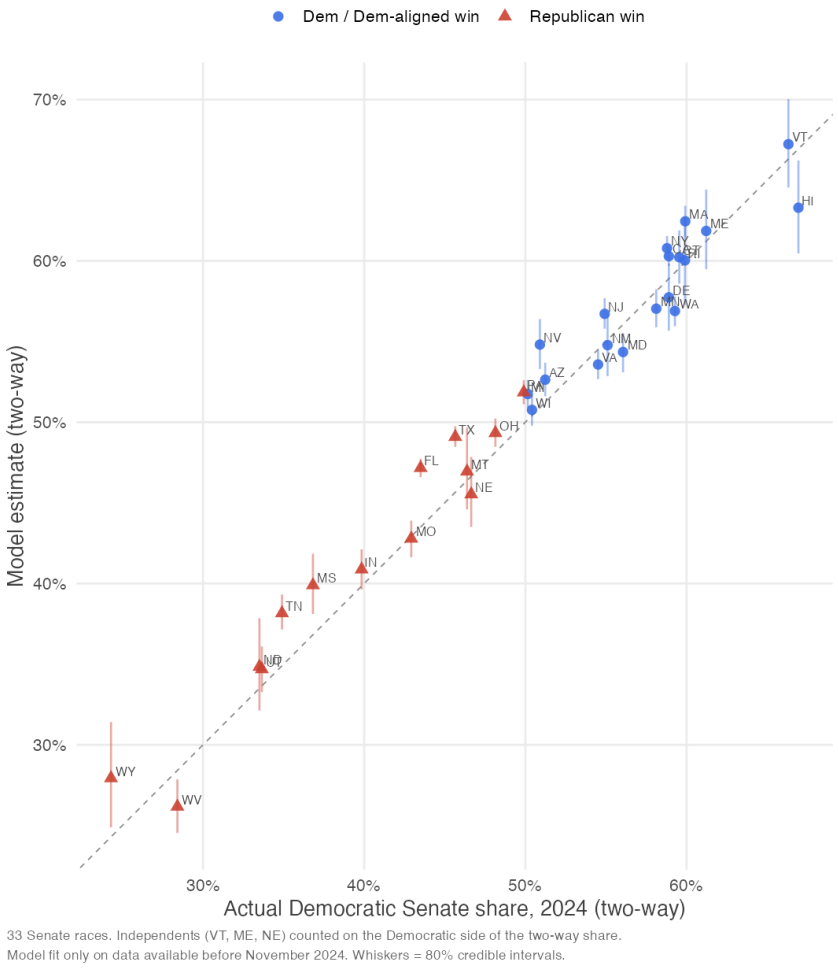

Figure 3: Senate backtest: predicted versus actual Democratic-aligned two-way share across 33 races, with 80% credible intervals. Independents in Vermont, Maine, and Nebraska are counted on the Democratic side. Color and point shape both encode the actual winner.

intervals cover the certified result in 36% of states (presidential), 64% of contested districts (House), and 42% of Senate races, and the 95% intervals cover 54%, 82%, and 61% respectively. The posterior intervals are conditional on a single estimated frame — none of the uncertainty in frame construction, imputation, calibration, or benchmark sampling is propagated through them — and the undercoverage is the visible symptom. This motivates the explicit error augmentation of Section 4.7, and it is why we describe the intervals throughout as posterior credible intervals conditional on the frame rather than as calibrated total-error intervals.

Winner probabilities fare better. Taking each race's simulated probability of a Democratic win and scoring it against the outcome, the Brier scores are 0.009 (presidential, 50 states), 0.016 (House, 393 districts), and 0.032 (Senate, 33 races) — for comparison, an uninformative 50/50 forecast scores 0.25.

4.7 Uncertainty: simulated aggregate outcomes

The point predictions above summarize posterior medians; the same posteriors support probabilistic statements about aggregate outcomes. We take 2,000 joint simulation draws per office from the vote models — each draw a complete map of state (or district) two-party shares, aggregated exactly as the point estimates are — and convert them to Electoral College votes, House seats, and Senate wins. Electoral votes are allocated winner-take-all by state (the Maine and Nebraska district splits are ignored; on a winner-take-all basis the actual 2024 count is Harris 225 rather than 226). House seat totals count the simulated contested districts plus the uncontested seats, whose winners are known, and a small number of unsimulated contested districts filled with their 2022 winner — pre-2024 information throughout. Senate control turns on the full chamber, so we add the 28 Democratic-aligned seats not up in 2024 — fixed, pre-2024 information — to the simulated wins among the 33 up-races. On the seat scale we report simulation medians: **221 Democratic House seats** (counting each district for its point-estimate winner instead gives 221) and a **median Democratic Senate caucus of 48 of 100** (20 up-race wins plus the 28 holdovers).

The posterior draws do not carry error from the NPORS party benchmark or its imputation onto the frame (Section 2.3) — one visible symptom is the interval undercoverage in Section 4.6. We therefore add that error explicitly. Every simulation receives Gaussian shocks on the two-party scale: a *common national shock* (sd 0.35 pp) representing sampling error in the pinned NPORS marginal — the standard error of the D−R margin at NPORS sample sizes is roughly 1 pp — multiplied by its pass-through to two-party vote, bounded by half the within-cell Democratic-versus-Republican conditional vote gap; an *independent state shock* (sd 0.35 pp) for the geographic imputation, calibrated so that the documented effect of swapping the imputation model (a 0.71 pp move in competitive-state bias) is a two-standard-deviation event; and, for the House, an additional *independent district shock* (sd 0.35 pp) for sub-state spread error. The national and state shocks are shared across offices within each simulation, so party error moves the presidential and Senate maps coherently.

The resulting distributions bracket the actual outcomes (Figures 4–6). The Electoral College simulation gives Harris a median of 239 electoral votes and a **7%** probability of reaching 270 (about 1% before the party-error component), against an actual 226: the model regarded a Harris Electoral College win as a clear underdog, and the party-error component matters most exactly here, in the tail, where a win requires correlated misses across several close states. The House simulation puts the median at 221 Democratic seats with an **83%** probability of a Democratic majority, against an actual 215 — the small Democratic level shift in the share estimates is amplified at the 218-seat threshold, where a two-seat move in the median swings the majority probability sharply. The Senate simulation puts the median Democratic caucus at 48 of 100 seats against an actual 47 — short of the 51 needed for control in all but a negligible share of draws, so the model correctly placed Republican Senate control as the near-certain outcome. Across all three the aggregate medians sit Democratic of the actual outcome — by thirteen electoral votes, six

House seats, and one Senate seat — consistent with the small positive level shift in the share estimates (+0.7 to +1.6 pp), while the actual result falls inside every distribution.

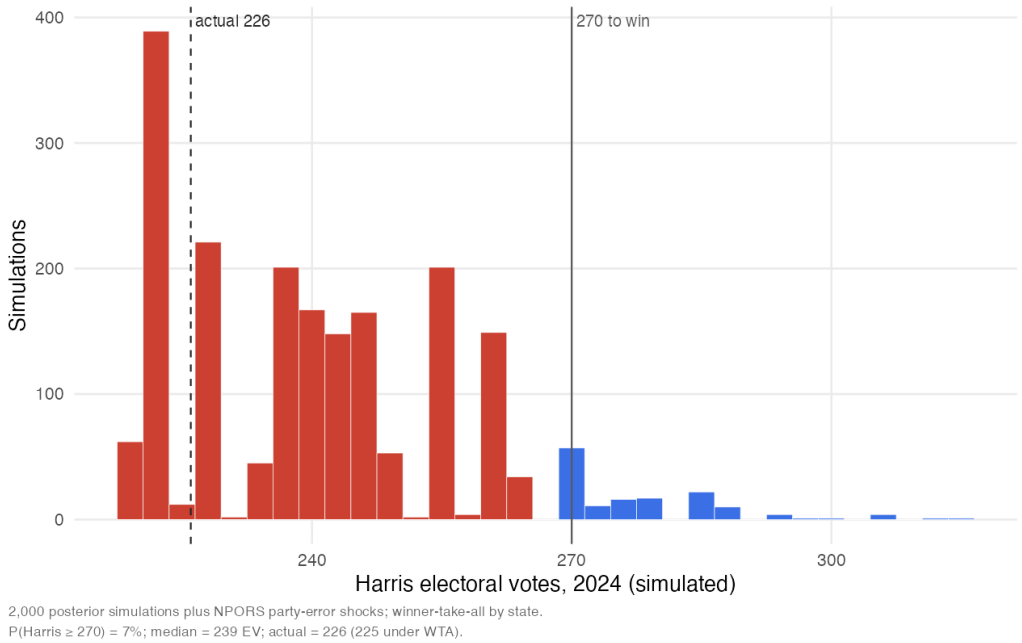

Figure 4: Simulated Harris Electoral College votes: 2,000 posterior simulations with NPORS party-error shocks, winner-take-all by state. P(Harris ≥ 270) = 7%; median 239; actual 226.

4.8 A second held-out cycle: the 2022 midterm

A single election is a single environment. To test whether the recovery generalizes, we repeat the exercise for the 2022 midterm — a materially different contest: no presidential race at the top of the ticket, the incumbent party defending rather than challenging, and a widely expected "red wave" that did not fully materialize. The protocol is identical: the frame is calibrated to the 2020 result, the 2022 returns are held out, and the House and Senate models are trained on pre-election 2022 data. There is no 2022 presidential contest, so the cycle tests the House and Senate only. For the House, six states redrew their congressional maps between the 2022 and 2024 elections (Alabama, Georgia, Louisiana, New York, North Carolina, and Ohio); because our district covariates and crosswalk are built on a single boundary vintage, those states are excluded from the 2022 House backtest to avoid comparing predictions and results defined on different lines, leaving 320 contested districts on boundaries common to both cycles.

Table 4: 2022 midterm backtest (second held-out cycle). House is district-level Democratic two-party share (CD-level model); Senate is state-level Democratic-aligned two-way share.

Office	Subset	n	RMSE	Bias	r	Winners
House	All contested	320	3.9 pp	+1.76 pp	0.98	304 / 320
House	Competitive (42.5–57.5%)	89	3.5 pp	+1.33 pp	—	74 / 89
Senate	All races	33	3.8 pp	+1.44 pp	0.96	33 / 33
Senate	Competitive (42.5–57.5%)	14	2.1 pp	+0.99 pp	—	14 / 14

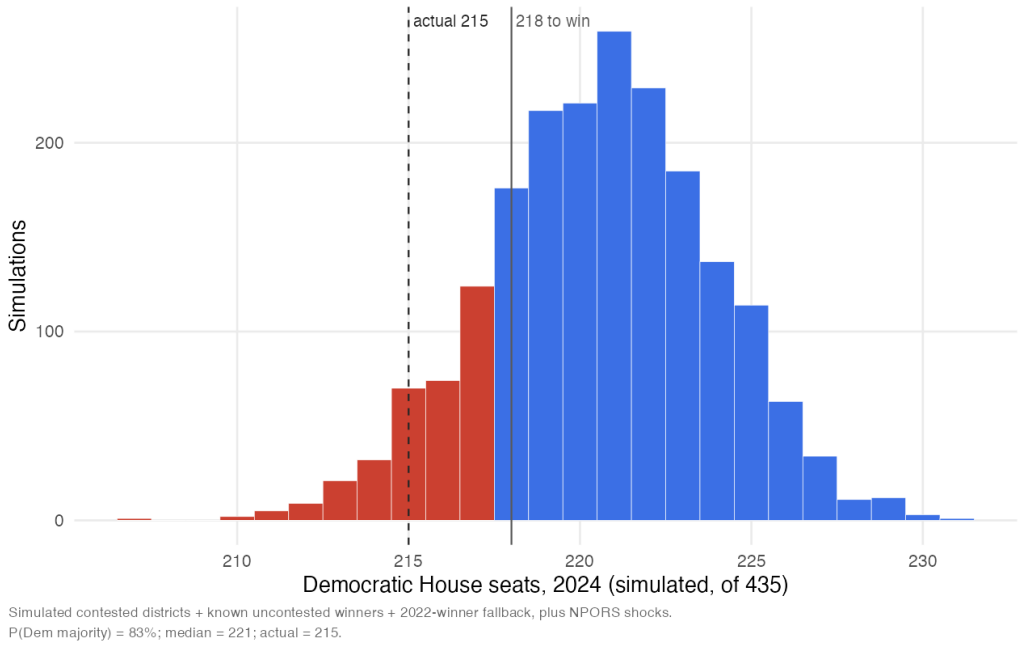

Figure 5: Simulated Democratic House seats, of 435: simulated contested districts plus known uncontested winners and 2022-winner fallback, with NPORS party-error shocks. P(Dem majority) = 83%; median 221; actual 215.

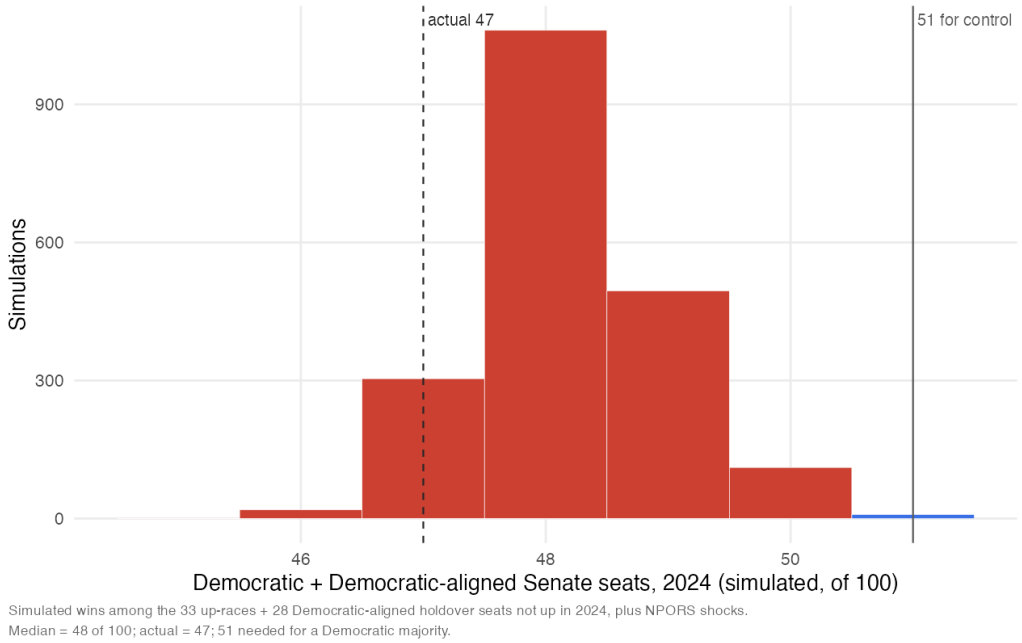

Figure 6: Simulated Democratic and Democratic-aligned Senate caucus (of 100): the simulated wins among the 33 up-races plus the 28 Democratic-aligned holdover seats not up in 2024, with NPORS party-error shocks. Median 48; actual 47; 51 needed for control.

The Senate recovers cleanly — 3.8 pp RMSE and every one of the 33 race winners called correctly (Figure 8). The House, on the 320 boundary-stable districts, is somewhat harder than 2024: 3.9 pp RMSE versus 3.4 pp, and 304 of 320 contested districts (95%) versus 98% (Figure 7). A midterm carries more incumbency and candidate-quality variation than the pipeline, anchored on presidential-vote geography, captures; the median absolute district error (2.6 pp) exceeds 2024's and the tail is heavier (a maximum single-district error of about 13 pp). The point is not that 2022 matches 2024 exactly, but that the same 2020-calibrated engine recovers a structurally different cycle without recalibration — cleanly for the statewide Senate contests and nearly as sharply for individual House districts.

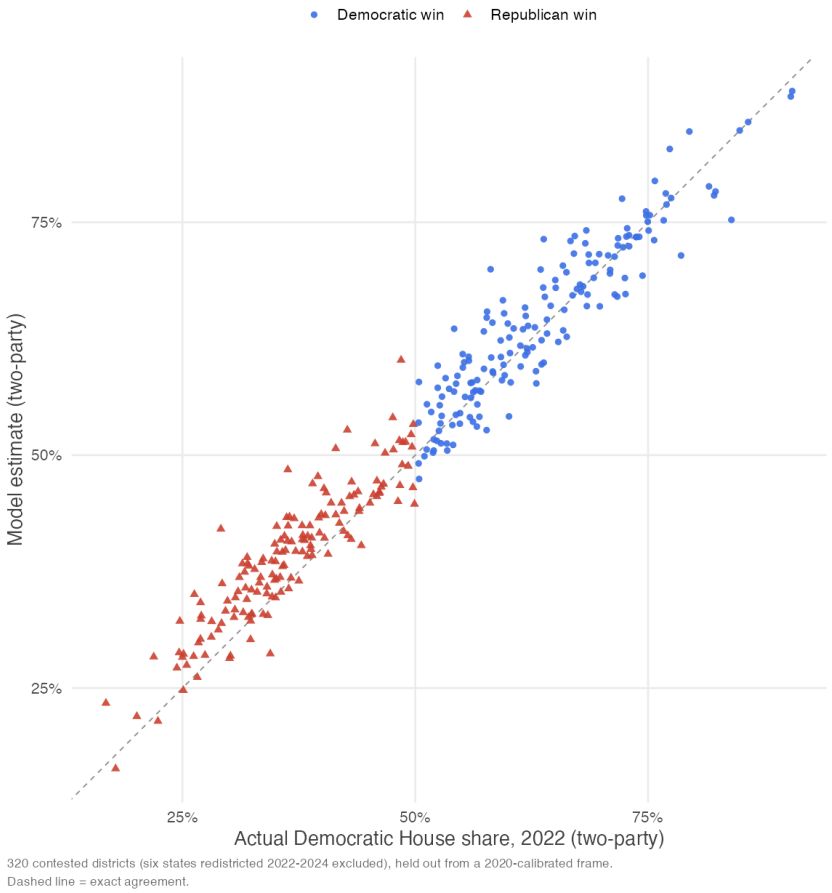

Figure 7: 2022 House backtest: predicted versus actual Democratic two-party share across 320 contested districts, from a 2020-calibrated frame with the 2022 returns held out. Six states redistricted between 2022 and 2024 are excluded (boundaries differ across cycles). Color and point shape both encode the actual winner.

5 The dynamic opinion layer

The frame serves as the post-stratification basis for the opinion estimates. MOSAIC post-stratifies survey-based opinion models against it to produce monthly estimates — presidential approval, issue handling, party trust, the generic congressional ballot, and party identification — at the national, state, congressional-district, and PUMA levels.

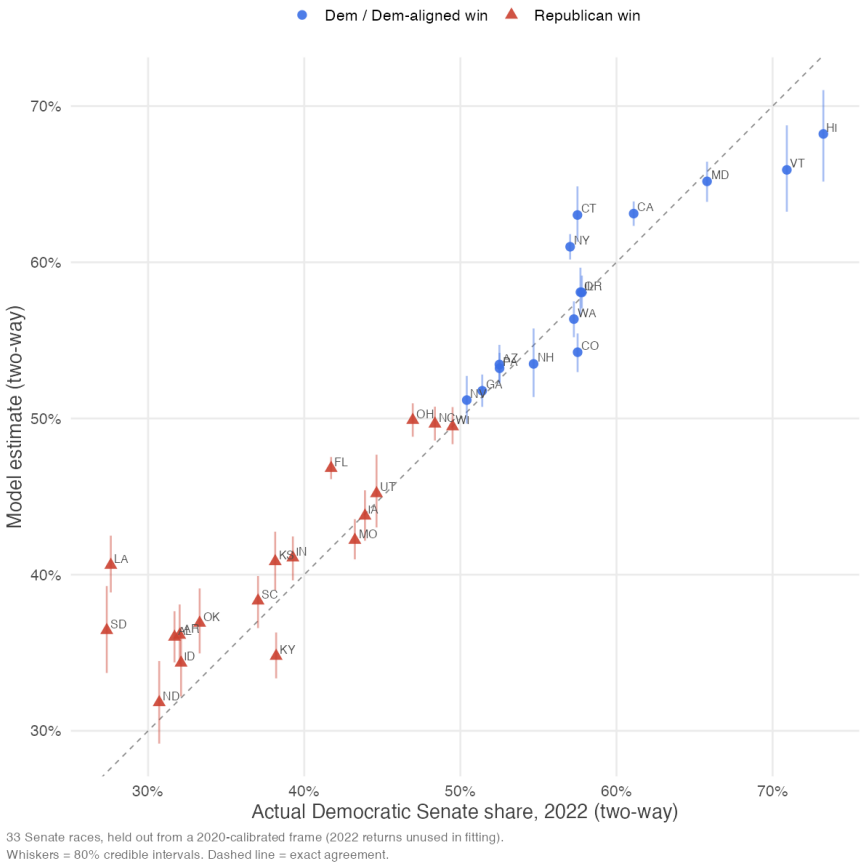

Figure 8: 2022 Senate backtest: predicted versus actual Democratic-aligned two-way share across 33 races, with 80% credible intervals. Color and point shape both encode the actual winner.

5.1 Survey data

The survey input is the Strength In Numbers Monthly U.S. Politics Poll, an online self-administered survey fielded on the Verasight probability-based panel. The universe is U.S. adults (18+), with registered voters reported as a subgroup; a typical wave is approximately 1,500 adults, and the pooled sample across the waves used in the models exceeds 12,000 interviews (waves fielded May 2025 through July 2026). Wave weights, produced by Verasight, match population benchmarks on age, sex, race, education, income, state, and a partisanship target.

5.2 Model specification

Time-varying outcomes share a dynamic Bernoulli-logit MRP implemented in a custom Stan program (the `binary_dynamic_mrp` family), run through `cmdstanr` under a weighted pseudo-likelihood: the wave weights enter the likelihood directly. For respondent i interviewed in wave $t[i]$, the linear predictor is

$$\begin{aligned}\text{logit Pr}(y_i = 1) \;=\; & \alpha_{t[i]} + \beta^{\text{race}}_{t[i],\,r[i]} + \beta^{\text{vote}}_{t[i],\,v[i]} + \beta^{\text{age}}_{t[i],\,a[i]} + \beta^{\text{edu}}_{t[i],\,e[i]} + \beta^{\text{inc}}_{t[i],\,k[i]} \\ & + \beta^{\text{sex}}_{t[i]}\, x^{\text{sex}}_i + \delta_{\text{state}[i],\,t[i]} + \beta^{d2p}_{t[i]}\, d_{p[i]} + \eta_i,\end{aligned}$$

where $\delta_{\text{state},t} = \delta_{\text{region},t} + \text{dev}_{\text{state},t}$ nests states in Census regions, $d_{p[i]}$ is the respondent's PUMA 2020 Democratic two-party share (so β^{d2p}_t is a partisanship-by-time elasticity), and η_i collects the static terms: the partial-pooled party effects ($\gamma^{\text{party}} + \gamma^{\text{party}\times\text{vote}} + \gamma^{\text{party}\times\text{race}}$), the demographic interaction families, and the contextual covariates. The model pools across all monthly waves and allows the overall level of each series to evolve while constraining higher-order interactions to remain static: every time-indexed term above evolves as a random walk across waves,

$$\theta_t = \theta_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \text{Student-}t(4,\, 0,\, \sigma_\theta),$$

with one walk per factor level and non-centered parameterization throughout for sampler stability. The Student-t ($df = 4$) innovations have fatter tails than Gaussian, so the latent series can accommodate abrupt between-wave changes while remaining smooth otherwise. The pairwise and three-way demographic interactions (sex × race, sex × age, race × education, race × income, race × region, race × vote, sex × vote, age × vote, education × vote, income × vote, and sex × age × white), the PUMA contextual covariates (white share, log median income, college share), and the state white-evangelical share are held static. Reflecting the party-in-frame design of Section 2.3, leaned party identification enters the outcome model as partial-pooled static effects — a party main effect plus party-by-vote and party-by-race interactions — and post-stratification runs over the vote-by-party-expanded frame. The one exception is the party-identification series itself: to avoid circularity, its outcome model omits the party terms and post-stratifies over the vote-only frame, so the published party-identification series is not constrained to reproduce the NPORS marginal.

The first-wave intercept and demographic walks are anchored on logit-scale group rates computed from the first wave's weighted survey data — initialization values that center the chains at a sensible starting point. The walk-innovation scales carry half-Normal priors: $\sigma \sim \mathcal{N}^+(0, 0.05)$ for the intercept, race, past-vote, and age walks; $\mathcal{N}^+(0, 0.03)$ for the education, income, sex, region, and partisanship-slope walks; and $\mathcal{N}^+(0, 0.02)$ for the within-region state-deviation walks. Interaction standard deviations carry Student-$t(3, 0, 0.5)$ priors (the party main effect, Student-$t(3, 0, 1.5)$); static region and state scales carry Student-$t(3, 0, 0.5)$ and Student-$t(3, 0, 0.25)$; and static coefficients carry Normal(0, 1) priors. "Don't know" responses are excluded from the model and reported separately on the topline. DC is excluded from

fitting (its composition destabilizes the geographic effects); its frame cells receive predictions through the demographic and regional terms and contribute to national aggregates, but no DC-specific estimate is published. Post-stratification runs separately for each month against the same calibrated frame, with aggregation to any geography exactly as in Section 2.4, yielding monthly national, state, district, and PUMA series with partial pooling across months.

5.3 Relation to prior tracking models

This design belongs to a tradition of dynamic Bayesian opinion measurement. Linzer (2013) modeled state-level presidential preferences as latent series evolving by random walk, partially pooled across states, and showed that sparse and irregular polling can yield stable estimates when time itself is part of the multilevel structure; the design insight is that smoothing belongs *inside* the model, with opinion at time t a latent quantity whose prior is its own recent past. The closest operational relative is the Civiqs tracking system (Civiqs, 2026), developed by Linzer: a dynamic Bayesian MRP in which demographic effects evolve as Gaussian random walks from day to day, post-stratified up to state and national figures. MOSAIC differs in cadence and target — monthly waves with informative walk priors rather than daily tracking, heavier-tailed innovations, and estimates maintained down to the PUMA level on the vote-and-party-reconstructed frame tested in this report. And unlike Linzer (2013), MOSAIC is measurement rather than forecasting: there is no election-day anchor, and the random walks are smoothing devices, not predictions about where opinion is headed.

5.4 Validation against independent state polls

We validate the opinion layer against independent state approval polls collected in the FiftyPlusOne polling database (May 2025 through July 2026). Each poll is matched to the model's estimate for the same state in the same month, on the same two-way approve scale (approve ÷ [approve + disapprove]), with the population base matched (adult samples to the adult estimate; registered-voter and "voter" samples to the registered-voter estimate). Likely-voter samples are excluded, as is Morning Consult, which accounts for more than half of all state approval polls in the database.

Across **230 polls** of President Donald Trump's approval rating in 2025 and 2026, the model tracks the cross-state ordering of approval closely (**Pearson** $r = 0.88$); the poll-level RMSE is **3.9 percentage points**, with a mean absolute difference of **3.1 pp** (Figure 9). The mean poll-minus-model difference is +1.3 pp — a small positive lean within the range of ordinary house effects. These figures are model-to-poll disagreement statistics rather than error against a known truth: each poll carries its own sampling error, house effects, and mode and wording differences, and the cross-state correlation partly reflects stable differences between strongly Republican and strongly Democratic states rather than within-state temporal tracking alone. Because the polls are themselves noisy, the raw correlation understates the model's agreement with true state opinion. Correcting for attenuation using each poll's sample size — the polls carry a mean sampling error of about 1.7 pp against a cross-state spread of 7.7 pp, a reliability of 0.95 — raises the correlation to **0.90** and lowers the implied model error to **3.5 pp**. Allowing for non-sampling error as well, under the common rule of thumb that total survey error runs about twice the sampling margin, puts the correlation near 0.98 and the implied error near 1.9 pp; we report the sampling-only adjustment as the conservative figure.

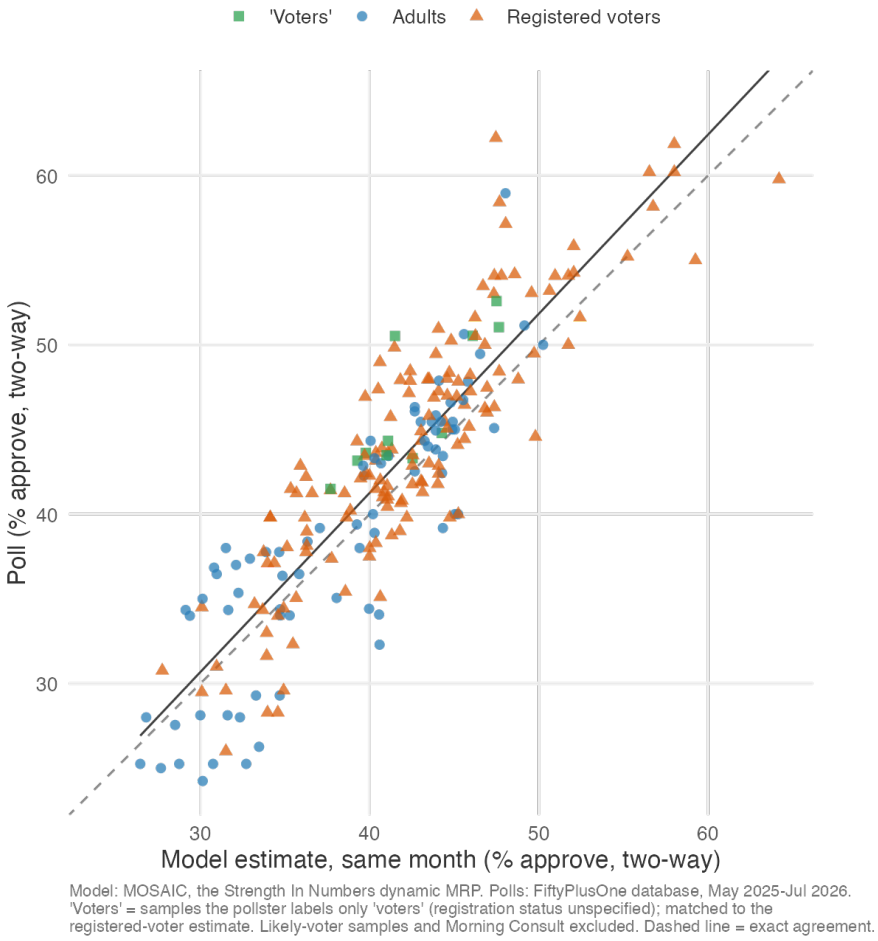

Figure 9: Independent state approval polls versus the model's estimate for the same state and month, on the two-way approve scale, May 2025 – July 2026. The dashed line marks exact agreement; the solid line is the least-squares fit. "Voters" denotes samples the pollster labels only as voters, with registration status unspecified; these are matched to the registered-voter estimate.

6 Limitations

Two cycles, not many. The backtest now covers two held-out elections — the 2024 general and the 2022 midterm (Section 4.8) — which are usefully different environments, and the engine recovers both without recalibration. But two cycles are still not a distribution over cycles: both sit within a single frame vintage, the 2022 House result is marginally weaker than 2024, and neither speaks to how the engine would behave under a sharply different turnout regime. The 2026 cycle will be the next out-of-sample test.

Specification choices used the backtest. As noted in Section 3, the party coding and imputation-model choices were selected by comparing configurations on this backtest, so the exercise is retrospective validation rather than a pre-registered test.

The estimand is citizen adults. The frame is restricted to citizen adults while the survey universe is all adults, and the validation polls' adult samples are compared against citizen-adult estimates; whether the NPORS party benchmark was restricted to citizens before pinning is likewise a frame-consistency question. The practical effect is small — noncitizens are a small share of adults and are concentrated in cells with low modeled registration — but the estimand should be read as citizen adults throughout.

Intervals are conditional on the estimated frame. The credible intervals propagate posterior uncertainty in the outcome models but treat the frame — ACS weights, imputations, voter-file match, party probabilities, geographic shifts, crosswalk, and calibration deltas — as fixed. The empirical undercoverage in Section 4.6 is the visible cost. Frame-level uncertainty propagation (multiple frame realizations via bootstrap or imputation draws) is the natural extension.

A frame anchored to one cycle. The production frame is calibrated to the certified 2024 electorate. It is the appropriate benchmark for the current period, but composition and turnout propensity drift over a cycle, and the frame will require recalibration as the electorate evolves toward 2026 and beyond.

Within-cell nonresponse is assumed ignorable. Post-stratification corrects for observable composition. Residual nonrandom selection *within* a demographic-geographic cell is not directly corrected; the sensitivity analysis noted in Section 3 bounds its plausible magnitude on the 2024 target and finds it small, but the assumption is not eliminated.

The opinion validation is state-level and single-outcome. Independent public approval polls exist at the state level, not the district level, and for presidential approval rather than the full battery; the concurrent validation in Section 5.4 therefore tests one outcome at one geographic level. The district-level evidence is the election backtest, which tests the underlying geographic engine rather than each opinion series directly.

7 Discussion

The two validations are complementary. The election backtest shows that the geographic and vote-reconstruction engine — the part of the pipeline that places opinion in space — recovers a held-out election to within about two points at the state level, for both the presidential and Senate contests, and about three and a half points at the district level, with all 50 state winners and 98% of contested district

winners called correctly; a second held-out cycle, the 2022 midterm, replicates that recovery on a different kind of election, cleanly for both the Senate and the House. The polling comparison shows that the opinion estimates the engine carries show substantial agreement with independent state-poll estimates of the same quantity. Neither test alone would be sufficient; together they cover the geographic engine and the opinion layer that rides on it.

The backtest is a retrospective test of recovery, not a claim about future outcomes. MOSAIC is measurement, not forecasting: the system is designed for descriptive small-area inference — what opinion looks like now, across places and groups, with posterior credible intervals whose conditioning is stated explicitly (Section 4.6).

A Turnout weighting and the likely-voter question

Every post-stratified vote estimate in this report weights each frame cell not only by its population but by its probability of voting. This appendix documents that turnout weight, backtests it against the standard alternative — a contemporaneous "likely-voter" model — and explains why the past-vote-calibrated weight is the more accurate choice.

A.1 How turnout enters the frame

The frame assigns each cell a voter mass equal to its population count times a modeled turnout probability, $N_c\, p_c^{\text{voted}}$ (Section 2.4). The turnout probability comes from the structural turnout stage of the vote reconstruction (Section 2.3): a multilevel model of validated turnout conditioning on demographics, geography, and — most importantly — past vote, calibrated so that each state's cell-weighted turnout matches its certified total from the anchoring presidential election (2020). Past vote dominates: 2020 voters carry turnout probabilities near 0.95, 2020 non-voters near 0.35. We call this the *past-vote turnout weight*, because its level is anchored to how the electorate actually turned out in the last presidential election.

The standard alternative is a *likely-voter* weight built from contemporaneous self-reported turnout intent — the approach behind most published likely-voter models. We fit its analog here (a multilevel model of survey turnout intent on the same demographic, geographic, and past-vote structure) and use its predicted propensity to replace p_c^{voted} in the frame. Because a midterm electorate is older and lower-turnout than a presidential one, and the structural weight is anchored to 2020 presidential turnout, an intent-based weight is a natural candidate to do better, especially in a midterm.

A.2 A backtest of the two turnout weights

We isolate the turnout weight cleanly. Each race uses the *same* canonical vote-choice model — unchanged, with turnout kept out of the regression — and we vary only the frame, replacing the past-vote weight with the intent-based likely-voter weight and re-post-stratifying. Table 5 reports both weights across all five held-out races.

The past-vote turnout weight is more accurate in four of the five races. The intent-based likely-voter weight degrades all three statewide contests — both Senate races and the presidential race — in both cycles, and is essentially neutral for the 2024 presidential-year House. It helps in exactly one place: the 2022 midterm House, where it lowers RMSE from 4.2 to 3.9 points and calls four more districts correctly.

Table 5: Turnout-weight backtest: the past-vote-calibrated weight versus an intent-based likely-voter weight, holding the vote-choice model fixed. Within each race the two rows share the identical canonical fit and simulation seed, so the only difference is the frame's turnout weight. This turnout-weight comparison is orthogonal to, and was computed before, the weighting choice of Appendix B: both rows here use the inverse-propensity-weighted vote-choice fits, so their absolute levels are the weighted ones and run slightly above the now-unweighted main-text results; the past-vote-versus-likely-voter contrast is unaffected.

Race	Turnout weight	RMSE	Bias	r	Winners	Comp. RMSE
2024 President	Past vote (used)	1.74	+0.88	0.990	49/50	1.37
	Likely-voter	2.02	+1.43	0.991	47/50	1.79
2024 House	Past vote (used)	3.64	+1.51	0.978	380/393	3.99
	Likely-voter	3.67	+1.96	0.982	380/393	3.89
2024 Senate	Past vote (used)	2.15	+0.81	0.982	31/33	2.05
	Likely-voter	2.27	+1.23	0.984	31/33	2.06
2022 House	Past vote (used)	4.18	+1.12	0.968	302/320	4.25
	Likely-voter	3.93	+1.67	0.975	306/320	3.94
2022 Senate	Past vote (used)	3.56	+1.54	0.965	33/33	2.86
	Likely-voter	3.78	+2.03	0.966	33/33	2.84

A.3 Why the past-vote weight wins

The pattern has a clean explanation. The turnout weight is calibrated to 2020 *presidential* turnout, which is close to the realized electorate in a presidential year. In 2024, therefore, the intent-based reweight has no miscalibration to correct — it only moves the estimate, and in every race it moves it the same direction, more Democratic, because it down-weights the low-propensity 2020-non-voter cells that lean Republican in these contests. So it adds bias without a compensating gain. Only in the 2022 midterm is the presidential-anchored weight genuinely wrong — it over-weights the voters who skip midterms — and only at the district level, where composition varies enough to matter, does correcting it outweigh the added bias; the statewide races' state random effects already absorb the composition, leaving only the added bias behind.

Beneath the cycle-specific story is a general point about turnout prediction. A turnout model can be unambiguously better at predicting *turnout* — who votes — and still produce a worse *vote-share* estimate. The vote-share bias is a turnout-weighted average of the vote-choice model's per-cell error,

$$B \;=\; \frac{\sum_c w_c\, \varepsilon_c}{\sum_c w_c}, \qquad \varepsilon_c = \hat{p}_c - p_c.$$

If the vote-choice model were conditionally unbiased ($\varepsilon_c = 0$), the turnout weights would not affect the bias at all; the fact that a better turnout weight *changes* the bias shows that ε_c is nonzero and correlated with turnout propensity. A better turnout model does not fix those per-cell errors — it reweights them, and reweighting toward the true high-propensity electorate can expose a vote-choice bias that the structural weight had partially averaged away (the smaller bias under the past-vote weight is, in part, error cancellation). Improving turnout and improving vote share are distinct objectives; they coincide only to the extent the vote-choice model is unbiased within the cells being reweighted. Consistent with this, the likely-voter weight *raises* the correlation r in most races — its ranking of places is if anything sharper — even as it worsens RMSE and bias, because it improves composition while exposing the conditional bias.

A.4 Choice and refinement

MOSAIC uses the past-vote-calibrated turnout weight p_c^{voted} for every office and cycle. It is the more accurate weight in backtesting — better in four of five held-out races and never materially worse — and it avoids the cross-cycle inconsistency of adopting an intent-based weight that helps only the midterm House while adding bias everywhere else. The single midterm-House gain is real and points to the natural refinement: anchoring the turnout weight to the appropriate cycle's turnout rate (a prior midterm for a midterm) rather than always to the last presidential election. We leave that to future work.

B Survey weighting in the vote-choice models

Every vote-choice model in this report is fit as an *unweighted* multilevel regression: survey respondents enter the likelihood with equal weight, and representativeness is supplied entirely by post-stratification over the frame. This appendix documents that choice, backtests it against the natural alternative — weighting each respondent by an inverse-propensity pseudo-weight that corrects the opt-in sample toward the population — and explains why the unweighted fit produces more accurate vote shares.

B.1 How the weights would enter

The Strength In Numbers survey is an opt-in online panel, so its respondents are not a random sample of the population even within demographic cells. A standard correction is an inverse-propensity weight (IPW): stack the survey against a synthetic sample drawn from the population frame, model the probability that a record is a survey respondent, and weight each respondent by $(1 - \hat{p})/\hat{p}$ so the weighted survey matches the population on the modeled margins. Passing that weight into the vote-choice regression — `outcome | weights(ipw)` $\sim \cdots$ — makes the fit a pseudo-population fit rather than a raw-sample fit.

The alternative, which MOSAIC adopts, is to fit the regression unweighted and let post-stratification do the reweighting. In a multilevel-regression-and-post-stratification model the two are partly redundant: the frame already reweights each cell's fitted probability to its population count, so the demographic and geographic composition the IPW weight targets is corrected a second time downstream. Whether adding the weight to the fit helps or hurts the final estimate is therefore an empirical question.

B.2 A backtest of weighted versus unweighted

We isolate the weight cleanly. Each race uses the *same* canonical vote-choice model over the *same* frame; the only change is whether the IPW pseudo-weight enters the likelihood, and the model is re-fit and re-post-stratified in each condition. Table 6 reports both across all five held-out races; the weighted rows reproduce the main-text results.

The unweighted fit is more accurate in four of the five races. It lowers RMSE and mean absolute error and calls more races correctly in both House cycles and both 2024 statewide contests; the gains are largest at the district level (House 2022 RMSE 4.17 $\rightarrow$ 3.50; House 2024 3.63 $\rightarrow$ 3.40), and it recovers wrong winner calls in the presidential race (Georgia) and the 2024 Senate (Montana) while netting additional correct districts in both House cycles. The lone exception is the 2022 Senate, where the weighted fit is modestly better on RMSE (3.29 versus 3.52) though both call all 33 races. Across the board the unweighted fit's bias is a fraction of a point *higher* even as its RMSE falls — the signature of removing variance rather than bias.

Table 6: Weighting backtest: inverse-propensity-weighted versus unweighted vote-choice fits, holding the model and frame fixed. Within each race the two rows share the identical model, frame, and seed, so the only difference is whether the IPW pseudo-weight enters the likelihood. House 2022 is scored on the 320 districts whose 2022 and 2024 boundaries agree.

Race	Weighting	RMSE	MAE	Bias	r	Winners
2024 President	IPW-weighted	1.74	1.35	+0.88	0.990	49/50
	Unweighted (used)	1.60	1.28	+0.66	0.992	50/50
2024 House	IPW-weighted	3.63	2.60	+1.49	0.978	380/393
	Unweighted (used)	3.40	2.43	+1.56	0.982	386/393
2024 Senate	IPW-weighted	2.15	1.77	+0.79	0.982	31/33
	Unweighted (used)	2.04	1.70	+0.81	0.985	32/33
2022 House	IPW-weighted	4.17	3.19	+1.10	0.966	302/320
	Unweighted (used)	3.50	2.71	+1.23	0.976	304/320
2022 Senate	IPW-weighted	3.29	2.30	+0.79	0.966	33/33
	Unweighted (used)	3.52	2.37	+0.87	0.966	33/33

B.3 Why unweighted wins

The pattern follows from what post-stratification already does. The IPW weight and the frame both correct the same thing — the survey's departure from the population on demographics, geography, and past vote — so applying both double-corrects the composition. What the weight adds beyond that is variance. Inverse-propensity weights are heavy-tailed: a few rare, up-weighted respondents dominate the cells they fall in, and in a small sample that noise passes straight into the fitted cell probabilities and the state random effects. The damage concentrates exactly where samples are thinnest — small, low-population states — and where the up-weighted low-propensity respondents happen to lean against the state. Montana is the clean example: with roughly two hundred in-state Senate respondents, the weighted fit put the Democrat at 50.2% (a wrong call), pulled up by heavily-weighted respondents in the 2020-non-voter and independent cells; unweighted, the same model over the same frame lands at 46.6% against an actual 46.4%, because the frame — not a handful of weights — sets the composition.

Beneath the specific case is a general one. Post-stratification consistently estimates the population mean when the cell probabilities are right; survey weights in the regression neither add nor remove bias if the model is conditionally unbiased, but they always inflate the variance of the fitted probabilities. When the frame is mature — calibrated past vote, imputed party, geographic covariates — the composition correction the weight was meant to supply is already present, so the weight is close to pure added noise. Hence the observed signature: lower RMSE with slightly higher bias, because dropping the weight removes variance while giving up the small error cancellation the weighting incidentally provided. The 2022-Senate exception is consistent with this reading — in that one cycle the weight's incidental composition shift helped a touch more than its added variance hurt, but not enough to change a call.

B.4 Choice

MOSAIC fits every vote-choice model unweighted and relies on post-stratification for representativeness. It is the more accurate configuration in backtesting — better in four of five held-out races, and never worse by more than a quarter-point of RMSE in the one exception, with no lost winner call anywhere. Residual

bias, which unweighting slightly raises, is addressed separately and downstream by the non-response (MUBP) correction, which is built for exactly the within-cell partisan skew that neither weighting nor post-stratification can reach.

C Leakage audit: keeping the held-out election held out

A backtest is only meaningful if the target election is genuinely invisible to the pipeline. We audited every input to the frame chain, the vote-choice models, and the weighting stage against that standard. Three channels were found by which post-election information reached the pipeline; all three have been closed, and every result in this paper comes from the corrected pipeline. We describe them, and what each was worth, because the size and *sign* of a leak are as informative as its existence.

The training sample was selected on post-election behaviour. The presidential and House models assigned `weight = commonpostweight` and then dropped rows where it was missing. That field is non-missing *exactly* for respondents who returned to complete the post-election wave — a perfectly degenerate cross-tab, with all 10,568 non-returners missing and all 49,432 returners present — so the filter conditioned the training sample on behaviour occurring after Election Day, discarding 17.6% of respondents. The discarded group was 7 pp more Democratic than the retained group, so the filter moved the raw training signal about 1.2 pp (presidential) and 1.7 pp (House) *toward* the eventual result. The weight never entered the likelihood; the filter was its only effect. Removing it and refitting *improved* the presidential backtest (RMSE 1.65 → 1.60, bias +0.79 → +0.66) and left the House essentially unchanged — post-stratification already corrects composition, so the recovered sample size mattered more than the shifted mean. Every outcome and covariate variable is drawn from the pre-election wave, which closed on 4 November 2024, the day before the election; after this fix nothing in the training data is dated later.

A post-election survey wave trained the party model. The party-identification donor pool defaulted to all available NPORS waves, including NPORS 2025 (fielded February–June 2025). The national party marginal is re-pinned by raking to pre-election waves, so only the within-cell structure $P(\text{party} \mid \text{demographics})$ could leak — but that is precisely the surface on which the 2024 realignment appears. Capping the donor at 2024 changes the 2024 Senate backtest by 0.001 pp of RMSE, with a mean absolute per-state change of 0.005 pp: real in principle, negligible in effect. It is closed regardless.

The 2022 backtest was post-stratified over a post-2022 party frame. This is the one that mattered. The 2022 models defaulted to the party frame whose marginal is raked to NPORS 2024 — party identification measured *after* the target election, used as a hard calibration constraint. Because party identification drifted roughly 3 pp Republican between 2022 and 2024, that constraint pulled the frame toward the realised 2022 result. Correcting it degrades both 2022 contests by almost identical amounts — Senate RMSE 3.5 → 3.8 and bias +0.87 → +1.44; House RMSE 3.5 → 3.9 and bias +1.23 → +1.76 — and the near-equality across two independent contests is the signature of a single shared cause rather than noise. The figures reported in Section 4.8 are the corrected ones. No winner call changes in either contest.

Researcher degrees of freedom. One further caveat belongs here rather than in a footnote. The choice to fit unweighted (Appendix B) was made by scoring weighted against unweighted *on the held-out elections*. That is a specification selected on the test set. The reasoning behind it is mechanical rather than fitted — post-stratification and inverse-propensity weighting correct the same quantity, so applying both adds variance without removing bias — and the choice is applied uniformly across all five contests rather

than tuned race by race. But it is not a clean out-of-sample decision, and a reader should discount the reported accuracy accordingly. The same applies, more weakly, to the turnout-weight comparison of Appendix A.

What we checked and found clean. Frame calibration targets are the 2020 certified returns throughout; the turnout model is trained on 2020 validated turnout from a voter-file match; the state and PUMA covariate files are programmatically stripped of every 2024 field, with the build script asserting it; the 2024 results files in the repository are read only by scoring code; the CES release carries post-election validated-vote columns (`TS_g2024` and related) that no script reads; and district incumbency was tested rather than trusted, matching the prior cycle's winners more closely than the target cycle's, as genuine on-ballot incumbency should. As an end-to-end check, the reconstruction beats a naive carry-forward of each district's 2020 presidential result by only 0.8 pp of RMSE, with a marginally *lower* rank correlation — the profile of a model that has learned demographic and past-vote structure rather than one that has seen the answer.

Disclosures

MOSAIC underlies Strength In Numbers Pro, a commercial subscription product operated by the author, who is the sole developer of the system and sole author of this report. The Strength In Numbers Monthly U.S. Politics Poll is fielded by Verasight on its probability-based panel under Verasight's panel consent procedures. The modeling pipeline is not currently public; the poll-comparison dataset and per-race backtest residuals underlying the figures are available from the author.

How to cite

> Morris, G. Elliott (2026). *Accurate Local-Level Estimation of U.S. Public Opinion: Methodology and Out-of-Sample Validation of MOSAIC, a Synthetic MRP Model Calibrated to Various Political and Census Demographic Targets.* Strength In Numbers Pro Technical Report v1.1. `https://pro.gelliottmorris.com/whitepaper/mosaic-methodology-v1.pdf`

References

Andridge, R. R., West, B. T., Little, R. J. A., Boonstra, P. S., & Alvarado-Leiton, F. (2019). Indices of non-response bias for proportions estimated from non-probability samples. *Journal of the Royal Statistical Society: Series C.*

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1).

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1).

Civiqs (2026). *Research Methodology.* `https://civiqs.com/methodology` (accessed July 2026).

Gabry, J., Češnovar, R., & Johnson, A. (2024). *cmdstanr: R Interface to CmdStan* [software].

Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23(2).

Ghitza, Y., & Gelman, A. (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, 57(3).

Kuriwaki, S., Ansolabehere, S., Dagonel, A., & Yamauchi, S. (2024). The geography of racially polarized voting: Calibrating surveys at the district level. *American Political Science Review*, 118(2), 922–939.

Lax, J. R., & Phillips, J. H. (2009). How should we estimate public opinion in the states? *American Journal of Political Science*, 53(1).

Linzer, D. A. (2013). Dynamic Bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association*, 108(501), 124–134.

Mercer, A. W., Kreuter, F., Keeter, S., & Stuart, E. A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly*, 81(S1), 250–271.

MIT Election Data and Science Lab (2021). *County Presidential Election Returns* [dataset]. Harvard Dataverse.

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, 12(4).

Pew Research Center (2020–2025). *National Public Opinion Reference Survey (NPORS)* [datasets]. Annual address-recruited probability sample of U.S. adults.

Pew Research Center (2025). *2023–24 Religious Landscape Study* [dataset].

Rosenman, E. T. R., McCartan, C., & Olivella, S. (2023). Recalibration of predicted probabilities using the "logit shift": Why does it work, and when can it be expected to work well? *Political Analysis*, 31(4), 651–661.

Schaffner, B., & Ansolabehere, S. (2025). *Cooperative Election Study 2024 Common Content* [dataset]. Harvard Dataverse.

Stan Development Team (2024). *Stan Modeling Language Users Guide and Reference Manual.*

U.S. Census Bureau (2024). *American Community Survey 5-Year Public Use Microdata Sample* [dataset].

U.S. Census Bureau (2025). *Current Population Survey, November 2024 Voting and Registration Supplement* [dataset].

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3).

Voting and Election Science Team (2021). *2020 Precinct-Level Election Results* [dataset]. Harvard Dataverse.

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1).