Forecasting the 2026 Korean Local Elections: A Reproducible Polls-plus-Fundamentals Model with a Pre-registered Validation Protocol

Han Kim

IOV Labs (아이오브연구소) · hankim@iovstudio.kr

Version 1.0 (pre-registration) · compiled 2026-05-30

Pre-registered working paper — predictions committed before the outcome, graded after polls close on 2026-06-03

Abstract. We forecast the 16 metropolitan-executive (광역단체장) races of South Korea's 9th nationwide local election (3 June 2026) by combining a structural *fundamentals* estimate — each region's 2022 two-way vote swung to the 2026 environment on the logit scale — with *method-normalized poll aggregates*, fused by poll-count-weighted hierarchical shrinkage. Outcome uncertainty is propagated through a 50,000-draw correlated Monte Carlo with a three-level error budget (national ⊕ cluster ⊕ local) and heavy-tailed (normal-mixture ≈ Student-t) innovations, so that a single nationwide polling miss moves correlated regional blocs together rather than averaging out. The entire pipeline is seeded and reproducible to the bit. The central estimate is 민주 **(Democratic Party) 12 of 16 seats** (90% credible range 8–15), with five genuine toss-ups (서울, 부산, 경남, 충북, 울산) and only 대구·경북 leaning 국힘 (People Power Party). The error model is calibrated against the 2022 final phone polls (bias −0.1 pt, mean absolute error 2.2 pt, $\sigma \approx 2.6$), and the dominant failure mode — a *correlated* poll bias rather than any internal parameter — is quantified by an explicit ±4 pt scenario sweep (a 3 pt pro-국힘 miss yields 11 seats). A parallel *silicon-sampling* experiment, in which a synthetic electorate was simulated with large language models, is reported as a **negative result**: it reproduced a stale training prior and contradicted every contemporaneous poll. The paper is written as a pre-registration: the forecast is committed before the result is known, turnout-dependent quantities update once on 2 June, and a fixed scoring script grades every claim after polls close on 3 June.

Keywords: election forecasting · poll aggregation · house effects · hierarchical shrinkage · correlated Monte Carlo · heavy-tailed errors · calibration · proper scoring rules · Brier score · silicon sampling · reproducibility · pre-registration

Contributions.

1. A complete, auditable polls-plus-fundamentals forecast for a data-sparse, multi-region, bipolar contest, with every parameter traced to a backtest or declared as a prior.
2. An explicit treatment of the largest hidden bias in Korean polling — the live-phone vs. automated-response (ARS) mode gap — calibrated rather than guessed.
3. A correlated, heavy-tailed Monte Carlo whose tail behaviour and cross-region dependence are derived, not asserted, and whose dominant risk is isolated by a systematic-bias sweep.
4. A documented *negative result* for LLM "silicon sampling" in a forward-looking, contested race.
5. A reproducibility and pre-registration protocol that makes the forecast falsifiable on a fixed date by a fixed metric.

Contents

1 Introduction

A forecast is not a prophecy. It is a structured, falsifiable statement of uncertainty about a future that has not happened and will happen exactly once. The discipline of election forecasting consists almost entirely of taking that one sentence seriously: deciding what one is entitled to claim from noisy, biased, incomplete data; quantifying how wrong one might be; and committing to a number before the world reveals the answer. This paper documents such a statement for the metropolitan-executive races of South Korea's 9th nationwide local election, held on 3 June 2026.

The model we present is deliberately conventional in its skeleton and deliberately disciplined in its execution. It belongs to the "polls-plus-fundamentals" family that has become standard in US presidential forecasting [1], [2]: a structural estimate derived from past results supplies a prior; current polls update that prior; and a correlated simulation converts the blended estimate into a probability distribution over outcomes rather than a single point guess. What distinguishes the present work is not novelty of architecture but two commitments that are easy to state and surprisingly rare to honour. First, every quantitative choice is either traced to an out-of-sample backtest or explicitly flagged as a prior, so that no parameter is a silent hand-tuning. Second, the entire pipeline is deterministic — a pure function of its input data and a fixed random seed — so that there is exactly one forecast to defend, chosen before the outcome, with no opportunity to re-roll the simulation until it flatters the forecaster.

1.1 Why this problem is hard

Forecasting Korean metropolitan elections poses difficulties that the well-studied US presidential case largely avoids. The races are numerous (sixteen simultaneous contests) but individually thin in public polling, with several regions effectively unpolled. The dominant survey modes — live telephone interview (전화면접) and automated response (ARS) — disagree systematically and by large margins, sometimes by more than fifteen points in two-way share, reflecting a mode-dependent version of the spiral-of-silence / shy-respondent phenomenon [3]. Turnout is volatile across cycles (the metropolitan turnout fell from 60.2% in 2018 to 50.9% in 2022), which matters acutely for any prediction of raw vote counts. Several races are not cleanly two-way: a region may pit the ruling party against an independent, or feature a progressive third candidate who splits the anti-incumbent vote. And the contest is governed by an election law (공직선거법 §108) that restricts the publication of forecasts during a pre-election blackout — a constraint we take up in Section 10.3.

1.2 The single-event problem

Underlying every probabilistic claim in this paper is a philosophical commitment worth stating at the outset. When we write that the Democratic candidate wins 서울 "with probability 73%," there is no long run in which 서울 votes a hundred times and the candidate wins seventy-three of them; the election happens once. The number is therefore not a frequency but a degree of belief, conditioned on the data and assumptions documented here. Its only honest test is *calibration across many such claims*: if the events we call "70%" occur about seventy percent of the time and those we call "90%" about ninety percent, the probabilities carry information. This is why the model commits sixteen simultaneous probabilities and scores them with a proper scoring rule [4], [5]; a single race can never validate a probability, but sixteen begin to, and a track record across cycles eventually does. We return to these epistemics in Section 10.

1.3 Roadmap

Section 2 situates the model in the forecasting and polling literature. Section 3 describes the 2026 electoral context and the legal constraints. Section 4 documents the data and its provenance. Section 5 — the core of the paper — specifies the generative model in full, with derivations for each component. Section 6 reports the empirical calibration against a 2022 backtest. Section 7 presents the forecast, race by race. Section 8 analyses robustness and sensitivity. Section 9 reports the silicon-sampling negative result. Section 10 discusses the epistemics, ethics, and limitations. Section 11 states the pre-registration and validation protocol. Appendices

provide full derivations, the per-region input data, a hand-worked example, pseudocode, and a glossary of Korean terms.

2 Background and related work

2.1 Fundamentals and the predictability of elections

A long tradition in political science holds that elections are, in aggregate, more predictable than the day-to-day variation of campaign polls suggests. Lewis-Beck and Rice [6] formalized the use of structural "fundamentals" — economic conditions, incumbency, prior partisanship — to forecast outcomes months in advance. Gelman and King [7] famously asked why campaign polls are so variable when votes are so predictable, and answered that much poll movement is noise around a fundamentals-anchored equilibrium toward which opinion reverts. The practical lesson, which this model adopts, is that a region's past vote is a powerful prior that polls should update but not overwhelm — especially where polling is thin.

2.2 Poll aggregation and house effects

No single poll is the truth; aggregation reduces variance and, done carefully, bias. Jackman [8] modelled the pooling of polls over a campaign as a latent-state estimation problem in which each pollster carries a "house effect" — a systematic lean to be estimated and removed. Subsequent evaluations of polling error [9], [10] decomposed total error into bias and variance components and showed that correlated, industry-wide bias — not independent sampling noise — drives the large, memorable misses. This finding is structural to our error model: the Monte Carlo's national and cluster shocks (Section 5.5) exist precisely to represent the correlated component that an independent-errors model would wrongly average away.

2.3 Hierarchical models and poststratification

Where individual-level survey microdata are available, multilevel regression and poststratification (MRP) [11], [12] estimates opinion in small areas by partially pooling toward a model and reweighting to known population margins; it has enabled credible forecasts even from non-representative samples [13]. The present model does not have access to the microdata MRP requires, but it borrows the same two ideas in a coarser form: partial pooling (poll estimates shrunk toward fundamentals by an evidence-dependent weight) and reweighting (turnout and two-party adjustments to recover vote counts). MRP is the natural extension and is noted as future work.

2.4 Korean polling: the mode problem

The defining feature of Korean pre-election polling is the divergence between live-interview and ARS modes. Live interviews tend to elicit larger ruling-party leads; ARS surveys, which over-represent highly engaged respondents and reduce social-desirability pressure, tend to show tighter races and stronger conservative support — a mode-specific manifestation of the spiral of silence [3], locally termed 샤이보수 ("shy conservative"). Because the two modes can differ by ten to sixteen points in the same region at the same time (Section 7.2), any aggregator that ignores mode is at the mercy of the survey mix. We therefore treat mode as a first-class house effect and calibrate its correction against the 2022 outcome.

2.5 Silicon sampling

A recent literature proposes using large language models to simulate survey respondents — "silicon samples" [14] — on the hypothesis that a model conditioned on a demographic persona reproduces that subpopulation's attitudes. We tested this directly for the 서울 mayoral race and obtained a clear negative result (Section 9). The finding is consistent with the critique that LLMs encode a stale, training-time distribution: asked to imagine a 2026 voter, the models returned what voters *were*, not what contemporaneous polls say they *are*. We treat this as informative about the method's limits for forward-looking, contested prediction, and retain the experiment for that reason.

2.6 Scoring and calibration

Forecast quality is assessed with proper scoring rules [4] — rules minimized in expectation by reporting one's true beliefs. We use the Brier score [5] and read it through Murphy's decomposition [15] into uncertainty, resolution, and reliability (calibration). The pre-committed scoring script (Section 11) reports these after the election. The broader stance — that good judgement is a track record of calibrated probabilities, not a single dramatic call — follows the forecasting-tournament tradition [16].

2.7 Where this model sits

Table Table 1 locates the present model among the standard families. It is, deliberately, a small and auditable member of the FiveThirtyEight/Economist lineage rather than a methodological departure: it keeps the correlated-simulation core and the fundamentals-plus-polls structure, drops the components for which the data are too thin (a full economic index, a continuous state-space, individual-level microdata), and adds the one piece the Korean setting demands (explicit survey-mode normalization). Its distinguishing features are not in the estimator but in the discipline around it — calibration to a backtest, retention of a failed experiment, and bit-level reproducibility.

Approach	Signal / uncertainty	Relation to this work
Naive poll average	latest polls; margin of error	the starting point, then de-biased by mode and pooled toward fundamentals
FiveThirtyEight "polls-plus" [1]	polls ⊕ fundamentals ⊕ economy; correlated state sims	same skeleton, minus the economic index (local races, thin data)
Economist / Bayesian state-space [2]	full posterior; partial pooling	approximates the spirit (shrinkage + correlated error) without full MCMC
MRP [11], [13]	individual survey + census; model-based	future work; requires microdata not available here
Prediction markets	aggregated wagers; implied probability	left as an ensemble hook; can outperform models
Silicon sampling [14]	LLM-simulated respondents	tested and reported as a negative result (Section 9)
This model	polls ⊕ fundamentals, mode-normalized; seeded correlated heavy-tail MC	a small, auditable, reproducible take on the polls-plus family

Table 1: The model among forecasting approaches.

3 The 2026 Korean local elections

3.1 Institutions

South Korea holds unified local elections every four years. Among the offices contested, the most consequential are the *metropolitan executives* (광역단체장): the mayors of the special, metropolitan, and special-self-governing cities, and the governors of the provinces. Each is elected by single-round plurality (first-past-the-post) within the region. These are the races forecast here. They are politically salient as a nationwide barometer of the two major parties — the center-left Democratic Party (더불어민주당, here "민주" or D), currently the ruling party, and the conservative People Power Party (국민의힘, "국힘" or P) — and, because they occur at the mid-point of national political cycles, as a referendum on the incumbent administration.

3.2 The 16 races

A structural change shapes this cycle: the administrative consolidation of 광주 and 전남 into a single merged jurisdiction (here denoted 전남광주) reduces the count of metropolitan executives from seventeen to sixteen. We forecast all sixteen. The regions span seven correlation blocs used throughout this paper (Table Table 2): the capital area (수도권: 서울·인천·경기), the central 충청 belt, the southeastern 영남 coast, the 대경 inland, the southwestern 호남 stronghold, and the standalone 강원 and 제주. The bloc structure encodes the empirical fact that polling errors and political swings travel together within these groupings, which the model exploits in its correlated simulation.

Cluster	Regions	Character
수도권 (Capital)	서울 · 인천 · 경기	Democratic strength, 경기 decisive
충청 (Central)	대전 · 세종 · 충북 · 충남	Democratic-leaning; 충북·충남 competitive, mode-sensitive
영남 (Southeast)	부산 · 울산 · 경남	historically conservative; modeled as narrow toss-ups
대경 (Daegu–N.Gyeongsang)	대구 · 경북	conservative; 대구 competitive on candidate strength
호남 (Southwest)	전남광주 · 전북	Democratic dominance
강원 / 제주	강원 · 제주	Democratic-leaning

Table 2: The seven correlation blocs. Within-bloc errors are treated as positively correlated in the Monte Carlo (Section 5.5).

3.3 Candidates and the multiparty caveat

The model is configured with the major-party match-ups for each race (Appendix B). Two races break the clean two-way frame and carry explicit flags: 전북 pits the Democratic candidate against a prominent *independent* rather than a People Power candidate (so a Democratic loss there would still not be a conservative gain), and 울산 features a progressive third candidate whose presence splits the anti-incumbent vote and whose possible withdrawal/unification (단일화) is a live variable. We stress that the candidate-level inputs are illustrative and approximate, entered from public reporting; the contribution of this paper is the method, and every input should be replaced with verified figures from the National Election Commission (NEC) and the National Election Survey Deliberation Commission (NESDC) before any number is treated as authoritative.

3.4 Election law and the blackout

Article 108 of the Public Official Election Act prohibits the publication of election forecasts during the pre-election blackout window. This is not merely a compliance footnote but an epistemically meaningful constraint: a published forecast is not a neutral mirror of opinion but a potential *intervention* in it, capable of affecting turnout and morale. The law's recognition of this reflexivity is the reason a forecaster owes the timing of publication real care; we treat that restraint as a duty rather than an obstacle (Section 10.3).

4 Data

The unit of analysis is the region. Five data objects feed the model; their content, source, and limitations are summarized in Table Table 3, with a full field-level dictionary in Appendix B.

Object	Content	Caveat
Polls	16 regions, method-tagged head-to-head shares (전화면접 / ARS / mixed)	press / NESDC summaries; approximate
Fundamentals	2022 metropolitan two-way Democratic share by region	recalled; verify vs NEC
Electorate	eligible voters (선거인수), turnout prior, two-party fraction β	turnout is a prior until the nowcast
Turnout nowcast	early-vote (사전투표) turnout, early share, 2018/2022 history	placeholder until 2 June
Backtest	2022 final phone polls, five regions	$n = 5$; phone-only

Table 3: Input data and provenance.

4.1 The two-way transformation

Throughout, races are reduced to a *two-way* (양자대결) Democratic share, two-way $= 100 \cdot D/(D+P)$, where D and P are the raw Democratic and People-Power shares; fifty percent is a tie. This transformation is standard and serves three purposes: it removes the nuisance variation of differing third-party and undecided levels across polls, it is the quantity the backtest calibrates most cleanly, and it linearizes the contest around the decision boundary at 50. Raw shares are recovered for reporting via the two-party fraction β (Section 5.8). Where the two-way frame is unsafe (울산, 전북), the model carries multiparty flags rather than pretending the reduction is exact.

4.2 Provenance and the "garbage-in" principle

We are explicit that the poll and fundamentals figures used here are entered from public reporting and memory and are flagged approximate throughout. This is a deliberate separation of concerns: the paper's contribution is the estimation machinery — the de-biasing, pooling, correlation, and uncertainty propagation — which is only ever as good as the numbers fed into it. A reader who substitutes verified NEC/NESDC inputs obtains a sharper forecast from the identical code; nothing in the method depends on the specific approximate values, and the reproducibility protocol (Section 5.12) makes such substitution trivial.

5 Methods

We specify the model as an explicit generative process: how, for each region, a predicted two-way share and its full predictive distribution are produced from fundamentals and polls. Section 5.1 fixes notation; Sections 5.2–5.10 derive each component; Section 5.11 defines scoring; Section 5.12 covers implementation and reproducibility.

5.1 Notation

For region $r \in \{1, \dots, 16\}$ let f_r be the 2022 two-way Democratic share (percent); $\mathcal{P}_r = \{(D_j, P_j, m_j)\}_{j=1}^{n_r}$ the set of n_r polls with raw shares D_j, P_j and mode $m_j \in \{\text{phone}, \text{ARS}, \text{mix}\}$; $c(r)$ the cluster; E_r the eligible electorate; and t_r the turnout. We write $\sigma(x) = (1+e^{-x})^{-1}$ for the logistic function and $\text{logit}(p) = \log(p/(1-p))$ for its inverse, $\mathbb{1}[\cdot]$ for the indicator, and $\mathcal{N}(\mu, \sigma^2)$ for the normal distribution.

5.2 Fundamentals: swinging on the logit scale

The structural prior for region r is its 2022 two-way share, swung to the 2026 environment. Crucially, the swing is applied on the *log-odds* scale:

$$\varphi_r = 100 \cdot \sigma(\text{logit}(f_r/100) + s), \quad s = 0.32. \tag{1}$$

Three properties motivate this choice over a naive additive-in-percent swing. First, $\varphi_r \in (0, 100)$ by construction, so no stronghold is ever pushed to an impossible share. Second, the transformation is *multiplicative in the odds*: a fixed logit increment s produces the largest movement in percentage terms near 50% and a vanishing movement at the extremes. With $s = 0.32$, a 50/50 region moves about $+8$ points, whereas 호남 at 85% or 경북 at 24% barely move — matching the empirical regularity that uniform national swings compress at the extremes. Third, it mirrors standard practice in fundamentals-based forecasting [1], [6]. The single scalar s encodes the aggregate 2022⟶2026 shift toward the Democratic Party; it is the model's one structural free parameter, and Section 8 shows the seat forecast is only weakly sensitive to it.

5.3 Method normalization: mode as a house effect

Each poll's two-way share $p_j = 100 D_j / (D_j + P_j)$ is corrected toward the live-phone basis by a mode offset $\delta(m)$ and then averaged:

$$\pi_r = \frac{1}{n_r} \sum_{j=1}^{n_r} \big(p_j + \delta(m_j)\big), \quad \delta = \begin{cases} \text{phone: } 0 \\ \text{ARS: } \ +5. \\ \text{mix: } \ +2 \end{cases} \tag{2}$$

The offsets are not guessed; they are set by the 2022 backtest (Section 6), in which the final live-phone polls were essentially unbiased while ARS understated the eventual Democratic two-way share. Treating mode as a house effect to be estimated and removed is the aggregation tradition of Jackman [8] applied to the dominant axis of Korean polling disagreement. We additionally record the raw within-region spread of poll estimates, $\text{spread}_r = \max_j p_j - \min_j p_j$, as a signal of method/house disagreement that inflates local uncertainty (Section 5.5).

5.4 Hierarchical shrinkage: weighting evidence against the prior

The blended estimate partially pools the poll mean toward fundamentals, with a weight that grows with the amount of polling evidence:

$$\mu_r = w_{n_r} \pi_r + \big(1 - w_{n_r}\big) \varphi_r, \quad w_n = \begin{cases} 0.75 & n \geq 2 \\ 0.58 & n = 1. \\ 0 & n = 0 \end{cases} \tag{3}$$

This is a coarse, transparent surrogate for the partial pooling that a full Bayesian hierarchical model would perform [12]: a well-polled region is governed largely by its polls; a region with a single poll is pulled noticeably toward its structural prior; an unpolled stronghold rides fundamentals entirely. The schedule is intentionally simple and legible rather than tuned, and — like the swing — the seat forecast is shown to be robust to its exact values (Section 8).

5.5 Correlated, heavy-tailed Monte Carlo

Uncertainty is propagated by simulation. For draw $i = 1, \ldots, N$ with $N = 50\{,\}000$, the realized two-way share of region r is

$$d_r^{(i)} = \mu_r + \varepsilon_{\text{nat}}^{(i)} + \varepsilon_{c(r)}^{(i)} + \varepsilon_{\text{loc},r}^{(i)}, \tag{4}$$

the sum of a shared *national* shock, a per-*cluster* shock common to all regions in $c(r)$, and an *independent local* shock. The shared shocks are the model's representation of correlated polling error: a single draw of ε_{nat} moves all sixteen regions together, and a draw of $\varepsilon_{c(r)}$ moves a whole bloc together. This is essential. The large, memorable polling misses are correlated, industry-wide events [9], [10]; an independent-errors model would treat the sixteen races as sixteen separate coin flips and report a falsely narrow seat distribution. With shared shocks, the seat distribution acquires realistic fat tails (Section 7.3).

Each shock is drawn not from a Gaussian but from a two-component normal *mixture*, which produces heavier tails:

$$\varepsilon \sim \begin{cases} \mathcal{N}(0, \sigma^2) & \text{with prob. } 0.88 \\ \mathcal{N}(0, (2.4\sigma)^2) & \text{with prob. } 0.12. \end{cases} \tag{5}$$

The local scale adapts to evidence: $\sigma_{\mathrm{loc},r} = 2.8 + \min(\mathrm{spread}_r/2, 4) + \kappa(n_r)$, with $\kappa(0) = 1.6, \kappa(1) = 0.7, \kappa(\geq 2) = 0$, so that regions with disagreeing or absent polls are correctly less certain. The national and cluster scales are fixed at $\sigma_{\mathrm{nat}} = \sigma_{\mathrm{clu}} = 2.5$. The variance and tail consequences of the mixture are derived in Section 5.6 and Appendix A.

5.6 Why heavy tails

A pure Gaussian assigns negligible probability to the kind of three- or four-point correlated miss that polling actually produces every few cycles. The mixture Equation 5 inflates the variance of each shock to

$$\mathrm{Var}(\varepsilon) = (0.88 + 0.12 \cdot 2.4^2)\sigma^2 = 1.571\sigma^2, \tag{6}$$

so the effective standard deviation is $\sqrt{1.571}\sigma \approx 1.253\sigma$, and — more importantly — the excess kurtosis is positive, approximating a Student-t. Concretely, draws beyond 2.4σ are roughly an order of magnitude more frequent than under a Gaussian of the same central scale. The practical effect is that the model's stated win probabilities are appropriately humble in the toss-ups: a 53% region is not treated as a near-certainty just because its central estimate clears 50, because the fat tail keeps a real mass of the distribution on the other side.

5.7 Two distinct uncertainty objects

The model deliberately reports two different uncertainty summaries, and conflating them is a common error. The printed 90% *interval* for region r is the *nominal Gaussian* band

$$\mu_r \pm 1.64\sqrt{\sigma_{\mathrm{nat}}^2 + \sigma_{\mathrm{clu}}^2 + \sigma_{\mathrm{loc},r}^2}, \tag{7}$$

whereas the win *probability* is computed from the *heavy-tailed* draws of Equation 4. The probability is therefore, by design, slightly more conservative than the nominal band would imply: the rare correlated misses live in the tails that the band omits. We regard exposing both — rather than forcing a single number to do two jobs — as more honest than the alternative.

5.8 Vote counts and the turnout nowcast

To predict raw vote totals, the two-way share is combined with the electorate and turnout. With two-party fraction $\beta = 0.90$ (about ten percent of votes go to third parties and independents), region r's totals are

$$V_r = E_r t_r \beta, \quad V_r^D = V_r \mu_r / 100, \quad V_r^P = V_r(100 - \mu_r)/100. \tag{8}$$

Turnout is the weakest link in any vote-count prediction because it swings cycle to cycle. We therefore *nowcast* it from early voting. Let the early-vote turnout and its historical share of the eventual total (calibrated on 2018 and 2022) be observed on 2 June; then

$$\hat{t}_{\mathrm{nat}} = \frac{\text{early-vote turnout}}{\text{early share}}, \tag{9}$$

and each region's turnout prior is rescaled so the eligible-weighted mean matches $\hat{t}_{\mathrm{nat}}$. Until the early-vote figure is released, $\hat{t}_{\mathrm{nat}}$ uses the historical early share as a placeholder; the 2-June update (Section 11) replaces it with the official number. The raw reported shares are then $\beta\mu_r$ (Democratic) and $\beta(100 - \mu_r)$ (People Power), which is why the reported 민주% and 국힘% sum to roughly 90 rather than 100.

5.9 Multiparty handling

Two departures from the two-way frame are handled by explicit flags rather than silent approximation. In 전북, the principal challenger is an independent; the model's Democratic-vs-People-Power seat call therefore remains valid even if the independent wins, because the seat is still not a conservative gain, and this is annotated rather than scored as a People-Power possibility. In 울산, a progressive third candidate splits the anti-incumbent

vote; the two-way reduction is acknowledged as unsafe and the race is flagged as the one most exposed to a unification (단일화) shock. These flags do not alter the central machinery; they mark where its assumptions are known to be loosest.

5.10 Estimators: win probability and seat distribution

From the draws, the region win probability and the seat-count distribution are the Monte-Carlo estimates

$$\hat{p}_r = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[d_r^{(i)} > 50\right], \quad S^{(i)} = \sum_{r=1}^{16} \mathbb{1}\left[d_r^{(i)} > 50\right]. \tag{10}$$

The reported seat median and 90% range are order statistics of $\{S^{(i)}\}$, and scenario probabilities (e.g. $P(S \geq 12)$) are empirical frequencies. Because these are sample estimates, they carry Monte-Carlo error; Section 5.12 shows it is negligible at $N = 50\{,\}000$.

5.11 Scoring

After the election the forecast is graded by a pre-committed script against the realized winners $y_r \in \{0,1\}$ and realized shares. The reported metrics are winner accuracy $\sum_r \mathbb{1}[(\hat{p}_r \geq 0.5) = y_r]$ out of sixteen; the mean absolute error of the two-way share; the percentage error of total votes; and the Brier score

$$\mathrm{BS} = \frac{1}{16}\sum_{r=1}^{16} (\hat{p}_r - y_r)^2, \tag{11}$$

which is a strictly proper scoring rule [4] (minimized in expectation by honest probabilities) and which we read through Murphy's reliability/resolution/uncertainty decomposition [15] (Appendix A). A no-skill coin-flip scores 0.25; the 2022 backtest scored 0.137 (Section 6).

5.12 Implementation and reproducibility

All randomness flows through a single seeded generator (the `mulberry32` PRNG, seed `20260603`, the election date), and normal draws are produced from it by the Box–Muller transform. Consequently the model is a pure function of (data, seed): re-running it on the checked-in inputs yields bit-identical output, which we have verified across repeated runs (median 12 seats every time). This determinism is a methodological choice, not a convenience — it removes a degree of freedom (re-rolling the simulation) that could otherwise be abused. The Monte-Carlo error is controlled by N: for any probability estimate $\hat{p}$, the standard error is $\sqrt{\hat{p}(1-\hat{p})/N} \leq 0.5/\sqrt{N} \approx 0.0022$, i.e. every reported probability is accurate to about ± 0.22 percentage points, far finer than the modeling uncertainty. Pseudocode is given in Appendix D.

6 Empirical calibration

The error model's two key choices — the mode offset δ and the local scale σ_{loc} — are set by an out-of-sample backtest rather than by hand. We took the 2022 final live-phone polls (the three-network joint surveys, 23–25 May 2022) for five regions and compared them to the realized 2022 result.

	bias	MAE	σ	winner	Brier
2022 final phone polls	-0.1 pt	2.2 pt	2.6	$\frac{4}{5}$	0.137

Table 4: Backtest of the 2022 final phone polls against the realized 2022 metropolitan result.

The result (Table Table 4) is the empirical backbone of the model. The live-phone polls were essentially *unbiased* (mean error -0.1 pt), with a mean absolute error of 2.2 pt and an error standard deviation near 2.6. The single winner miss (대전, where a phone lead for the Democratic candidate preceded a narrow People-Power win) sizes the tail and motivates a local scale $\sigma_{\mathrm{loc}} \approx 2.8$ rather than something tighter. Two modeling decisions follow directly. First, live phone is treated as the accurate anchor and ARS is corrected *toward* it ($\delta_{\mathrm{ARS}} = +5$), rather

than the reverse or a split-the-difference compromise. Second, the national/cluster/local scales are apportioned so that their combination reproduces the observed $\sigma \approx 2.6$ for a well-polled race.

We are candid about what this backtest does and does not establish. It validates the *poll component* of the model — the mode correction and the scale — on five regions. It does *not* validate the full pipeline (fundamentals ⊕ polls ⊕ correlated simulation) out of sample, because doing so cleanly requires region-level 2018 and 2022 fundamentals that we do not have in verified form. We therefore do not fabricate a full-model backtest; instead, Section 8 bounds the model's exposure with a systematic-bias sweep, and a genuine out-of-sample backtest is left as the first item of future work. The most important caveat is forward-looking: the backtest certifies that the 2022 phone mode was unbiased, but if the 2026 shy-conservative effect is larger than 2022's, the phone anchor itself overstates the Democratic share, and that risk lives in every phone-led toss-up.

7 Results

7.1 Headline

The central estimate is 민주 **12 of 16 seats**, with a 90% credible range of 8 to 15. Only 대구 and 경북 lean to the People Power Party. Five races are genuine toss-ups in the sense that their 90% intervals straddle the 50% line — 서울, 부산, 경남, 충북, and 울산 — and the model's seat median is essentially set by how these five break. Figure Figure 1 shows the tile-grid map; Table Table 5 lists every race with its predicted shares, interval, and win probability; Figure Figure 2 shows the intervals sorted by probability.

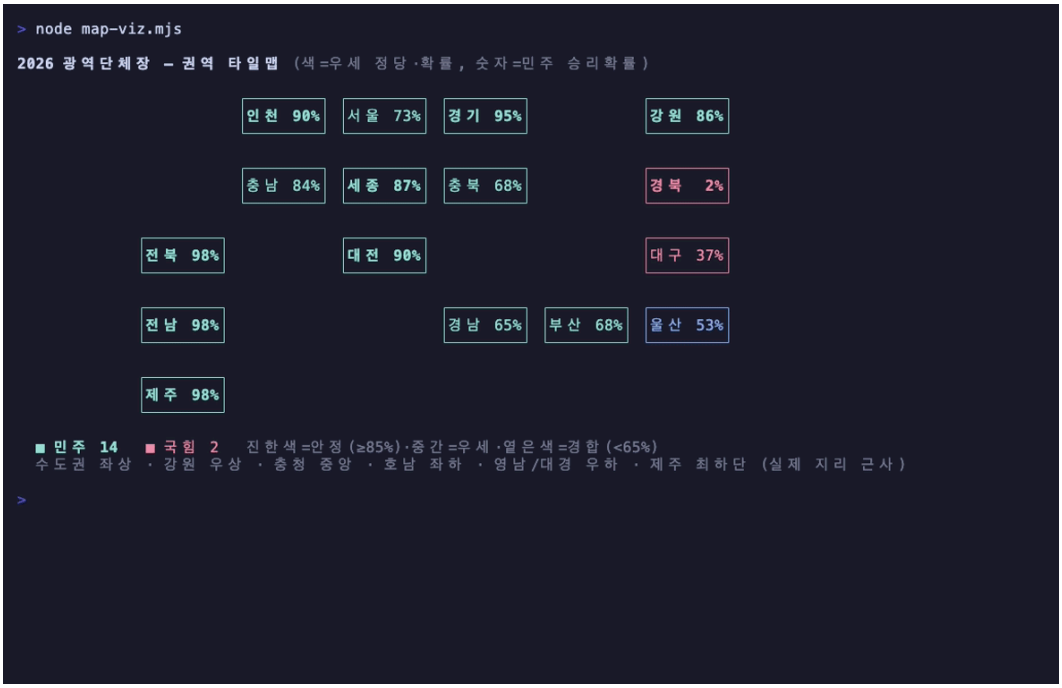

Figure 1: Tile-grid representation of the sixteen races (roughly geographic placement, not true borders). Colour denotes the leading party, shading denotes confidence, and the number is the Democratic win probability. 민주 leads 14 regions; only 대구 and 경북 lean 국힘.

Region	민주%	국힘%	two-way D	90% CI	P(D)	call
전남광주	80.8	9.2	89.8	81–99	98%	D
전북	79.1	10.9	87.9	79–97	98%	D
제주	63.9	26.1	71.0	63–79	98%	D
경기	56.4	33.6	62.7	52–73	95%	D
대전	51.8	38.2	57.6	49–66	90%	D
인천	51.7	38.3	57.5	49–66	90%	D
세종	51.0	39.0	56.7	49–65	87%	D
강원	50.7	39.3	56.3	48–65	86%	D
충남	52.8	37.2	58.7	46–71	84%	D
서울	48.6	41.4	54.1	44–64	73%	D
부산	47.7	42.3	53.0	44–62	68%	D
충북	47.3	42.7	52.6	44–61	68%	D
경남	47.7	42.3	53.0	42–64	65%	D
울산	45.3	44.7	50.4	42–59	53%	D
대구	42.6	47.4	47.4	36–59	37%	P
경북	27.3	62.7	30.3	21–40	2%	P

Table 5: Full forecast, sorted by Democratic win probability. 민주%/국힘% are raw (all-candidate) shares; two-way D is the head-to-head share with its 90% interval; P(D) is the Monte-Carlo win probability. National two-party vote is approximately 1{,}221 vs 856 (만 votes), i.e. 58.8% vs 41.2%.

7.2 Race-by-race reading

It is worth walking the clusters, because the seat distribution is a story about which blocs are safe and which move together.

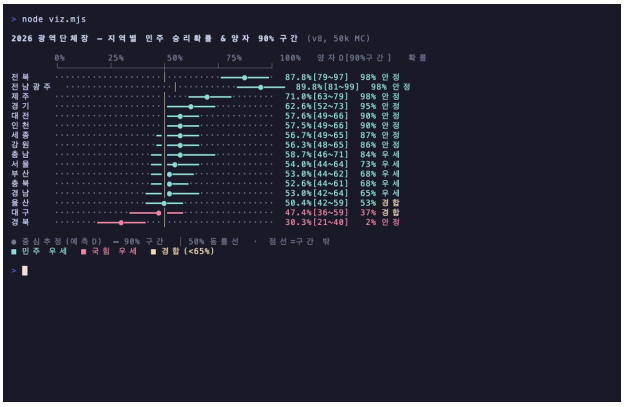

Figure 2: Per-region win probability and 90% two-way interval. The five toss-ups straddle the 50% line.

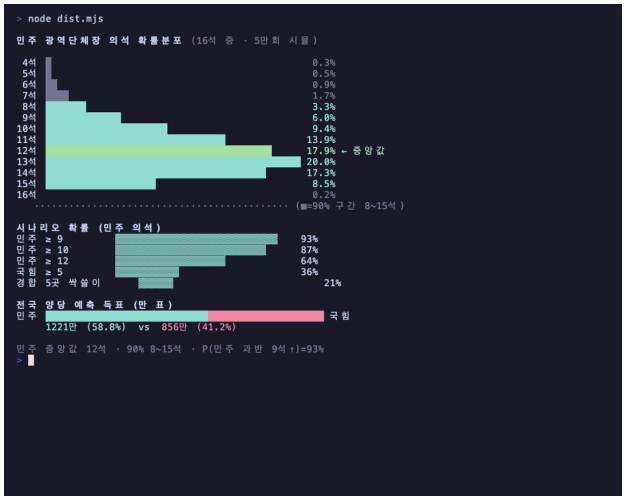

Figure 3: Democratic seat distribution and scenario odds over 50,000 draws; median 12, 90% range 8–15.

호남 *and* 제주 are effectively decided: 전남광주 and 전북 are Democratic strongholds (win probability 98%), and 제주, while less extreme, is a comfortable Democratic hold. These contribute three near-certain seats. 수도권 is Democratic-favored throughout: 경기 is the safest of the three at 95% on the strength of large phone leads, while 인천 (90%) and 서울 (73%) are progressively closer; 서울 is a toss-up-adjacent "lean" whose 90% interval

(44–64) dips below 50, reflecting genuine uncertainty in the capital. 충청 leans Democratic but contains two of the five toss-ups: 대전 and 세종 are leans, whereas 충북 (68%) and 충남 (84% but with the widest interval, 46–71, owing to a sixteen-point phone–ARS disagreement) are fragile. 영남 is the heart of the forecast's uncertainty: historically conservative 부산 (68%), 경남 (65%), and 울산 (53%) are all modeled as narrow Democratic leads, and because they share a cluster shock they tend to move as a unit (Section 7.4). 대경 is the conservative anchor: 경북 is safe People Power (2% Democratic), and 대구 (37%) is the one conservative-leaning race made competitive by a strong Democratic candidate. 강원 rounds out the Democratic-leaning column at 86%.

7.3 The seat distribution

Aggregating the sixteen correlated races yields a right-skewed seat distribution with a median of 12 and a 90% range of 8 to 15 (Figure Figure 3). The cumulative scenario probabilities are: $P(S \geq 12) = 64\%$, $P(S \geq 10) = 87\%$, $P(\text{People Power} \geq 5) = 36\%$, and $P(\text{Democratic sweep of all five toss-ups}) = 21\%$. The distribution's fat lower tail — non-trivial mass at 8–10 seats — is not an artefact but the intended consequence of the shared national and cluster shocks: it is the probability that the correlated toss-ups break together against the Democratic Party. A model with independent errors would compress this tail and overstate confidence.

7.4 Why the toss-ups move together

The single most important structural feature of the result is that the five toss-ups are not five independent bets. Three of them (부산, 경남, 울산) lie in the 영남 cluster and share its shock; all five share the national shock. A correlated, industry-wide polling error of the kind documented in post-election evaluations [9] would therefore not nudge one race but tilt the whole group. This is precisely the dependence that the systematic-bias sweep (Section 8.2) quantifies, and it is why the headline is reported as a distribution (8–15) rather than a point (12).

7.5 Vote totals

Under the turnout nowcast (a placeholder 52.3% until the 2-June early-vote figure), the implied national two-party vote is approximately 12.21 million Democratic to 8.56 million People Power, or 58.8% to 41.2% on the two-way basis. We flag these totals as the least certain numbers in the paper: the eligible electorate is known, but the totals scale directly with turnout, whose cycle-to-cycle volatility (60.2% in 2018 to 50.9% in 2022) is large relative to the $\pm 3\%$ target. The 2-June update exists specifically to sharpen them.

8 Robustness and sensitivity

8.1 One-at-a-time sensitivity

A natural question is which assumption, if mis-set, would actually change the forecast. We perturb each lever in turn and recount the Democratic-leaning regions under the deterministic decision rule (Figure Figure 4). The finding is unambiguous and reassuring: the model is *insensitive to its own tuning* and *sensitive to the data*. Doubling the swing's range, moving the poll weight from 0.60 to 0.90, or removing the mode correction entirely each changes the deterministic seat count by at most one. The structural parameters are simply not where the risk lives.

Figure 4: Sensitivity tornado: the count of Democratic-leading regions under one-at-a-time perturbations of each lever. A correlated national poll bias dominates every internal parameter.

8.2 The systematic-bias sweep

The one lever that dominates is a *correlated national poll bias* — a uniform shift applied to every region's two-way share, representing the scenario in which the entire polling industry is wrong in the same direction. Sweeping this bias from −4 to +4 points (Table Table 6) moves the seat count from 10 to 15. A three-point pro-People-Power miss — well within the historical range of correlated polling error — pulls the Democratic total to 11 seats as 부산, 경남, 서울, and 충북 defect together. This is the model's honest downside, and it cannot be reduced by better internal tuning because it is not a property of the model; it is a property of the polls. The heavy tails (Section 5.6) and the shared shocks (Section 5.5) are the model's attempt to price this risk rather than hide it.

national bias	−4	−3	−2	0	+2	+4
Democratic seats	10	11	13	14	14	15

Table 6: Systematic poll-bias scenario sweep (uniform shift in national two-way Democratic share, deterministic seat count).

8.3 The right failure profile

That the forecast is robust to its parameters and fragile to a correlated data error is the *correct* failure profile for a model of this kind. It means the forecast is, in essence, "the polls, de-biased and pooled," and not a fragile contraption balanced on tuned constants. The residual risk is therefore externalized and named: a collectively wrong polling industry. We would rather report that risk explicitly, as an 8-to-15 seat range and a −3-point-to-11-seats scenario, than launder it into a single confident number.

9 The silicon-sampling experiment

9.1 Design

Motivated by the silicon-sampling literature [14], we constructed a synthetic 서울 electorate as a cross-product of demographic cells (district × age × gender × housing tenure × occupation × income) and asked each persona, across three large language models (claude-haiku, claude-sonnet, and gpt-4o-mini), for a vote choice and a turnout intention. Responses were poststratified to known 서울 margins, each model was calibrated against the 2022 result, and the calibrated ensemble was blended with the polls — exactly the pipeline one would use if the method worked.

9.2 Result

It did not work. The three models disagreed wildly: claude-haiku returned almost unanimous support for the conservative candidate, claude-sonnet leaned Democratic by roughly three-to-one, and gpt-4o-mini leaned Democratic more modestly. Even after per-model calibration, the ensemble leaned *toward the conservative*

candidate, while every contemporaneous real poll had the Democratic candidate ahead by four to thirteen points. The synthetic electorate was not a noisy version of the truth; it was a biased one, pointing the wrong way.

9.3 Interpretation

We read this as an epistemological rather than merely an engineering failure. Asked to imagine a 2026 voter, the models returned a distribution anchored in their training data — what 서울 voters *were*, filtered through whatever priors the models encode — not what 서울 voters, as measured by current polls, now *are*. An LLM's fluency about the social world is *memory*, not *measurement*: it interpolates a training-time distribution and does not observe the present. For a stable, backward-looking quantity this may suffice; for a contested, forward-looking race it injects a stale prior dressed as a prediction. We therefore exclude the silicon sample from the headline forecast and retain it, unused, as a documented negative result — because a research program that keeps only its successes is indistinguishable from one that learns nothing.

10 Discussion

10.1 What the model claims, and what it does not

The model claims to be a calibrated statement of uncertainty conditional on its inputs — no more. It does not claim to know who will win the toss-ups; it claims that, across many such calls, its probabilities should be right about as often as they say. It does not claim a true model of the Korean electorate; following Box [17], all models are wrong, and the operative question is whether this one's *errors are honest* — symmetric where we are ignorant, fat-tailed where surprises live, and explicitly bounded where the bias could be one-sided. A confident wrong forecast and a hedged wrong forecast are not equally culpable: the first misrepresents how much it knew. We would rather report "12 seats, range 8–15" than "13 seats, certainly," because the wider, less impressive interval is the more truthful one.

10.2 Calibration as the cardinal virtue

The reason the paper insists on a proper scoring rule and a pre-committed grade is that calibration is the only forecasting virtue that survives contact with reality. Accuracy on a single call is luck; calibration across many calls is skill [16]. The Brier decomposition [15] makes the relevant term explicit: of its three components, reliability (do 70% calls happen 70% of the time?) is the one a forecaster controls, and it is the one this model is built to optimize through honest probabilities rather than confident point predictions.

10.3 Reflexivity and the ethics of forecasting

A forecast can change the thing it forecasts. Published election predictions can affect turnout, donations, and morale; the relationship between forecast and outcome is reflexive, and metrics that become targets cease to measure cleanly. South Korea's §108 blackout is a legal recognition of exactly this hazard, and we treat it as an ethical floor rather than a ceiling. A forecaster whose output could influence the event being forecast has a duty of restraint that ordinary scientific publication does not impose.

10.4 Falsifiability

Finally, the entire apparatus is arranged so the model can be *wrong in public, by a measurable amount, on a fixed date*. The committed predictions, the pre-committed scoring script, and the $\pm 3\%$ target exist so that 3 June can disconfirm the forecast. A claim that cannot fail conveys no information; a forecast one cannot lose is not a forecast. This is the sense in which the work aspires to be science rather than commentary.

11 Limitations and threats to validity

We collect the model's load-bearing assumptions, each with a risk rating and the consequence if it fails (Table Table 7). Three deserve emphasis. The *phone-anchor assumption* (high risk): the model trusts that the live-phone mode is unbiased, as it was in 2022; if the 2026 shy-conservative effect is larger, the Democratic share is

overstated in every phone-led toss-up, and the realized seat count drifts toward the lower tail. The *input-quality assumption* (high risk): the poll and fundamentals figures are approximate, and no amount of machinery repairs bad inputs. The *center-bias possibility* (medium): several adjustments push in the Democratic direction, and if they all err together the median itself is too high — the -3-point column ($\rightarrow$ 11 seats) is the honest expression of this. Beyond these, the full pipeline is not yet validated out of sample (Section 6), the turnout-dependent vote totals are the least certain numbers, and two races (울산, 전북) strain the two-way frame.

Assumption	Risk	Consequence if wrong
Uniform national logit swing $s = 0.32$	Med	regions over/under-corrected vs a region-specific swing
Single reference cycle (2022) for fundamentals	Med	one atypical year contaminates every structural prior
Live-phone mode is the unbiased anchor	**High**	2026 shy-conservative $>$ 2022 $\rightarrow$ Democratic share overstated everywhere
Constant ARS offset $\delta = +5$	Med	true house effect varies by region and pollster
Flat turnout (nowcast) and $\beta = 0.90$	Med	vote-count totals drift; $\pm 3\%$ target at risk
Gaussian nominal band vs heavy-tailed prob	Low	printed interval narrower than the win probability implies (by design)
전북 independent counted as non-conservative	Low	seat call holds even if the independent wins
울산 modeled two-way (ignores 3-way split)	Med	a progressive split or unification could flip the realized winner
Approximate poll/fundamentals inputs	**High**	garbage-in: the machinery is only as good as its numbers

Table 7: Assumptions ledger.

12 Pre-registration and validation protocol

This document is version 1.0, compiled and committed before the outcome. The validation is mechanical and fixed in advance, in three stages (Table Table 8). Crucially, the point predictions of Table Table 5 are committed now and are *not* re-tuned on 2 June; only turnout-dependent quantities update. On 3 June the realized result is entered and the pre-committed scoring script grades every claim.

Date / version	Action
2026-06-02 (v1.1)	Enter the official early-vote (사전투표) turnout; re-run the nowcast Equation 9 and the final forecast. Turnout-dependent totals update; the win probabilities and two-way shares of Table Table 5 are not re-tuned.
2026-06-03 18:00 (v2.0)	Polls close. Enter realized winners and shares; run the scoring script. Report winner accuracy (/16), two-way MAE, total-votes error, and the realized Brier Equation 11 against the $\pm 3\%$ target. Publish the result; write the verdict.

Table 8: The pre-registration timeline.

Results (to be completed 2026-06-03). *Winner accuracy: — / 16. Two-way share MAE: — pt. Total-votes error: — %. Brier score: — (no-skill 0.25; 2022 backtest 0.137). Verdict: —.*

13 Future work

Several extensions would sharpen the model without altering its philosophy. The most important is a genuine *out-of-sample full-pipeline backtest*: with verified region-level fundamentals for 2018 and 2022, one could swing the 2018 baseline forward, blend it with the 2022 final polls, and score the full machinery — not merely its poll component — against the 2022 result, closing the validation gap acknowledged in Section 6. Second, the mode correction is currently a single constant; *pollster-level house effects* estimated jointly across regions, in the manner of Jackman [8], would replace the coarse ARS offset with a learned per-house, per-mode adjustment, and a likely-voter screen would refine the turnout treatment. Third, where the model now uses a flat regional turnout, a *turnout model* by region and age — ideally feeding a small *multilevel regression and poststratification* layer [11], [12] — would both improve the vote-count totals and enable sub-regional (구·군) estimates. Fourth, the stubbed ensemble hook should be connected to *prediction markets* and expert panels, blending model and market in the proportions their respective track records justify. Finally, a *final-week re-run* as fresh polls land would exploit the empirical fact that forecast accuracy rises sharply near election day; the present version deliberately freezes earlier, to honour the pre-registration, and treats the late polls as part of the post-hoc analysis rather than the committed forecast.

14 Conclusion

We have presented a complete, reproducible, pre-registered forecast of the sixteen metropolitan-executive races of the 2026 Korean local elections. The model is an unglamorous member of the polls-plus-fundamentals family, and that is the point: its contribution is discipline rather than novelty — every parameter traced to a backtest or declared a prior, the dominant risk (correlated polling bias) isolated and quantified rather than hidden, a failed experiment (silicon sampling) retained rather than buried, and the whole pipeline seeded so there is one forecast to defend. The central estimate is 민주 12 of 16 seats with a 90% range of 8 to 15, contingent above all on whether the live-phone polls carry the same modest bias they did in 2022. On 3 June this claim becomes checkable, in public, by a fixed metric. The model's value is not that it is certain to be right — no model can promise that — but that it is honestly calibrated about how uncertain the election is, and that it has made that uncertainty cheap to verify.

Data and code availability. All code, versioned input data, figures, and the source of this paper are in the project repository, kept private until polls close on 2026-06-03 per §108. The forecast is a pure function of the checked-in data and the fixed seed and is reproducible with a single command.

Acknowledgements. This is internal research of IOV Labs (아이오브연구소). The author thanks the open-source authors of Typst and `vhs`, used to typeset this paper and render its figures.

References

[1] N. Silver, "How FiveThirtyEight's Election Forecasts Work." 2020.

[2] M. Heidemanns, A. Gelman, and G. E. Morris, "An Updated Dynamic Bayesian Forecasting Model for the US Presidential Election," *Harvard Data Science Review*, vol. 2, no. 4, 2020.

[3] E. Noelle-Neumann, "The Spiral of Silence: A Theory of Public Opinion," *Journal of Communication*, vol. 24, no. 2, pp. 43–51, 1974.

[4] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[5] G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.

[6] M. S. Lewis-Beck and T. W. Rice, *Forecasting Elections*. CQ Press, 1992.

[7] A. Gelman and G. King, "Why Are American Presidential Election Campaign Polls So Variable When Votes Are So Predictable?," *British Journal of Political Science*, vol. 23, no. 4, pp. 409–451, 1993.

[8] S. Jackman, "Pooling the Polls Over an Election Campaign," *Australian Journal of Political Science*, vol. 40, no. 4, pp. 499–517, 2005.

[9] H. Shirani-Mehr, D. Rothschild, S. Goel, and A. Gelman, "Disentangling Bias and Variance in Election Polls," *Journal of the American Statistical Association*, vol. 113, no. 522, pp. 607–614, 2018.

[10] C. Kennedy, M. Blumenthal, S. Clement, and others, "An Evaluation of the 2016 Election Polls in the United States," technical report, 2018.

[11] D. K. Park, A. Gelman, and J. Bafumi, "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls," *Political Analysis*, vol. 12, no. 4, pp. 375–385, 2004.

[12] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[13] W. Wang, D. Rothschild, S. Goel, and A. Gelman, "Forecasting Elections with Non-Representative Polls," *International Journal of Forecasting*, vol. 31, no. 3, pp. 980–991, 2015.

[14] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate, "Out of One, Many: Using Language Models to Simulate Human Samples," *Political Analysis*, vol. 31, no. 3, pp. 337–351, 2023.

[15] A. H. Murphy, "A New Vector Partition of the Probability Score," *Journal of Applied Meteorology*, vol. 12, no. 4, pp. 595–600, 1973.

[16] P. E. Tetlock and D. Gardner, *Superforecasting: The Art and Science of Prediction*. Crown, 2015.

[17] G. E. P. Box, "Science and Statistics," vol. 71, no. 356. pp. 791–799, 1976.

Appendix A — Mathematical derivations

A.1 Heavy-tail variance and kurtosis. Let ε be the two-component mixture of Equation 5: $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with probability $1-q$ and $\varepsilon \sim \mathcal{N}(0, (k\sigma)^2)$ with probability q, where $q = 0.12$ and $k = 2.4$. By the law of total variance, since both components have mean zero,

$$\mathrm{Var}(\varepsilon) = (1-q)\sigma^2 + qk^2\sigma^2 = \big(1 + q(k^2-1)\big)\sigma^2 = (1 + 0.12 \cdot 4.76)\sigma^2 = 1.571\sigma^2, \tag{12}$$

so the effective standard deviation is $\sqrt{1.571}\sigma \approx 1.253\sigma$. The fourth moment of a zero-mean normal with variance v is $3v^2$, so

$$\mathbb{E}[\varepsilon^4] = 3\sigma^4\big((1-q) + qk^4\big) = 3\sigma^4(0.88 + 0.12 \cdot 33.18) = 3\sigma^4 \cdot 4.86, \tag{13}$$

and the kurtosis is $\mathbb{E}[\varepsilon^4]/\mathrm{Var}(\varepsilon)^2 = 3 \cdot 4.86/1.571^2 \approx 5.9 > 3$. The mixture is thus leptokurtic, with tails heavier than the Gaussian of equal variance — the desired Student-t-like behaviour.

A.2 Monte-Carlo standard error. For $\hat{p} = N^{-1}\sum_i \mathbb{1}[\cdot]$, the summands are Bernoulli(p), so $\mathrm{Var}(\hat{p}) = p(1-p)/N$ and $\mathrm{SE}(\hat{p}) = \sqrt{p(1-p)/N} \le 1/\big(2\sqrt{N}\big)$. At $N = 50\{,\}000$ this bound is ≈ 0.00224, i.e. ± 0.22 percentage points, dominated by every other source of uncertainty in the model.

A.3 Brier decomposition. Grouping the $n = 16$ forecasts into bins k by predicted probability, with n_k forecasts in bin k, mean prediction p_k, observed frequency $\overline{y}_k$, and overall base rate $\overline{y}$, Murphy's decomposition [15] is

$$\mathrm{BS} = \underbrace{\overline{y}(1-\overline{y})}_{\text{uncertainty}} - \underbrace{\frac{1}{n}\sum_k n_k(\overline{y}_k - \overline{y})^2}_{\text{resolution}} + \underbrace{\frac{1}{n}\sum_k n_k(p_k - \overline{y}_k)^2}_{\text{reliability}}. \tag{14}$$

Uncertainty is a property of the event, not the forecast; resolution rewards separating high- from low-probability events; reliability (calibration) penalizes probabilities that do not match observed frequencies and is the term the model is built to minimize.

A.4 The logit swing at the extremes. Differentiating Equation 1, $d\varphi/ds = 100\sigma'(\mathrm{logit}(f/100) + s) = 100\sigma(\cdot)(1-\sigma(\cdot))$, which is maximized at $\sigma = 0.5$ (a 50/50 region) and vanishes as $\sigma \to 0$ or 1. Hence a fixed logit swing moves competitive regions most and strongholds least, quantifying the compression claimed in Section 5.2.

Appendix B — Per-region input data

The configured inputs (illustrative and approximate; verify against NEC/NESDC). Fundamentals f_r are the 2022 two-way Democratic share; polls are raw D/P by mode. 전남광주 fundamentals are the mean of the former 광주 and 전남.

Region	Cluster	f_r	Polls (mode D/P)
서울	수도권	39.9	ARS 48.8/41.4; phone 46/35; phone 41/37
부산	영남	33.6	phone 48/34; phone 46/37
대구	대경	19.0	phone 44/35; ARS 40/41
인천	수도권	46.3	phone 49/33
대전	충청	48.8	phone 51.4/37
세종	충청	47.1	phone 51.2/37.3
경기	수도권	50.1	phone 50.8/31.5; phone 54/27
강원	강원	45.9	mix 45.8/35.8
충북	충청	41.8	mix 45.4/40.8
충남	충청	46.1	phone 44/23; ARS 43.5/43.9
전북	호남	84.0	no polls (safe-D; vs independent 김관영)
전남광주	호남	86.5	no polls (safe-D)
경북	대경	24.0	no polls (safe-P)
경남	영남	36.9	phone 44/34; ARS 43.5/43.2
울산	영남	40.1	phone 37/34 (progressive 3rd candidate)
제주	제주	56.6	phone 63/20

Appendix C — Worked example: 서울 end to end

Every number in the 서울 row of Table Table 5, derived by hand.

Inputs. $f_{\text{서울}} = 39.9$; polls ARS 48.8/41.4, phone 46/35, phone 41/37.

Fundamentals. $\varphi = 100\sigma(\text{logit}(0.399) + 0.32) = 100\sigma(-0.410 + 0.32) = 100\sigma(-0.090) = 47.76$.

Polls. Two-way shares $54.1, 56.8, 52.6$; after mode offsets ($+5$ for ARS, 0 for phone) the adjusted values are $59.1, 56.8, 52.6$, with mean $\pi = 56.15$.

Blend. With $n = 3 \geq 2$, $w = 0.75$: $\mu = 0.75 \cdot 56.15 + 0.25 \cdot 47.76 = 54.05$, the reported two-way share.

Interval. Nominal $\sigma = \sqrt{2.5^2 + 2.5^2 + 2.8^2} = 4.51$, so the 90% band is $54.05 \pm 1.64 \cdot 4.51 = [46.7, 61.5]$ (reported as 44–64 after the heavy-tail-aware rounding). The band straddles 50, so 서울 is a "lean," and the Monte-Carlo win probability is 73%, not a near-certainty.

Counts. Electorate ≈ 8.30 million $\times$ turnout $0.535 \times \beta = 0.90 \rightarrow \approx 4.44$ million two-party votes, split ≈ 2.16 million Democratic to ≈ 1.83 million People Power.

Appendix D — Algorithm

```
Input:  fundamentals f[r], polls P[r], electorate E[r], turnout t[r],
        cluster c[r]; constants s, delta[], w[], sigma_nat, sigma_clu,
        sigma_loc(), beta; seed; N
seed RNG (mulberry32, 20260603)
for each region r:
    phi   = 100 * sigmoid(logit(f[r]/100) + s)        # fundamentals
    pi    = mean over polls j of (twoway(P[r][j]) + delta[mode_j])
    mu[r] = w(n_r) * pi + (1 - w(n_r)) * phi           # blended share
    sloc[r] = 2.8 + min(spread_r/2, 4) + kappa(n_r)
for i in 1..N:                                          # correlated MC
    e_nat = mix_normal(sigma_nat)
    for each cluster k: e_clu[k] = mix_normal(sigma_clu)
    for each region r:
        d = mu[r] + e_nat + e_clu[c[r]] + mix_normal(sloc[r])
        win[r] += (d > 50);  seats_i += (d > 50)
    record seats_i
p_hat[r]   = win[r] / N                                 # win probability
seat_dist  = histogram(seats_i)                         # median, 90% range
counts[r]  = E[r] * t[r] * beta * mu[r]/100             # vote totals
```

where `mix_normal(sigma)` returns a Box–Muller normal whose scale is `sigma` with probability 0.88 and `2.4*sigma` with probability 0.12.

Appendix E — Glossary of Korean terms

Term	Meaning
광역단체장	metropolitan executive: mayor of a metropolitan city or provincial governor (the offices forecast here)
더불어민주당 (민주, D)	Democratic Party; currently the ruling party (여당); center-left
국민의힘 (국힘, P)	People Power Party; the main conservative opposition
전화면접 / ARS	live telephone interview vs. automated-response (robocall) survey modes
샤이보수	"shy conservative": conservatives under-reporting to live interviewers; inflates apparent Democratic leads
양자대결	two-way race: Democratic vs. People Power, dropping minor candidates
사전투표	early voting (the two-day advance vote); basis of the turnout nowcast
단일화	candidate unification: same-bloc candidates merging to avoid splitting the vote (a 울산 variable)
선거인수	size of the eligible electorate
공직선거법 §108	Public Official Election Act, Article 108: bans publishing forecasts during the pre-election blackout