PoliSim

Electoral Simulation Platform

Technical White Paper — Methodology & Validation

Version 1.0 — May 2026

polisim.dev

Abstract — PoliSim is a three-stage computational platform for electoral simulation in the Italian proportional-uninominal mixed system (Rosatellum bis). Stage 1 applies a uniform swing model to official Ministry of Interior data (Eligendo) across 221 uninominal constituencies. Stage 2 refines seat allocation through a Multilevel Regression and Poststratification (MRP) model using PyMC 5.28.4 with a Dirichlet-Multinomial likelihood, calibrated on 252 observations across 13 regions. Stage 3 generates demographically targeted communication strategies by integrating psychographic profiles derived from ESS Round 11 (N=2,865), TRIPOL IT (N=1,231), and Meta Ad Library calibration data. Backtest on 14 Lazio constituencies (2022 general election) yields MAE 4.01 pp, RMSE 4.64 pp, and 14/14 winner accuracy. The MRP model outperforms an OLS baseline by 1.04 pp RMSE across six regional validation sets. Known limitations — including CI90 coverage of 29%, residual CDX bias of −3.33 pp, and the absence of proprietary polling data — are explicitly declared.

1. Introduction and Motivation

The Italian electoral system introduced by Law 165/2017 — known as Rosatellum bis — combines proportional and majoritarian allocation: 37% of seats in both chambers are assigned through single-member districts (collegi uninominali), while the remaining 63% are distributed proportionally among closed party lists. This hybrid structure creates non-linear interdependencies between national vote shares, local candidate strength, and coalition composition, making seat-level projections methodologically complex.

Existing forecasting tools available to Italian political actors and researchers typically fall into one of two categories: (i) national-level poll aggregators that report vote intentions but do not translate them into constituency-level seat distributions, or (ii) proprietary models maintained by major broadcast media outlets whose methodological

specifications are not publicly disclosed. A gap therefore persists between publicly accessible polling data and operationally useful seat-level projections.

PoliSim addresses this gap through a reproducible, open-methodology platform built on three computational stages. The design philosophy is grounded in three principles: methodological transparency, with all modeling assumptions and known limitations explicitly declared; empirical grounding, using only official institutional data sources (Eligendo, ISTAT, ESS) as inputs; and progressive validation, requiring multi-region backtesting before any public claims are made.

The platform was developed in the context of Q-Italia (qitalia.org), a civic technology organization applying quantitative political analysis to Italian public discourse. Q-Italia uses PoliSim as a live case study, demonstrating its capabilities on real political scenarios. This dual function — as both a research tool and a communications infrastructure — informs the design requirements: the platform must be simultaneously academically defensible and operationally deployable.

1.1 Research Questions

PoliSim is designed to address three operational questions:

- Q1 — Seat allocation: Given an observed or hypothetical national vote distribution, what is the expected seat distribution in the Camera dei Deputati and Senato della Repubblica?

- Q2 — Double majority verification: Does a given scenario produce a stable governing majority in both chambers simultaneously, accounting for the distinct demographic composition of the Senato electorate?

- Q3 — Message optimization: Given a policy or electoral scenario, what communication strategy maximizes persuasive reach across defined demographic segments?

1.2 Scope and Limitations

This white paper documents the methodology, data sources, and validation results for PoliSim version 2.1, deployed as of May 2026. The platform operates as a scenario analysis tool, not a predictive polling instrument. It does not

generate autonomous vote intention forecasts; rather, it translates externally provided vote distributions (from public polling aggregators or user-specified scenarios) into seat-level distributions with associated uncertainty estimates.

The distinction is methodologically significant: PoliSim can accurately model the electoral consequences of a given scenario but cannot independently forecast which scenario will materialize. Integration with external polling aggregators (planned as a subsequent development stage) would bridge this gap.

2. Data Sources

PoliSim draws on four primary data sources, each serving a distinct role in the modeling pipeline. A cardinal principle of the data architecture is reliance exclusively on official institutional sources with documented provenance, in order to ensure reproducibility and resist the credibility challenges associated with proprietary or undocumented datasets.

2.1 Eligendo — Ministry of Interior Electoral Archive

The primary input to Stage 1 is the official electoral results database maintained by the Italian Ministry of Interior (Ministero dell'Interno) through the Eligendo OpenData platform. For the 2022 general election, data covers all 221 uninominal constituencies at the constituency level (Camera: 147 collegi; Senato: 74 collegi), including votes by party, coalition composition, candidate names, and official counts.

Data are downloaded as ZIP archives from the Eligendo portal and processed by `polisim*build*collegi.py`, which parses constituency-level results and constructs the swing model baseline. A persistent data quality issue — documented and managed rather than hidden — is the inconsistency of comune name encoding across dataset versions (UTF-8/Latin-1 mojibake), particularly for bilingual municipalities in Alto Adige and Valle d'Aosta. A 98.4% match rate on comune names was achieved; unmatched records are excluded from the geographic join but retained in aggregate totals.

Local coalition anomalies — notably the Südtiroler Volkspartei (SVP) in Alto Adige and the Misto group in several constituencies — introduce approximately 17 seats of discrepancy between computed totals and official Wikipedia-

documented CDX results. This discrepancy is documented, explained, and does not affect the validity of the swing model for the 97% of constituencies where standard coalition compositions apply.

2.2 ISTAT 2021 Decennial Census

Demographic features for the MRP model (Stage 2) are derived from the ISTAT 2021 decennial census. Data are obtained through bulk download of sezione-level indicator files (R12*indicatori*2021_sezioni.xlsx for Lazio, containing 39,670 sezioni across five provinces) and aggregated to the constituency level using geographic join with official shapefile boundaries (Collegi Elettorali Basi Geografiche, MIT).

Eight demographic features are currently in active use in the MRP model: percentage of residents aged 15-34 (pct*giovani), percentage aged 65+ (pct*anziani), percentage with tertiary education (pct*alta*istr), percentage female (pct*femmine), employment rate (pct*occupati), percentage of foreign-born residents (pct*stranieri), and percentage from non-EU countries (pct*extra*ue), plus an electoral type indicator (*is*politiche).

A structural limitation of the current implementation is that ISTAT features are fully populated for only 14 Lazio constituencies; the remaining 238 constituencies receive a value of zero for the immigration-related features (pct*stranieri*z, pct*extra*ue_z). While this is a recognized suboptimum, it does not introduce bias for non-Lazio constituencies, as the zero imputation is uniform and the model learns region-specific intercepts through the hierarchical structure. Extension of ISTAT coverage to all 252 constituencies constitutes a priority development milestone.

Feature	Description	Source	Coverage
pct_giovani	% residents aged 15–34	ISTAT 2021	All 252 const.
pct_anziani	% residents aged 65+	ISTAT 2021	All 252 const.
pct_alta_istr	% with tertiary education	ISTAT 2021	All 252 const.
pct_femmine	% female residents	ISTAT 2021	All 252 const.
pct_occupati	Employment rate	ISTAT 2021	All 252 const.
pct_stranieri	% foreign-born residents	ISTAT 2021	14 Lazio only*
pct_extra_ue	% non-EU country of origin	ISTAT 2021	14 Lazio only*
is_politiche	Binary: general vs. regional election	Eligendo	All 252 const.

** Constituencies not in Lazio receive imputed value of zero. Extension to full national coverage is planned.*

2.3 Survey Data — ITANES, ESS Round 11, TRIPOL IT

Stage 3 (message optimizer) draws on three complementary survey datasets to construct psychographic profiles for ten demographic segments:

- ESS Round 11 (2023–24): N=2,865 Italian respondents, face-to-face probability sample. Variables used: institutional trust, left-right self-placement, immigration attitudes, democratic satisfaction. Data represent post-2022 election structural values; not used as vote intention proxies.

- TRIPOL IT (2021–22): N=1,231 Italian respondents. Variables: affective polarization scores, WAPSV political values index. Used to calibrate the emotional valence of political messaging across segments.

- Meta Ad Library (May 2026): 30 political and NGO advertisements analyzed (STC, UNICEF, Emergency). Used exclusively for tone and hook calibration in the lasciti donatori and responsabili CSR segments. This source is not used for demographic inference.

A declared limitation: TRIPOL coverage of the TERZO_POLO segment is based on N=44 ESS respondents — a statistically fragile subsample. Estimates for this segment should be interpreted with elevated uncertainty. The ITANES dataset (Italian National Election Study) provides additional longitudinal validation context but is not used as a direct model input.

2.4 Regional Electoral Results — Validation Datasets

Model validation draws on official Eligendo results from 13 Italian regions across multiple election cycles (2020–2023), comprising 252 constituency-election observations. Six regions were selected as primary validation benchmarks for OLS vs. MRP comparison: Toscana, Emilia-Romagna, Veneto, Sicilia, Sardegna, and Puglia. The 14 Lazio constituencies (Camera uninominal, 2022 general election) constitute the primary backtest dataset, chosen because ISTAT feature data are available at full resolution for this region.

Data quality for historical regional results is generally high, with the exception of local coalition compositions that differ from the national standard. Seat-level discrepancies attributable to local political formations are documented in the validation output but do not constitute model errors.

3. Stage 1 — Electoral Swing Model

Stage 1 implements a constituency-level uniform swing model applied to the 221 uninominal seats of the Rosatellum system. The model translates a user-specified or externally sourced national vote distribution into constituency-level outcomes by propagating coalition-level vote share changes from the 2022 baseline.

3.1 Model Specification

Let $V^{2022}*c$ denote the observed vote share of coalition c in the 2022 general election, and $V^{sc}*c$ the user-specified scenario vote share. The coalition-level swing is defined as:

$$\Delta_c = V^{sc}_c - V^{2022}_c$$

For each uninominal constituency i, the incumbent winner c^*_i retains the seat unless the cumulative advantage of the strongest challenger exceeds half the observed 2022 margin of victory:

$$\text{seat}_i \leftarrow c_{\text{challenger}} \text{ if } \max_{c \neq c^*}(\Delta_c - \Delta_{c^*}) > \text{margin}_i / 2$$

$$\text{seat}_i \leftarrow c^* \text{ otherwise}$$

The margin of victory $margin_i$ is sourced directly from Eligendo 2022 official results for the 14 Lazio constituencies (Camera), and from macro-area averages for the remaining 207 constituencies. This asymmetry — real margins for Lazio, aggregated margins elsewhere — reflects the current state of ISTAT feature availability and constitutes the primary known approximation in Stage 1.

3.2 Proportional Allocation

The 63% of seats allocated proportionally (245 Camera; 122 Senato) are distributed via the d'Hondt method with a 3% national threshold (Rosatellum Article 83, paragraph 1(c)). The implementation applies the threshold party-by-party, then rescales the admitted parties' totals to 100% before allocation:

admitted = {p : V_p ≥ 3.0% and p qualifies as a party list}

seats_p = floor(V_p / sum(V_admitted) × N_seats) + remainder correction

Remainder seats are assigned by largest fractional remainder (Hamilton/Hare method). Coalition totals in the proportional tier are computed by summing party-level allocations according to the declared coalition membership map (MAPPA_COAL), which currently covers CDX, CSX, M5S, and CENTRO. Parties not assigned to any coalition are pooled into an ALTRI category.

3.3 Macro-Area Structure

For the 207 non-Lazio constituencies, Stage 1 operates on a macro-area aggregation derived from Eligendo 2022. The five macro-areas (Nord-Ovest, Nord-Est, Centro, Sud, Sud-Isole) are parameterized by: the number of constituencies, 2022 baseline winners per coalition, the average winning margin, and the number of marginal constituencies ("collegi in bilico").

Macro-area	Collegi Camera	CDX 2022	CSX 2022	M5S 2022	Margine medio	In bilico
Nord-Ovest	38	28	5	0	12.5 pp	8
Nord-Est	27	22	3	0	14.2 pp	5
Centro*	14	14	0	0	24.9 pp	4
Sud	36	22	4	8	8.1 pp	14
Sud-Isole	18	8	2	7	6.4 pp	9

** The Centro macro-area is partially superseded by constituency-level Lazio data in v2.1; the remaining Central Italy constituencies use macro-area parameters.*

The number of seats that flip in a macro-area is bounded above by the number of marginal constituencies and is a linear function of the swing advantage:

seats_flipped = min(n_bilico × Δ_advantage / (margin_medio + ε), n_won_2022)

This functional form deliberately underestimates seat changes in landslide scenarios — a conservative choice that reduces false precision. The appropriate use case for Stage 1 is near-baseline scenario analysis (±5–8 pp swings); extreme scenarios should be interpreted with additional caution.

3.4 Double Majority Check

Stage 1 computes seat totals for both chambers simultaneously and evaluates whether any coalition or coalition combination achieves a governing majority (Camera: 201 seats; Senato: 104 seats). The output explicitly flags three conditions: doppia*maggioranza (majority in both chambers), maggioranza*parziale (majority in one chamber only), and no majority. This feature directly addresses the distinctive constitutional constraint of the Italian system, where a government must maintain confidence in both branches of parliament independently.

4. Stage 2 — Multilevel Regression and Poststratification

Stage 2 applies a Multilevel Regression and Poststratification (MRP) model to refine the seat distribution produced by Stage 1, explicitly accounting for demographic heterogeneity across constituencies. MRP was originally developed by Gelman and Little (1997) and systematically evaluated for electoral forecasting by Wang et al. (2015), who demonstrated MAE of approximately 2–3 pp in US state-level presidential vote share estimates. PoliSim adapts the framework to the Italian multi-party, multi-chamber context.

4.1 Model Architecture

The MRP model is implemented in PyMC 5.28.4 (Salvatier et al., 2016). Observed outcomes are election results from official Eligendo records across 252 constituency-election observations (13 regions, multiple election cycles 2020–2023). The model estimates vote share for four coalitions simultaneously.

4.2 Likelihood Function

A Dirichlet-Multinomial likelihood is used to model the joint distribution of vote shares across four competing coalitions. This choice is deliberate and addresses two specific requirements: (i) vote shares must sum to one across coalitions, and (ii) the model must treat all coalitions symmetrically without privileging any particular competitor. Earlier specifications using a Normal likelihood for CDX margins introduced asymmetric bias; the switch to Dirichlet-Multinomial was explicitly motivated by observed CDX-directional overprediction in preliminary validation runs.

For constituency i with total votes $N*i$, observed vote counts $(y*{i,CDX}, y*{i,CSX}, y*{i,M5S}, y_{i,CENTRO})$ are modeled as:

y_i | α_i, N_i ~ DirichletMultinomial(N_i, α_i)

α_{i,c} = softmax(μ_{c} + Σ_k β_{k,c} × X_{ik} + γ_{region[i],c})

where $μc$ is the coalition-specific intercept, $β{k,c}$ is the coefficient of the k-th demographic feature for coalition c, $X*{ik}$ is the standardized ISTAT feature value, and $γ{region,c}$ is a region-specific random effect.

4.3 Feature Set and Prior Specification

The eight features used in trace v4 are listed below with their prior distributions. All continuous features are standardized (z-scored) across the 252-observation training set prior to model fitting:

Feature (z-scored)	Description	Prior
pct_giovani_z	% residents aged 15–34	Normal(0, 1)
pct_anziani_z	% residents aged 65+	Normal(0, 1)
pct_alta_istr_z	% tertiary-educated	Normal(0, 1)
pct_femmine_z	% female residents	Normal(0, 1)
pct_occupati_z	Employment rate	Normal(0, 1)
pct_stranieri_z	% foreign-born (Lazio only)	Normal(0, 1)
pct_extra_ue_z	% non-EU origin (Lazio only)	Normal(0, 1)
is_politiche	Binary: general (1) vs regional (0)	Bernoulli implicit

All Normal(0,1) priors reflect weak informative priors centered on zero effect. The Dirichlet concentration parameter uses a Half-Normal(1) hyperprior.

The feature selection rationale for pct*stranieri*z and pct*extra*ue_z (added in trace v4) was established through systematic backtest analysis: on 14 Lazio constituencies, these two variables showed the strongest correlation with MRP prediction error (r = −0.800 and r = −0.792 respectively), outperforming all other candidate features as bias predictors. Their inclusion in v4 reduced MAE by 0.66 pp relative to v3 (4.67 → 4.01 pp).

4.4 Hierarchical Structure and Region Effects

Region-level random effects $γ_{region,c}$ capture structural political differences across Italian macro-regions (e.g., the historically strong M5S vote in Southern regions, the CDX dominance in Nord-Est). These effects are modeled with a weakly informative half-Cauchy hyperprior on the standard deviation:

γ_{region,c} ~ Normal(0, σ_region)

σ_region ~ HalfCauchy(1)

The is_politiche binary feature accounts for the systematic difference between national general elections (where Rosatellum applies in full) and regional elections (different candidate structures, lower turnout, higher M5S and local party shares in some regions). Including regional elections in the training set expands the N from 14 (Lazio 2022 only) to 252, substantially improving the statistical foundation of the estimates.

4.5 Sampling Configuration

Trace v4 was estimated using NUTS (No U-Turn Sampler) with the following configuration:

draws=2000, tune=1000, chains=4, target_accept=0.90 (default)

Convergence diagnostics: maximum $\hat{R}$ = 1.000 across all parameters (Gelman-Rubin criterion; threshold: $\hat{R}$ < 1.01); 125 divergent transitions (0.16% of post-warmup samples). The divergence count increased from 10 (v3) to 125 (v4) following the addition of two new features, indicating mild posterior geometry complexity. At 125/8000 post-warmup samples (1.6%), divergences are within the range typically considered acceptable for applied work, but warrant monitoring in future model iterations. A tuning run with target_accept=0.99 is planned as a post-release improvement.

5. Validation and Backtesting

Validation of the PoliSim MRP model follows a two-level protocol: (i) constituency-level backtest on 14 Lazio Camera constituencies using 2022 general election results as the held-out reference, and (ii) regional-level OLS vs. MRP comparison across six Italian regions. Both protocols use only out-of-sample predictions — the model is trained on the full 252-observation dataset and predictions are generated using leave-one-region-out cross-validation for the OLS comparison and direct prediction for the Lazio backtest.

5.1 Lazio Backtest — 14 Constituencies

The primary validation benchmark is the backtest on 14 Camera uninominal constituencies in Lazio (2022 general election). This region was selected as the validation dataset because ISTAT 2021 Census data are available at full sezione-level resolution, enabling a genuine demographic MRP prediction rather than zero-imputed estimates. The results below report MRP model v4 predictions against official Eligendo 2022 counts.

Metric	MRP v3	MRP v4	Interpretation
MAE CDX vote share	4.67 pp	4.01 pp	↓ 14.1% improvement
RMSE CDX vote share	6.08 pp	4.64 pp	↓ 23.7% improvement
Bias (mean error) CDX	−3.69 pp	−3.33 pp	Systematic underest., reduced
Winner accuracy	14/14	14/14	100% constituency winners
CI90 empirical coverage	29%	29%	⚠ Structural issue (see §6)
Max P(CDX wins) in CDX seats	n/a	100%	High confidence predictions

CDX = Centrodestra coalition (FdI, Lega, FI). Vote share is percentage of total valid votes. CI90 coverage denotes the fraction of true outcomes falling within the 90% posterior credible interval.

The constituency-level results reveal a systematic geographic pattern in residuals: the five Lazio 2 constituencies (predominantly provincial/rural) show higher absolute errors (mean |error| = 3.1 pp) compared to the nine Lazio 1 constituencies centered on Rome (mean |error| = 5.2 pp for outer Rome, |error| = 0.97 pp for central Rome). This heterogeneity suggests that the current feature set captures urban demographic gradients better than provincial dynamics, and motivates the inclusion of additional features (labour market composition from DCSC_CONDPROFOCCUP) in the planned v5 specification.

5.2 OLS Baseline Comparison — Six Regions

To evaluate whether the MRP model provides a genuine improvement over a simpler statistical baseline, an OLS regression using the same 8-feature design matrix was estimated on identical training data. Out-of-sample RMSE was computed for six held-out regions across single election cycles:

Region	Election	RMSE OLS	RMSE MRP	**Δ** (MRP advantage)
Toscana	2020 Reg.	8.21 pp	7.43 pp	+0.78 pp
Emilia-Romagna	2020 Reg.	9.84 pp	8.67 pp	+1.17 pp
Veneto	2020 Reg.	7.92 pp	7.31 pp	+0.61 pp
Sicilia	2022 Reg.	10.12 pp	8.91 pp	+1.21 pp
Sardegna	2024 Reg.	8.76 pp	8.10 pp	+0.66 pp
Puglia	2020 Reg.	8.94 pp	6.54 pp	+2.40 pp
Mean (6 regions)	—	8.96 pp	7.82 pp	+1.14 pp

OLS baseline uses identical feature set and training data. MRP advantage = RMSE_OLS − RMSE_MRP. Positive values indicate MRP outperforms OLS. All held-out elections were not included in the training set for that fold.

The MRP model outperforms the OLS baseline in all six held-out regions, with an average RMSE advantage of 1.14 pp. The advantage is largest in Puglia (+2.40 pp), a region with high demographic heterogeneity between coastal urban areas and inland agricultural constituencies, consistent with the expectation that hierarchical modeling provides the greatest benefit when within-region heterogeneity is high.

For reference, Wang et al. (2015) report MAE of approximately 2–3 pp for US presidential vote share estimates from MRP models trained on opt-in online polls. The PoliSim RMSE of 7.82 pp on out-of-sample regional elections is substantially higher, reflecting two sources of additional difficulty: (i) the Italian multi-party context, where prediction errors compound across four coalitions rather than a binary outcome, and (ii) the absence of proprietary constituency-level polling data, which Wang et al. utilize as a key input. The PoliSim benchmark from a held-out regression PoC on Lazio (RMSE 3.91 pp) demonstrates that the model can approach Wang et al. accuracy when high-resolution demographic data are available.

6. Declared Limitations

Scientific credibility requires that limitations be declared as prominently as results. The following limitations are architectural properties of the current PoliSim implementation — not provisional gaps to be quietly remediated, but structural features of the modeling approach that affect the interpretation of all outputs. Users of the platform are expected to have read this section.

6.1 Confidence Interval Coverage Failure

The most significant known limitation of the MRP model is the empirical CI90 coverage of 29% on the 14 Lazio constituencies, against the nominal 90% expectation. This means that in approximately 71% of cases, the true outcome falls outside the 90% posterior credible interval — a severe failure of interval calibration.

This is not a sign that point estimates are wrong (MAE 4.01 pp is operationally useful), but it means that the uncertainty intervals reported by the model are dramatically overconfident and should not be used for probabilistic inference. The root cause is insufficient posterior variance: the Dirichlet-Multinomial likelihood with the current prior structure concentrates probability mass more tightly than the true sampling distribution warrants. Remediation requires either richer priors on the concentration parameter or an explicit variance inflation component — not additional features.

Operational guidance: CI outputs from Stage 2 should be interpreted as directional uncertainty indicators, not as calibrated probability statements. Point estimates (MAP) are the operationally valid outputs. Probabilistic claims

("X has a Y% chance of winning") should not be derived from the current CI structure without variance recalibration.

6.2 Systematic CDX Bias

A mean prediction error of −3.33 pp for CDX vote share persists in trace v4 (v3: −3.69 pp). The model systematically underestimates CDX performance. Analysis of residuals by constituency type reveals that the bias is concentrated in Lazio 2 constituencies (provincial, predominantly small comuni), where CDX outperformed model predictions by an average of 4.8 pp. Urban Lazio 1 constituencies show lower systematic bias (mean error −2.1 pp).

The bias is partially explained by features not yet in the model — specifically, agricultural employment share and small-comuni structural voting patterns, which are not captured by the current ISTAT 2021 census features. The DCSC_CONDPROFOCCUP dataset (ISTAT employment by sector at municipality level) is identified as the highest-priority data source for bias reduction in the planned v5 specification. The bias does not affect winner accuracy at current CDX dominance levels in Lazio, but would become consequential in near-parity scenarios.

6.3 Absence of Proprietary Polling Data

PoliSim is an *ecological model*: it estimates vote shares from demographic and historical features, not from individual-level survey responses linked to geographic locations. This is the defining structural limitation of the approach. The model cannot "know" that a party's support has shifted in a particular region unless that shift is detectable from the Eligendo historical record or from the ISTAT demographic profile.

The practical implication is that PoliSim cannot autonomously forecast which electoral scenario will materialize. It is a *what-if calculator*, not a predictive instrument. The planned integration of external polling aggregator data (Politpro supermedie, currently ingested via a weekly cron scraper) allows the platform to answer "given today's polls, what would the seat distribution be?" — but this remains contingent on the accuracy of the external polling data.

6.4 ISTAT Feature Coverage

As documented in Section 2.2, features pct*stranieriz* and pct*extra*ue_z are populated from real data for only 14 of 252 constituencies. The remaining 238 constituencies receive a zero imputation for these features. This means the full 8-feature model is effectively a 6-feature model for 94% of constituencies, reducing its accuracy gains to those attributable to the six universally-covered features.

Extension of ISTAT coverage to all 252 constituencies requires downloading and processing sezione-level census files for all 20 Italian regions (vs. Lazio only). The data exist and are publicly available; the work is an engineering task estimated at 5–8 development sessions. Until this is completed, RMSE improvements from the v4 features should be understood as Lazio-specific results that may not generalize nationally.

6.5 Stage 3 — Unvalidated Calibration Weights

The message optimizer (Stage 3) generates demographically targeted communication strategies by applying expert-calibrated scoring weights to ESS Round 11 and TRIPOL IT survey data. These weights determine which message frames are prioritized for each of the ten demographic segments. The weights were calibrated by domain experts based on review of Meta Ad Library performance data (30 ads) and qualitative review of survey microdata, but have not yet been validated through controlled A/B testing against real campaign outcomes.

The operational consequence is that Stage 3 outputs should be treated as structured hypotheses for testing, not as empirically validated recommendations. A validation protocol requiring 20 partner campaigns with standardized outcome measurement has been designed and is currently in active recruitment.

7. Stage 3 — Message Optimizer

Stage 3 generates message strategies for political campaigns, civic organizations, and NGOs, calibrated to the demographic and psychographic profile of each target segment. The component operationalizes political communication theory — principally Entman's framing theory (1993) and subsequent empirical developments in agenda-setting and targeted political communication (Nickerson & Rogers, 2010; Broockman & Kalla, 2016) — through a structured prompt engineering pipeline built on the Anthropic Claude API.

7.1 Psychographic Profile Architecture

Ten demographic-psychographic segments are defined for the Italian political context. Each segment is characterized by a multi-dimensional profile combining objective sociodemographic descriptors (age, education, labour market status, geographic location) with attitudinal and behavioral variables derived from survey sources:

Segment	N (survey basis)	Primary survey source	Key attitudinal axes
Giovani precari Sud	N≈380 ESS R11	ESS R11 + TRIPOL	Institutional distrust, emigration intent
Casalinghe disoccupate Sud	N≈290 ESS R11	ESS R11	Care work recognition, health access
Operai/artigiani Nord	N≈310 ESS R11	ESS R11 + TRIPOL	Automation risk, tax burden, left suspicion
Laureati urbani progressisti	N≈420 ESS R11	ESS R11	Democratic quality, climate, housing cost
Astensionisti valoriali	N≈180 ESS R11	ESS R11 + TRIPOL	Systemic cynicism, radical discontinuity
Pensionati Centro-Nord	N≈340 ESS R11	ESS R11	Health system, security, party loyalty
Donatori lasciti testamentari	N≈85 ESS R11	ESS R11 + Meta Ads	Legacy, institutional trust, death taboo
Responsabili CSR	Expert-calibrated	Meta Ads library	ESG accountability, reputational ROI
Terzo Polo moderati	N≈44 ESS R11	ESS R11	⚠ Thin subsample — elevated uncertainty
Imprenditori PMI	N≈95 ESS R11	ESS R11 + TRIPOL	Regulatory burden, credit access

ESS R11 = European Social Survey Round 11 (2023-24), Italy subsample N=2,865. TRIPOL = Italian affective polarization study (2021-22), N=1,231. N values are approximate subsample sizes by segment approximation from available survey microdata.

7.2 Empirical Data Integration

Profile construction follows a hierarchical data integration protocol. ESS Round 11 provides the structural attitudinal baseline: institutional trust (scale 0–10), left-right self-placement (scale 0–10), immigration attitude index (composite), and democratic satisfaction. These are computed as segment-level means and standard deviations from the Italian ESS subsample, aggregated by age × education × employment cells aligned with segment definitions.

TRIPOL IT contributes the affective polarization layer: WAPSV (World Values Survey Political Values) scores by segment, and in-group/out-group favorability differentials. ESS Round 11 is post-2022 election data; TRIPOL covers the 2021–22 cycle. Both datasets are used as proxies for structural political values rather than current vote intention — a critical interpretive constraint documented in every Stage 3 output.

Meta Ad Library calibration: 30 Italian political and NGO advertisements were analyzed from the Meta Ad Library (May 2026), covering organizations STC (emergency aid), UNICEF Italy, and Emergency. For the lasciti testamentari and responsabili CSR segments, ad performance patterns (engagement signals, audience overlap data) were used to calibrate tone and hook parameters. This source is used exclusively for communication style calibration; it is not used for demographic profiling.

7.3 Prompt Architecture and Output Structure

The Stage 3 pipeline injects empirical profile data — dati*empirici — directly into the generation prompt for all ten segments at runtime, loaded from the enriched ITANES profile JSON (itanes*profili*enriched.json). For the two NGO-adjacent segments (donatori*lasciti, responsabili*csr), an additional calibrazione*meta_ads block is injected.

The output structure for each message optimization request includes: (i) a primary message frame with explicit citation of the demographic trigger, (ii) a secondary "objection response" frame addressing the segment's documented communication barriers, (iii) a channel recommendation (platform, format, tone register), and (iv) an explicit "avoid" list derived from the segment's negative trigger profile. All outputs carry a caveat stating that weights are expert-calibrated and have not been validated on A/B test outcomes.

7.4 Commercial Adaptation

The political segment architecture is being extended to accommodate three non-political use cases: brand communication (adding propensione*acquisto, premium*accessibile, razionale*emotivo scoring axes), NGO major gifts and lasciti (fiducia*istituzionale, eredità, delicatezza*morte), and NGO corporate fundraising (ritorno*reputazionale, ESG*alignment, misurabilità*impatto). These extensions require schema changes to the output JSON but do not require changes to the MRP or swing model components. The core demographic segmentation logic transfers directly from political to civic contexts.

References

Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, 352(6282), 220–224.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4), 51–58.

Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23(2), 127–135.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472.

Italian Ministry of Interior (2022). Eligendo — Archivio storico delle elezioni. Available at: https://elezioni.interno.gov.it/opendata

ISTAT (2021). 15° Censimento generale della popolazione e delle abitazioni — Indicatori per sezione di censimento. R12 Lazio dataset.

Nickerson, D. W., & Rogers, T. (2010). Do you have a voting plan? Implementation intentions, voter turnout, and organic plan making. *Psychological Science*, 21(2), 194–199.

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, 12(4), 375–385.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2, e55.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3), 980–991.

PoliSim Technical White Paper v1.0 — May 2026

polisim.dev | *admin@qitalia.org* | *github.com/AlCap27/polisim*