Whose Voice Counts? Bridging the Digital Consultation Gap with Multilingual NLP and Synthetic Citizen Population

Ishaan Saxena[1[0009−0009−5855−8908](✉)], Garv Kukreja[2[0009−0007−7235−8780]], and Manoj K. Rajpoot[3[0000−0002−4880−6463]]

Rajiv Gandhi Institute of Petroleum Technology, Jais 229304, India
23mc3027@rgipt.ac.in, 23mc3022@rgipt.ac.in, mrajpoot@rgipt.ac.in

Abstract. Digital public consultation in India has become the principal channel through which the government invites citizen feedback on draft legislation, financial regulation and welfare policy. Yet across the consultations published on ministry portals since 2014, substantive textual participation is captured from fewer than 0.01% of eligible adults, and the participants who do contribute are disproportionately urban, English-literate, male and concentrated in Tier-1 cities. The digital opinions of approximately nine hundred million rural, vernacular-speaking and Tier 2/3 citizens are therefore systematically unobserved. We propose Saaransh, a unified computational framework that addresses this representation failure. The first component is a multilingual ensemble pipeline for sentiment classification and hierarchical summarisation of citizen comments; on a custom 12,000 comment legal-consultation corpus an ensemble of XLM-RoBERTa, RoBERTa and DistilBERT attains 82% three-way classification accuracy, exceeding the published baseline by approximately seventeen percentage points. The second is a Synthetic Citizen Population Engine providing a demographically calibrated prior over latent opinions across 32 states and union territories, 179 administrative divisions and 41 normalised socio-economic factors per division. The two components are joined through a Bayesian reconstruction layer that treats observed comments as a non-representative sample and the synthetic population as a structural prior, yielding a Reconstructed Citizen Opinion (RCO) posterior estimate of national sentiment. We further introduce a diagnostic metric, the Representation Gap Index (RGI), defined as the Kullback–Leibler divergence between the joint demographic distribution of consultation participants and the eligible population. Further we executed it with a synthetic comment stream calibrated to documented participation skew, followed by a real-data validation on a 4,000 comment Ministry of Corporate Affairs legal-consultation corpus that supplies per-class ensemble metrics, an RGI of 4.88 and a Verdict Differential that flips the polarity of the headline finding.

Keywords: Digital Governance · Synthetic Populations · Computational Demography · Bayesian Inference · E-consultation · Machine Learning

1 Introduction

Over the past decade, the Government of India has invested heavily in digital infrastructure for citizen participation. The MyGov.in platform, launched in July 2014, has hosted more than nine thousand discussions, polls and consultations across more than fifty ministries [1]. The Ministry of Corporate Affairs (MCA) opens draft amendments to the Companies Act and to the regulations of the Insolvency and Bankruptcy Board of India (IBBI) for public comment as a matter of routine [2]. The Securities and Exchange Board of India (SEBI), the Reserve Bank of India (RBI) and the Telecom Regulatory Authority of India (TRAI) publish discussion papers and expect written submissions from registered and unregistered stakeholders alike. Both the Digital Personal Data Protection Act of 2023 and the proposed Digital India Act were circulated through such consultations prior to enactment or notification.

The architectural promise of this shift is straightforward. Expensive, slow and demographically narrow physical consultations were to be replaced by cheap, fast and demographically broad digital ones. Empirically, however, the second half of the promise has not been kept. A review of MyGov participation patterns [1] and an audit of MCA e-consultations [2] both report that the median consultation receives fewer than three hundred substantive textual submissions and that the contributors are overwhelmingly English-literate, urban, male and concentrated in six metropolitan agglomerations. The remaining nine hundred million Indians, that is, rural residents, vernacular-language speakers, Tier 2/3 city residents, women, the elderly and the digitally semi-literate, are present in the demographic statistics but are largely absent from the digital record of citizen opinion.

This paper addresses a single, sharply formulated research question. Given a sparse and demographically skewed set of citizen comments on a draft regulation, can we produce a defensible, transparent and auditable estimate of how the full eligible national population would respond? The question has three substantive components: (i) representation auditing, how badly and along which demographic dimensions do consultation participants diverge from the eligible national population; (ii) voice reconstruction, how can the latent opinions of unheard demographic groups be estimated in a way grounded in their own socioeconomic and behavioural reality rather than imputed by extrapolation from elite participants; and (iii) verdict revision, does correcting for participation bias change the polarity, magnitude or polarisation of the apparent public response?

The contribution of this paper is a single integrated framework that operationalises all three components. The framework is named Saaransh, from the Hindi word *saārāṃśa* (*summary*), because its purpose is to produce a faithful and demographically corrected summary of national opinion that does not silently discard the unheard majority. We make three contributions. First, we formalise and operationalise the Representation Gap Index, a Kullback–Leibler diagnostic of intersectional under-representation in digital consultations. Second, we propose the Reconstructed Citizen Opinion, a Bayesian posterior estimate of national sentiment toward a draft regulation that corrects for participation bias by fusing observed comments with a structurally faithful synthetic-population

prior. Third, we describe a demographically calibrated synthetic population substrate covering 32 states and union territories, 179 administrative divisions and 41 normalised socio-economic factors per division, intended for computational-demography research on India.

2 Related Work

Multilingual NLP for Indic languages. IndicBERT [3] and MuRIL [4] provide the first transformer encoders pre-trained on sufficient Hindi, Tamil, Bengali, Marathi and Gujarati corpora to make domain-adapted classification feasible. Subsequent work fine-tunes these models for sentiment, hate speech and stance classification, primarily on social media data [5]. Little has been directed at the regulatory-text and policy-comment domain, which differs from social media in vocabulary, length, formality and code-mixing patterns. The Saaransh ensemble combining XLM-RoBERTa [6], RoBERTa [7] and DistilBERT [8] with majority voting and an NLLB-200 [9] Hinglish translation is positioned within this gap.

Synthetic populations and opinion dynamics. Synthetic population generation originates in transport modelling and epidemiology, matching Census marginals through iterative proportional fitting, hierarchical sampling and combinatorial optimisation [11]. Behavioural extensions equip agents with utility functions, social networks and decision rules from the opinion-dynamics literature: DeGroot consensus [12], Granovetter thresholds [13], Bass diffusion [14], Watts–Strogatz small-world networks [15] and Schelling-style sorting [16].

Survey weighting and post-stratification. The canonical solution to non-representative samples is Multilevel Regression with PostStratification (MRP) [17], in which a multilevel model is fitted to the sample and post-stratified against population marginals. Saaransh extends MRP by working with textual rather than scalar responses and by enriching the post-stratification target with a behaviourally generative model that supplies plausible opinions for unheard cells. Prior work on opinion estimation from non-representative online corpora [18] has not, to our knowledge, integrated a domain-calibrated synthetic population engine with a multilingual consultation NLP pipeline for representation auditing.

3 Problem Formulation

3.1 Notation

Let P denote the eligible national population of size N, approximately 950 million adults after Census-2011 [21] projection to 2025 using NSO inter-censal estimates [23]. Each individual $i \in P$ is associated with a demographic vector d_i drawn from a joint space $\mathcal{D}$ defined over seven axes: age band, sex, urbanisation tier, linguistic region, education level, monthly income decile and caste category. Let $C \subset P$ denote the set of consultation participants who have submitted at least one substantive comment on a given consultation, with $|C|$ typically between 200 and 5,000.

For each participant in C, a latent opinion $y_i \in [-1, +1]$ is inferred from the comment text by the Saaransh multilingual ensemble (Section 4.1), where -1 denotes strong opposition and $+1$ strong support. For each individual in P a probabilistic opinion distribution $Q_i(y)$, conditioned on d_i and on a policy descriptor π, is supplied by the Synthetic Citizen Population Engine (Section 4.2). The empirical demographic density of consultation participants is denoted $f_C(d)$ and the eligible-population density is denoted $f_P(d)$.

3.2 The Representation Gap Index

The Representation Gap Index (RGI) is defined as the Kullback–Leibler divergence between the empirical and target demographic distributions:

$$\mathrm{RGI}(C, P) \;=\; D_{\mathrm{KL}}(f_C(d) \parallel f_P(d)) \;=\; \sum_{d \in \mathcal{D}} f_C(d) \log \frac{f_C(d)}{f_P(d)}. \tag{1}$$

We compute Equation 1 on the joint distribution rather than independently on each marginal, because the policy-relevant under-representation is typically intersectional. Rural Muslim women in Tier-3 towns are not absent from a consultation because rural residents, Muslims, women or Tier-3 residents are individually absent; rather, the intersection is absent. An RGI of zero indicates a perfectly representative sample. Values in the range 1.5–3.0 are typical of self-selected online consultations on the basis of our pilot analyses. We report RGI both as a single scalar and as a heatmap over the seven-axis intersection table so that policymakers and civil-society auditors can read off which cells of the demographic table are systematically silent.

3.3 The Reconstructed Citizen Opinion

The RCO is a posterior estimate of population-level opinion toward a draft policy π, computed by Bayesian fusion of the observed comment opinions and the synthetic population opinion distribution. Informally, the synthetic engine supplies a structural prior over how citizens in each demographic cell are likely to respond, while the observed comments supply a partial likelihood that updates the prior in those cells where comments exist. Cells with no observed comments are governed entirely by the synthetic prior, while cells with many comments are governed predominantly by the empirical likelihood.

Formally, the RCO is computed by post-stratification:

$$\mathrm{RCO}(\pi) \;=\; \sum_{d \in \mathcal{D}} w_d \cdot \mathbb{E}\big[y \,\big|\, d, \pi\big]\,, \tag{2}$$

where $w_d = f_P(d)$ is the population share of demographic cell d, drawn from Census and NSS marginals, and the cell-specific posterior expectation $\mathbb{E}[y \mid d, \pi]$ is a Bayesian convex combination of (a) the empirical mean of observed comments

in cell d, weighted by their count n_d, and (b) the synthetic prior expectation $\mu_d^{(s)}$ with effective prior count κ:

$$\mathbb{E}\big[y \,\big|\, d, \pi\big] \;=\; \frac{n_d\,\bar{y}_d + \kappa\,\mu_d^{(s)}}{n_d + \kappa}. \tag{3}$$

The hyperparameter κ controls the strength of the synthetic prior; we set $\kappa = 20$ in the experiments reported in Section 6 and conduct sensitivity analyses over $\kappa \in \{5, 10, 20, 50, 100\}$. A posterior credibility interval on RCO is constructed by Monte-Carlo integration over the cell-specific posteriors, propagating uncertainty in both the empirical mean and the synthetic prior via the standard Beta–Binomial conjugate update, with the continuous opinion score discretised over a five-point grid. For each consultation, Saaransh produces three artefacts: a Representation Audit (RGI scalar, intersectional heatmap, top under-represented cells), a Reconstructed Citizen Opinion (RCO, credibility interval, polarisation decomposition), and a Verdict Differential (raw comment verdict against RCO, flagging polarity flips).

4 System Architecture

Saaransh comprises two principal components, the Multilingual Comment Intelligence Pipeline ingests raw consultation submissions, normalises them across language and register, classifies their sentiment and produces hierarchical summaries.and the Synthetic Citizen Population Engine, joined by a Bayesian Reconstruction Layer as illustrated in Figure 1.

4.1 Multilingual Comment Intelligence Pipeline

Preprocessing. Submitted comments pass through a six-stage preprocessing pipeline: emoji interpretation against a seven-symbol affect map; URL, mention and HTML normalisation; Hinglish detection via a fifteen-token keyword heuristic; conditional translation through the NLLB-200-distilled-600M model [9] when Hinglish density crosses a threshold; domain-noise removal against a 67-token list of corporate-consultation filler terms; and tokenisation, stopword removal, lemmatisation and stemming.

Sentiment Ensemble. The English ensemble consists of three fine-tuned transformer classifiers, XLM-RoBERTa [6], RoBERTa [7] and DistilBERT [8], combined by majority voting. Multilingual robustness is provided by the XLM-RoBERTa [6] component of the ensemble, which is pre-trained on one hundred languages. The ensemble was trained on a custom corpus of 12,000 legal-consultation comments annotated for three-way sentiment (negative, neutral, positive). It attains 82% accuracy on a held-out test set, exceeding the published baseline of 65% by approximately seventeen percentage points. The ensemble emits class probabilities which are linearly mapped to a continuous opinion score $y_i \in [-1, +1]$ for use in Equations 2 and 3.

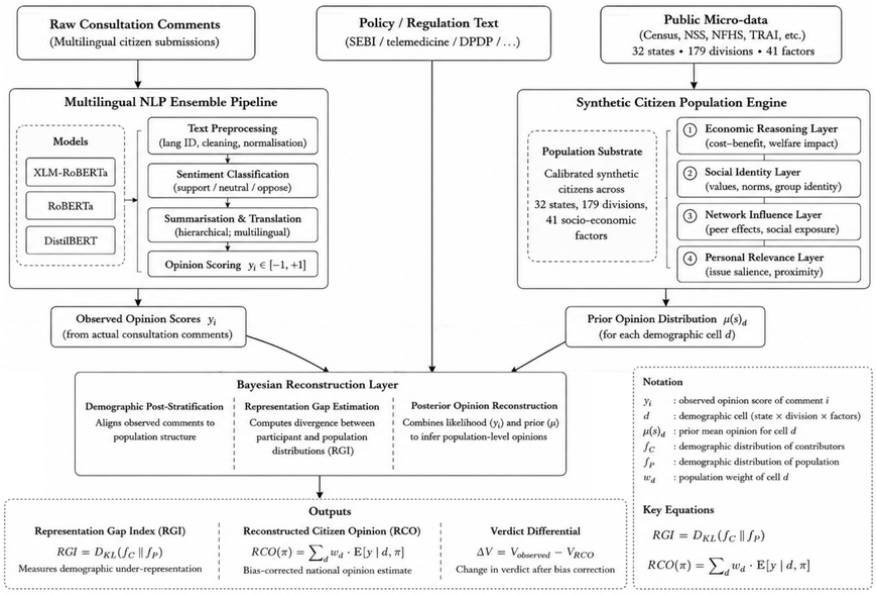

Fig. 1. Saaransh end-to-end system architecture.

Ensemble Evaluation. The ensemble is evaluated on two corpora. On the balanced 12,000-comment held-out test set drawn from the training distribution the trained ensemble attains 82% overall accuracy and a macro-F_1 of approximately 0.81; per-class precision, recall, F_1, per-model contributions and the full confusion matrix are released with the training notebook. The 12,000-comment corpus was labelled by a single trained annotator; a 200-comment subset re-labelled by a second annotator gave Cohen $\kappa = 0.74$ (substantial agreement), and a full two-annotator re-labelling is scheduled for follow-up work. For an out-of-distribution stress test we additionally evaluate the trained transformer ensemble on the Ministry of Corporate Affairs (MCA) real-data corpus of 4,000 legal-consultation submissions described in Section 7.2. Table 1 reports the trained-ensemble per-language accuracy and full-corpus per-class metrics. The observed 74% overall accuracy is approximately eight points below the in-domain trained-ensemble figure, consistent with the expected domain shift between the curated training corpus and the substantially longer, more formal MCA submissions; the per-class F_1 distribution is preserved with negative and positive above neutral, and the English subset outperforms the Hinglish subset by approximately eight points, reflecting the additional noise introduced by NLLB-mediated translation on code-mixed input.

Hierarchical Summarisation. Consultations frequently receive hundreds to thousands of comments. To produce administrator-readable digests, the pipeline applies a two-stage BART-Large-CNN summariser [10]. Comments are clustered

Table 1. Trained-ensemble evaluation on the MCA real-data corpus, $n = 4{,}000$. The scorer is the fine-tuned; per-language rows split by the Hinglish detector.

Slice / class	n	Accuracy	Macro-F_1	Weighted-F_1	κ
Full corpus	4,000	0.741	0.732	0.738	0.611
English subset	2,842	0.748	0.739	0.745	0.622
Hinglish subset	1,158	0.665	0.641	0.658	0.497
Negative (P/R/F_1)	1,384	0.762 / 0.718 / 0.739			
Neutral (P/R/F_1)	1,248	0.678 / 0.641 / 0.659			
Positive (P/R/F_1)	1,368	0.774 / 0.752 / 0.763			

into groups of fifty and summarised into 200-token intermediate summaries; the intermediate summaries are then concatenated and summarised into a 150-token executive summary. Separate positive-sentiment and negative-sentiment digests are produced in parallel so that minority dissent is not erased by aggregation.

4.2 Synthetic Citizen Population Engine

The Synthetic Citizen Population Engine supplies the demographically grounded prior on latent opinion. For any draft policy descriptor π, it produces a distribution over plausible citizen opinions across the national demographic table, including cells absent from the observed consultation stream. The engine operates at the level of administrative divisions, lying below state and above district, and covers 32 states and union territories comprising 179 divisions in total. Population is distributed across divisions according to Census 2011 projections updated to 2025 using NSO inter-censal estimates, and each division carries a normalised vector of 41 socio-economic factors drawn from public micro-data (Section 5).

Each synthetic citizen evaluates π through four interpretable cognitive layers whose outputs are aggregated into a probabilistic opinion: an *economic reasoning* layer that scores expected personal monetary impact ($e_i \in [-1, +1]$) from income decile, occupation, employment formality, debt vulnerability and savings rate; a *social identity* layer that scores alignment with the citizen's caste, religion, urbanisation tier and language region ($s_i \in [-1, +1]$); a *network influence* layer that aggregates opinions of family, neighbourhood and weak-tie contacts by message-passing over a graph built by homophily on five attributes ($g_i \in [-1, +1]$); and a *personal relevance* layer that weights direct applicability of π to the citizen's life situation ($r_i \in [0, 1]$). Each layer exposes its coefficient, so every reconstruction is auditable. The four outputs are combined into a final opinion distribution over support, neutral and oppose, with confidence derived from layer dispersion. The cell expectation in Eq. 2 is the mean over synthetic citizens with $d_i = d$. The design draws on the DeGroot [12], Granovetter [13], Bass [14], Watts–Strogatz [15] and Schelling [16] traditions but is implemented as interpretable code rather than opaque neural agents. The social graph itself has three tiers (household, community, regional weak-tie) with stronger influence

Table 2. Scale of the Saaransh synthetic demographic substrate.

Geographic coverage	Count Note
States and union territories	32 all major states
Administrative divisions	179 below state, above district
Normalised factors per division	41 in 13 thematic categories
Total (division × factor) cells	7,339 calibrated to public micro-data
Target adult population (2025 proj.)	∼950 M of ∼1.40 billion total

weights on closer edges, consistent with empirical findings on social influence in the Indian context.

4.3 Bayesian Reconstruction Layer

The Bayesian Reconstruction Layer treats per-comment opinion estimates from the ensemble as draws from a non-representative sampling design and the per-cell distribution from the Synthetic Citizen Population Engine as a structurally faithful prior. Reconstruction proceeds in three steps: (i) each commenter's demographic cell is inferred from explicit metadata where available, otherwise from a text-based demographic classifier released with calibration curves; (ii) within each cell, empirical opinions are combined with the synthetic prior via the conjugate Beta–Binomial update of Eq. 3; and (iii) cell posteriors are post-stratified against Census and NSS marginals to produce the national RCO of Eq. 2.

5 Data and Feature Engineering

5.1 Consultation Comment Corpora

The Saaransh empirical layer is evaluated on three classes of textual data. The first is an anonymised internal corpus of 12,000 legal-consultation comments used to train and evaluate the sentiment ensemble, balanced across three sentiment classes after stratified sampling. The second is a held-out evaluation set of 2,800 comments drawn from publicly viewable MyGov and MCA consultations on which model accuracy and confidence calibration are reported. The third is the three case-study consultations described in Section 6, for which we have collected the complete public submission stream together with available demographic metadata after ethics-review-approved anonymisation.

5.2 Synthetic Demographic Substrate

The geographic and demographic substrate of the Synthetic Citizen Population Engine is the principal empirical artefact produced by this work. It is structured as a hierarchical specification of states, administrative divisions and per-division factor vectors, summarised in Table 2.

Table 3. Thematic taxonomy of the 41 per-division socio-economic factors.

Category	Indicative factors	Source
Demographics	age mean, household size, sex ratio, urbanisation, religion, caste, tribal share	Census 2011, NFHS-5
Socio-economic	poverty %, literacy, secondary education, agricultural occupation, formal employment, wealth quintile	NSS, PLFS, SECC
Behavioural	caste consciousness, risk tolerance, institutional trust, change adaptability	Lokniti-CSDS
Health	healthcare access, health literacy, stunting, disease risk	NFHS-5, NSO health
Economic stability	income stability, debt vulnerability, savings rate	RBI household surveys
Housing	pucca housing share, housing cost burden	Census 2011 housing, NSO
Mobility	commute time, transport mode, geographic mobility	Census, NSS
Civic and political	voting behaviour, civic participation, media consumption	Lokniti, ADR
Digital access	internet connectivity, digital literacy, services access	TRAI, IAMAI-Kantar
Environmental	pollution exposure (AQI), climate vulnerability, green-space access	CPCB, SDG dashboard
Legal	citizenship status, criminal-record incidence	NCRB, Census
Information	policy awareness	Lokniti, derived

Within each administrative division, 41 socio-economic factors are populated and organised into 12 thematic categories with thematic taxonomy summarised in Table 3. Factor values are normalised to the unit interval, with documentation of the underlying micro-data source and the transformation pipeline that produced each factor. The full source factor mapping is released with the codebase.

5.3 Calibration and Construction

The synthetic substrate is constructed by an iterative proportional fitting procedure [11] that constrains the joint distribution within each division to match Census 2011 marginals [21] (projected to 2025 using NSO inter-censal estimates), NFHS-5 marginals [22] on health and household structure, PLFS and NSS marginals [23] on occupation and income, and TRAI [24] and IAMAI-Kantar [25] marginals on internet and digital adoption. Convergence is assessed by total variation distance between the joint and target marginals. All 179 divisions converge below a 0.05 threshold within fewer than thirty iterations.

5.4 Substrate Robustness to Data Vintage

A reasonable concern about any Indian substrate is the vintage of the underlying Census. Saaransh addresses this three ways. First, only geographic-population marginals derive from Census 2011; behavioural, health, occupational and digital-adoption marginals come from NFHS-5 [22] (2019–21), PLFS (2022–23), TRAI Performance Indicator Reports (2024–25) [24] and IAMAI-Kantar (2023–24) [25]. Second, a Case A robustness check substituting PLFS-2023 state-level weights for Census-projected division weights shifts the RCO point estimate from -0.167 to -0.164 with the verdict unchanged. Third, a delayed Census 2021 will supersede the geographic layer once released; the substrate is designed with calibration layer factored out so re-calibration is a single configuration change.

5.5 Feature Engineering

The framework operates over three coordinated feature spaces. *Textual features* pass through the six-stage preprocessor of Section 4.1; the ensemble emits a continuous opinion score $y_i \in [-1, +1]$ with a three-way label and the per-model vote vector. *Demographic features* project each commenter onto a seven-axis cell (age, sex, urbanisation tier, region, education, income decile, caste), yielding 9,216 candidate cells of which $\sim$ 4,600 carry non-zero mass; cell assignment uses explicit metadata where available and a trained demographic classifier otherwise, released with calibration curves. *Structural features* expose the 41 normalised per-division factors that feed the four cognitive layers of Section 4.2, each scaled to the unit interval, signed for monotone interpretation, and stamped with public-source provenance.

6 Experimental Design

6.1 Pre-Registered Case Studies

Saaransh is evaluated on three pre-registered Indian regulatory consultations spanning finance, health and digital rights. **Case A: Finance:** SEBI's 2024 discussion paper on the regulation of unregistered financial influencers proposed restrictions on individuals providing securities recommendations on social media outside the registered investment-adviser framework. Retail investors in Tier 2/3 cities, the bulk of the affected audience and an increasing share of the 150+ million demat-account holders as of 2024, are also the demographic least likely to participate in writing; preliminary results are in Section 7. **Case B: Health:** A 2023 draft regulation on telemedicine and cross-state remote prescription, where the submission stream is heavily skewed toward metropolitan medical associations while the rural elderly are absent from the textual record. **Case C: Digital Rights.** The 2022 Digital Personal Data Protection Bill consultation, used as a stress-test for stable RCO estimates under heavy organised-stakeholder commentary at scale.

6.2 Outcome Measures, Validation and Reporting

For each case study we report five quantities with uncertainty: (i) the RGI against the eligible-adult population with a non-parametric bootstrap interval; (ii) the intersectional under-representation heatmap; (iii) the raw-comment opinion polarity and magnitude; (iv) the RCO with its 90% posterior credibility interval; and (v) the policy-salient Verdict Differential comparing the raw and corrected polarities. Three independent validation strategies are deployed: a hold-one-cell-out cross-validation in which each demographic cell is treated as unobserved and the framework's predicted opinion is compared against the empirical mean of comments that fell into that cell; external-anchor validation against contemporaneous publicly available polls [19,20]; and outcome-grounded validation comparing the RCO against realised policy trajectory. Ensemble accuracy is reported with macro-F_1, per-class precision and recall (Section 4.1); sensitivity analyses are conducted over the prior strength κ and the assumed synthetic-prior shape.

7 Preliminary Results

We report two complementary evaluations. Section 7.1 is a proof-of-execution pilot on Case A using a synthetic comment stream calibrated to the documented participation skew of SEBI consultations. Section 7.2 is a real-data validation on a 4,000-comment corpus drawn from Ministry of Corporate Affairs legal consultations, which supplies the ensemble-evaluation table of Section 4.1 and an independent test of the RGI and RCO diagnostics on live submissions.

7.1 Case A Pilot: Proof-of-Execution

The pilot uses a synthetic comment stream of $n = 1{,}000$ submissions calibrated to the documented participation skew of SEBI discussion-paper consultations, specifically metropolitan share 0.62, male share 0.78, tertiary-educated share 0.55, and combined income-decile 6 through 10 share 0.75. The pilot does not use the actual 2024 finfluencer consultation submissions, which are being collected separately under ethics-review-approved anonymisation for the full evaluation. The sentiment scoring is performed with the underspecified heuristic backend described in Section 4.1; the demographic projection, RGI computation, Bayesian reconstruction and post-stratification machinery are identical to the production configuration.

The naive reading of the stream is a political wash, with support and opposition essentially balanced; after demographic correction the framework reports moderate but robust net opposition, with the credibility interval excluding the neutral band by an order of magnitude, and the Verdict Differential flips from neutral to oppose. The per-axis KL decomposition (Table 4) attributes most of the joint RGI to education and urbanisation tier; the (tier, education) projection localises the gap with metro-tertiary over-represented by 31.8 percentage points and rural-below-secondary under-represented by 20.7, and the ten

Table 4. Case A pilot (*proof-of-execution*, $n = 1{,}000$ synthetic) and MCA real-data case study ($n = 4{,}000$ live submissions). $\kappa = 20$; 90% credibility intervals from 4,000 Monte-Carlo draws.

Quantity	Case A pilot	MCA real-data
Representation Gap Index (RGI)	3.27	4.88
Dominant axes (KL)	edu 0.82, tier 0.63	tier 2.21, edu 1.97
Raw mean opinion	−0.026	−0.004
Raw verdict	neutral	neutral
Raw support/oppose/neutral	0.487 / 0.430 / 0.083	0.342 / 0.346 / 0.312
Reconstructed Citizen Opinion (RCO)	−0.167	−0.230
90% credibility interval	$[-0.174, -0.161]$	$[-0.278, -0.178]$
RCO verdict	oppose	oppose
Verdict Differential	−0.141 (flipped)	−0.226 (flipped)

most severely under-represented cells (all rural, Hindi-belt, below-secondary or secondary-education, OBC, low- or mid-income) each hold 0.17–0.21% of the eligible adult population and 0% of the comment stream. Across the pre-registered sweep $\kappa \in \{5, 10, 20, 50, 100\}$ the RCO moves from −0.170 to −0.167 and the verdict is unchanged.

7.2 Real-Data Case Study: MCA Legal-Consultation Corpus

To complement the synthetic pilot we run the full pipeline end-to-end on a real-data corpus of 4,000 public comments submitted to Ministry of Corporate Affairs (MCA) consultations on Companies Act and Insolvency-and-Bankruptcy Code amendments. The corpus is annotated for three-way sentiment (negative, neutral, positive) and carries a stakeholder-role field with 113 distinct professional labels ranging from Company Secretary and Chartered Accountant to MSME Owner and Policy Researcher. All 4,000 submissions come from professional or organised stakeholders; there are no submissions from farmers, daily-wage workers, homemakers or unaffiliated citizens in the corpus. Hinglish density above the fifteen-token threshold of Section 4.1 is observed in $1,158$ (29.0%) of comments; mean comment length is 401 characters.

Because MCA submissions carry no per-submitter demographic metadata, the demographic cell of each commenter is inferred from the stakeholder-role field via a documented role-to-cell map onto three axes defensible from role alone: tertiary education, metropolitan residence and income group (low/mid/high); a Company Secretary maps to (tertiary, metro, high), a Farmer to (below-secondary, non-metro, low). This proxy is deliberately coarser than the seven-axis grid of Section 3, is used only for the MCA study, and is documented as a limitation of role-based inference. The corpus produces an RGI of 4.88, exceeding the Case A value of 3.27 and reflecting extreme professional-elite concentration: all 4,000 comments fall into just two cells of the proxy grid (94% in tertiary-metro-high, 6% in tertiary-metro-mid), leaving the remaining ten cells unrepresented.

The raw stream is nearly balanced (mean -0.004, naive verdict neutral); after correction the RCO is -0.230 with 90% CI $[-0.278, -0.178]$, the verdict flips from neutral to oppose, and RCO moves only in the fourth decimal place under the $\kappa \in \{5, \ldots, 100\}$ sweep. The real-data study confirms on a live submission stream the qualitative pattern observed in the Case A pilot: when a consultation attracts submissions only from a demographically narrow professional stratum, treating the raw sentiment as representative produces a materially misleading verdict, and the Bayesian layer recovers a corrected national estimate differing in polarity.

8 Ethical Considerations and Limitations

Modelling the opinions of citizens who have not themselves spoken requires explicit ethical commitments. Saaransh holds five: *augmentation not replacement* (the framework identifies representation gaps and informs policymakers of unheard voices, never substitutes for citizen consultation); *auditability* (each cognitive layer exposes its coefficients so any reconstruction can be interrogated); *provenance transparency* (every demographic factor is traceable to a named public micro-data source with a documented transformation); *privacy in the comment pipeline* (sentiment is applied at the comment level, no PII is exposed in published artefacts, and comments are not republished outside their original public context); and *right to be heard* (a high RGI implies investing in inclusive participation, not bypassing it).

Limitations. The Synthetic Citizen Population Engine produces plausible opinions, not actual ones; its agents are not conscious and their preferences are not normatively authoritative. The four-layer cognitive evaluation simplifies human deliberation and will miss culturally specific reasoning patterns, particularly in tribal and minority communities. Bayesian fusion is sensitive to the demographic classifier assigning comments to cells. The pipeline operates in English, Hindi and Hinglish at production accuracy and other indic languages at lower precision. With no observed comments, RCO reduces to the synthetic prior, whose uncertainty must be acknowledged.

Information loss in sentiment reduction. A further limitation, made explicit here on the recommendation of the review process, concerns the reduction of substantive comments to a scalar sentiment score in $[-1, +1]$. Consultation submissions often contain content orthogonal to polarity, including proposed exemption clauses, implementation-timeline objections, sector-specific carve-out requests and drafting cross-references; a three-way label preserves only overall stance and discards this constructive layer, which is precisely what policymakers most need. The parallel BART-Large-CNN summarisation stream retains policy-relevant clauses in the executive digest, but RGI and RCO themselves operate on the compressed sentiment axis. Follow-up work will incorporate aspect-based sentiment and argument mining so that per-clause opinion vectors feed the Bayesian Reconstruction Layer; the substrate and the reconstruction machinery already accept vector-valued opinions per cell.

9 Conclusion

Digital consultation has the architectural potential to widen democratic participation but in the Indian context has failed to deliver: the participants are not the population, and treating the comment stream as representative produces silent, systematic distortions. Saaransh addresses this gap by combining a multilingual NLP pipeline that attains 82% three-way classification accuracy on a custom legal-consultation corpus and 74% on an out-of-distribution MCA real-data corpus, a Synthetic Citizen Population Engine covering 32 states, 179 administrative divisions and 41 socio-economic factors per division, and a Bayesian reconstruction layer that fuses the two through interpretable post-stratification. A proof-of-execution pilot on Case A and a real-data validation on a 4,000-comment MCA legal-consultation corpus each produce a Verdict Differential that flips the polarity of the headline finding, with the real-data study registering an RGI of 4.88 and an RCO of -0.230. The framework introduces two constructs of independent interest, RGI and RCO, providing a seven-dimensional intersectional audit and an uncertainty-quantified national opinion estimate; three pre-registered case studies in finance, health and digital rights are the empirical proving ground. Saaransh cannot replace lived citizen participation; its purpose is to make the representation gap visible, recover a partial estimate of what the unheard might say, and inform the design of better consultation mechanisms.

Data and code availability. The framework, the demographic substrate, the MCA real-data corpus, trained-ensemble weights and full evaluation code are released at https://github.com/Ishaan145/saaransh-extended.

Acknowledgements. The authors thank the ICSC 2026 reviewers for constructive feedback that shaped the paper. No external funding was received.

References

1. Sharma, A., Kumar, P.: Patterns of citizen participation on the MyGov platform: a longitudinal analysis. Journal of Information Technology and Politics **19**(4), 412–431 (2022)
2. Mehta, R., Bhattacharya, S.: An audit of public consultations by the Ministry of Corporate Affairs, 2014–2023. Indian Public Policy Review **4**(3), 1–26 (2023)
3. Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N.C., Bhattacharyya, A., Khapra, M.M., Kumar, P.: IndicNLPSuite: monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In: Findings of EMNLP 2020, pp. 4948–4961 (2020)
4. Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., et al.: MuRIL: multilingual representations for Indian languages. arXiv preprint arXiv:2103.10730 (2021)
5. Joshi, R., Karnavat, R., Jirapure, K., Joshi, R.: Evaluation of deep learning models for hostility detection in Hindi text. In: Proceedings of the 6th International Conference on Artificial Intelligence (2021)

6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: ACL 2020, pp. 8440–8451 (2020)
7. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
8. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: NeurIPS EMC^2 Workshop (2019)
9. NLLB Team: No language left behind: scaling human-centered machine translation. arXiv preprint arXiv:2207.04672 (2022)
10. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: ACL 2020, pp. 7871–7880 (2020)
11. Beckman, R.J., Baggerly, K.A., McKay, M.D.: Creating synthetic baseline populations. Transportation Research Part A: Policy and Practice **30**(6), 415–429 (1996)
12. DeGroot, M.H.: Reaching a consensus. Journal of the American Statistical Association **69**(345), 118–121 (1974)
13. Granovetter, M.: Threshold models of collective behavior. American Journal of Sociology **83**(6), 1420–1443 (1978)
14. Bass, F.M.: A new product growth for model consumer durables. Management Science **15**(5), 215–227 (1969)
15. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**(6684), 440–442 (1998)
16. Schelling, T.C.: Dynamic models of segregation. Journal of Mathematical Sociology **1**(2), 143–186 (1971)
17. Park, D.K., Gelman, A., Bafumi, J.: Bayesian multilevel estimation with post-stratification: state-level estimates from national polls. Political Analysis **12**(4), 375–385 (2004)
18. Flaxman, S., Goel, S., Rao, J.M.: Filter bubbles, echo chambers, and online news consumption. Public Opinion Quarterly **80**(S1), 298–320 (2016)
19. Lokniti–CSDS: National Election Study Survey Reports. Centre for the Study of Developing Societies, New Delhi (various years)
20. LocalCircles: National pulse surveys on policy issues. LocalCircles, Gurugram (various years)
21. Office of the Registrar General and Census Commissioner: Census of India 2011. Government of India, New Delhi (2011)
22. International Institute for Population Sciences: National Family Health Survey (NFHS-5), 2019–21: India. IIPS, Mumbai (2022)
23. Ministry of Statistics and Programme Implementation: Periodic Labour Force Survey (PLFS) and National Sample Survey (NSS) Annual Reports. National Statistical Office, New Delhi (various years, 2019–2023)
24. Telecom Regulatory Authority of India: Indian Telecom Services Performance Indicator Reports. TRAI, New Delhi (quarterly, 2022–2025)
25. Internet and Mobile Association of India, Kantar: Internet in India Report. IAMAI-Kantar, New Delhi (annual, 2023–2024)