Toward Collective-Centric Evaluation of Preference Inference for Participatory Democracy

Pierre-Antoine Lequeu[1]*, Salim Hafid[2]*, Paul Lerner[1], Nazanin Shafiabadi[1], Laurène Cave[3], David Mas[4], Jean-Philippe Cointet[2], Benjamin Piwowarski[1], François Yvon[1]

[1]Sorbonne Université, CNRS, ISIR, Paris, France
[2]Sciences Po, Médialab, Paris, France
[3]Sorbonne Université, STIH/CERES, Paris, France
[4]Make.org
lequeu@isir.upmc.fr, salim.hafid@sciencespo.fr

Abstract

To scale up collective decision-making, participatory democracy platforms such as Polis and Remesh enable online deliberation among thousands of participants. However, at this scale, participants cannot review every opinion submitted by others, producing highly sparse voting data that misrepresent patterns of consensus, conflict, and minority support. Platforms therefore increasingly rely on *Preference Inference (PI)* models to predict missing votes. Yet this automation is not neutral: inferred preferences can artificially amplify, suppress, or reorder existing patterns of support, ultimately reshaping how the outcomes of a deliberation are interpreted. More generally, we lack a systematic understanding of how existing *PI* methods affect the *collective* preference landscape. To address this gap, we benchmark several existing *PI* approaches in this context. Moving beyond conventional *user-centric* evaluations centered on the accuracy of individual predictions, we introduce a *collective-centric* evaluation framework that measures whether inferred votes preserve salient properties of the broader preference landscape. We further contribute the largest multilingual dataset of its kind: four consultations spanning over 90k participants, 1M votes, and 22 languages. Our experiments show that models with comparable predictive accuracy can differ substantially in the degree to which they preserve the collective structure. These results demonstrate that accuracy alone is insufficient for evaluating *PI* in democratic settings. By contributing a novel comprehensive and *collective-centric* evaluation benchmark for the task of *PI*, this work aims to support the development of AI systems that scale deliberation without compromising the integrity of its democratic outcomes.[1]

Code — https://github.com/LequeuISIR/PrefInferenceEval

1 Introduction

Participatory democracy grounds the legitimacy of collective decisions in the capacity of citizens to participate in forming those decisions (Habermas 2015; Rawls 1997; Dryzek and List 2003). Participatory democracy platforms such as Polis and Remesh promise to extend such participation beyond small, resource-intensive forums to large-scale consultations used in public policy design, participatory budgeting, and peace-building (Konya et al. 2025; Fishkin et al. 2019; Landemore 2020; Small et al. 2021). In these consultations, participants submit propositions in free form text and express support for or opposition to propositions submitted by others, resulting in a collection of textual propositions and a participant–proposition voting matrix (see Figure 1).

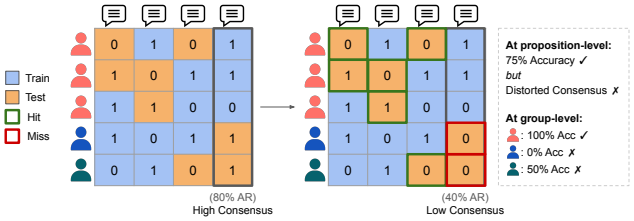

Figure 1: **Accuracy fails to reveal distortions induced by *Preference Inference*.** In this example, 75% of held-out test votes (unobserved in deployment) are correctly inferred. Despite this accuracy, inference drops a consensual proposition's approval rate (AR) from 80% to 40% and distributes errors unevenly across socio-demographic groups. This motivates evaluating *proposition-level* support and *participant-level* outcomes alongside accuracy.

However, given the volume of inputs, each participant can only engage with a small fraction of all propositions, leading to a highly sparse voting matrix. This sparsity creates the need for *Preference Inference (PI)*: the task of predicting whether a participant would support or oppose a proposition they have not reviewed (Konya et al. 2022). These predictions carry concrete implications as they reconstruct the *landscape of preferences* subsequently used to identify opinion groups, select representative propositions, and produce summaries for facilitators and decision-makers (Small et al. 2021; Revel et al. 2025; Fish et al. 2026). Prior work on deliberation summarization demonstrates that different models can produce substantially different levels of proportional representation and may erase minority voices (Zhu et al. 2025; Hafid,

[1]datasets available at https://huggingface.co/collections/democratic-commons/citizens-consultations

*These authors contributed equally.

Berriche, and Cointet 2026). Distortions introduced during *PI* may thus propagate throughout the deliberative chain and alter the resulting collective representation.

Since *PI* errors may be distributed unevenly (for instance, as a function of support size or internal group-level correlations), certain propositions risk being poorly modeled, generating spuriously consensual or non-consensual outcomes (Revel and Pénigaud 2026), or under-representing specific socio-demographic and interest groups. *PI* should therefore neither erase genuine political conflict to produce an appearance of agreement nor create an illusion of disagreement by fragmenting groups or distorting propositions around which participants genuinely agree. Deliberative theory offers a useful middle ground through the notion of *meta-consensus*: deliberation need not eliminate disagreement, but should accurately structure the values, judgments, and preferences over which disagreement persists (Dryzek 2012). A democratic *PI* model must therefore operate between two failures: artificial consensus and artificial conflict (Revel and Pénigaud 2026). We argue that it can achieve this by preserving the salient structure of the expressed *landscape of preferences* at two levels: how *propositions* are positioned through patterns of support and opposition, and how *participants* are organized into behavioral and socio-demographic groups.

Meeting this dual objective creates a corresponding evaluation problem: how can we determine whether an inference model preserves, rather than distorts, the proposition- and participant-level structures of the preference landscape? A preference matrix is expected to exhibit a clustered structure because political preferences are often organized by relatively stable shared belief systems (Converse 2006). Yet, *PI* systems are commonly assessed using classification or recommendation metrics that aggregate errors across participant–proposition pairs (Konya et al. 2022), overlooking whether inference alters proposition-level support patterns, group-specific approval profiles, or vote-based community structure. More broadly, research on fairness in inference systems has emphasized the need to connect abstract metrics to the normative requirements of particular socio-technical contexts (Deldjoo et al. 2024). As illustrated in Figure 1, a model may achieve high accuracy while still distorting the preferences of several socio-demographic groups. We therefore argue for complementing standard *user-centric* metrics with *collective-centric* metrics designed to assess whether inference preserves the landscape of preferences.[2]

To address this gap, this work makes three contributions:
(1) We construct a theoretical and empirical framework for evaluating how Preference Inference reshapes the *landscape of preferences* derived from a voting matrix. We formalize this landscape along two complementary dimensions grounded in social sciences: the positioning of *propositions* through patterns of support and opposition, and the grouping of *participants* through voting behavior and socio-demographic attributes. Building on this formalization, we develop an evaluation benchmark that complements standard predictive metrics by measuring whether inferred preferences preserve proposition-level approval rates (AR) and participant-level preference structures, including vote-based communities and socio-demographic group patterns.
(2) We contribute a new multilingual dataset comprising four consultations obtained via a partner platform. To our knowledge, it is the largest corpus of its kind, containing a total of almost 90k participants, 1M votes, and 22 distinct languages. To support reproducibility and facilitate the use of our metrics, we make our code and dataset publicly available.
(3) We conduct an extensive evaluation of a comprehensive set of Preference Inference models, showing that models with high predictive accuracy can still substantially distort patterns of consensus and preferences of particular socio-demographic groups.

[2]Algorithmic preference inference also introduces an accountability challenge when preferences do not stem directly from participation, though addressing this limitation lies beyond our scope.

2 Related Work

Participatory Democracy Platforms

Preference Inference operates within the broader pipeline through which participatory democracy platforms organize, represent, and summarize large-scale public input. These platforms typically represent participants' expressed preferences as a participant-proposition voting matrix, and use different inference strategies to deal with sparsity. Polis imputes missing entries using proposition-level mean votes before applying dimensionality reduction and clustering (Small et al. 2021), whereas Remesh combines language model representations with latent factors learned from observed voting behavior (Konya et al. 2022, 2023). The resulting inferred matrices support downstream tasks such as opinion group identification or automated summarization. Such systems have seen various high-stakes deployments: Remesh by the UN for peace-building (Ovadya 2023), and Polis in Taiwan to shape Uber and taxi regulations (Konya et al. 2023).

From a social-science perspective, Preference Inference goes beyond a mere technical operation: it attributes preferences that participants have not explicitly expressed, thereby shaping how citizens and groups become represented within the consultation and how collective preferences are made visible to decision-makers (Revel and Pénigaud 2025). PI should therefore be understood as both a predictive mechanism and a form of algorithmic representation. Inspired by Innerarity's discussion of the normative implications of predictive analytics (Innerarity 2023), we focus on one aspect of algorithmic fairness: whether prediction accuracy differs systematically across socio-demographic groups and vote-based communities. Building on this perspective, we introduce novel metrics to assess whether inferred votes preserve proposition-level support patterns and whether preferences are inferred with comparable accuracy across groups. Moreover, because multilingual data remain underexplored in existing PI benchmarks, we evaluate our framework on a new multilingual dataset comprising four consultations, over 90K participants, 1M votes, and 22 languages.

Preference Models for Collective Representation

Models Although Preference Inference is a cornerstone task for participatory democracy platforms, it long predates

these systems and has a substantial history across several research traditions. Early probabilistic models from psychometrics and conjoint analysis represented preferences through latent utilities, pairwise comparisons, and the attributes of alternatives (Thurstone 1927; Green and Srinivasan 1978). Later work introduced adaptive survey methods that collect informative responses without requiring every participant to evaluate every proposition (Salganik and Levy 2015). More contemporary approaches draw on recommender systems and related latent-variable methods for user-item data, including matrix-factorization and collaborative filtering (Koren, Bell, and Volinsky 2009), latent-factor models (Johnson et al. 2014), and regularized matrix completion (Bilich et al. 2023). These methods estimate missing responses from shared low-dimensional structure across participants and items. Recent enhancements additionally incorporate the semantic content of propositions through language model representations (Konya et al. 2022), following recent advances in NLP (Sorensen et al. 2025) and stance detection (Allaway and McKeown 2020; Barriere and Balahur 2023). Blair, Procaccia, and Tambe (2026) go further by training language models so that distances accurately represent participants' preferences. Across methods, Preference Inference must typically contend with sparsity (Koren, Bell, and Volinsky 2009) and cold-start participants and propositions (Lee et al. 2019). In this work, we show that multilingual content poses an additional challenge.

Evaluation In recommender systems, Preference Inference models are generally evaluated using classification or ranking metrics aggregated across user-item pairs, with recent work expanding to user- and item-side fairness (Deldjoo et al. 2024; Zhao et al. 2025; Ge, Delgado-Battenfeld, and Jannach 2010). However, participatory democracy demands more than the equal predictive performance common in recommender systems, as group-level parity does not guarantee faithful collective representation. For example, suppose Group A supports proposition 1 and opposes proposition 2, while Group B holds the opposite preferences. A model that predicts 50% support for both propositions in both groups will show no disparity when evaluated using a conventional group-fairness metric, because both groups are inferred equally inaccurately. Yet the inference has erased the central *structure* of the consultation: the groups originally held opposite preferences, but now appear identical. To address this gap, we introduce metrics that evaluate whether inference preserves such group-specific approval profiles and vote-based community structure. Put simply, in addition to assessing whether groups are predicted equally well, our metrics determine whether the inferred voting matrix still represents what the groups believe and how they differ.

3 Theoretical Framework

Notation Consider a democratic process $\mathcal{D}$ involving a set of N participants $\mathcal{U} = \{u_i\}_{i=1}^{N}$ and a set of M propositions $\mathcal{T} = \{t_j\}_{j=1}^{M}$. Participants' opinions are represented by a voting matrix $\mathbf{V} \in \{-1, +1\}^{N \times M}$, where $v_{ij} = +1$ if participant u_i supports proposition t_j, and $v_{ij} = -1$ if they oppose it. Because participants vote on only a subset of propositions, the democratic process provides a partially observed matrix $\mathbf{V}_{\text{obs}} \in \{-1, +1, \perp\}^{N \times M}$, where $v_{ij}^{\text{obs}} = v_{ij}$ when the vote is observed and $v_{ij}^{\text{obs}} = \perp$ otherwise.

Problem formalization: Preference Inference Given the partially observed voting matrix $\mathbf{V}_{\text{obs}}$, a Preference Inference model $\mathcal{M}$ predicts the missing entries and produces a completed voting matrix $\hat{\mathbf{V}}^{\mathcal{M}} \in \{-1, +1\}^{N \times M}$, defined by

$$\hat{v}_{ij}^{\mathcal{M}} = \begin{cases} v_{ij}^{\text{obs}}, & \text{if } v_{ij}^{\text{obs}} \neq \perp, \\ \mathcal{M}(\mathbf{V}_{\text{obs}}, u_i, t_j), & \text{if } v_{ij}^{\text{obs}} = \perp. \end{cases} \tag{1}$$

Preference inference therefore aims to estimate the unobserved votes while leaving the observed votes unchanged. In this work, we evaluate not only the accuracy of these predictions but also whether $\hat{\mathbf{V}}^{\mathcal{M}}$ preserves the collective properties of the underlying voting matrix $\mathbf{V}$.

Characterizing the Landscape of Preferences

We define the **landscape of preferences** as the set of collective structures derived from the voting matrix. We characterize it through two complementary perspectives: how **propositions** are positioned relative to participants' preferences, and how **participants** are organized through their voting behavior and social attributes.

Propositions On the proposition side, the voting matrix induces aggregate patterns of support and opposition. We define the *Approval Rate* of a proposition t_j, denoted $\text{AR}(t_j) \in [0, 1]$, as the proportion of positive votes it receives. Approval rates position propositions along the landscape of preferences: some propositions are broadly supported, some are broadly rejected, and others are contested, with approval rates close to 0.5. These proposition-level patterns are central to how consultations are interpreted, since they indicate which ideas attract consensus, opposition, or disagreement. A Preference Inference model should therefore not artificially inflate, suppress, or reorder these patterns.

Participants On the participant side, the voting matrix induces both behavioral and socio-demographic structures. First, participants may form vote-based *communities*[3] (Mason 2022): groups of users with similar voting profiles, characterized by high within-group similarity and lower between-group similarity. These communities capture distinct perspectives within the consultation and reflect the plurality of preferences expressed by participants. Second, participants may belong to socio-demographic groups defined by attributes such as age, gender, ethnicity, or location. Because inference may affect these groups differently, Preference Inference should be evaluated with respect to whether it preserves their position and representation in the landscape of preferences, rather than amplifying, attenuating, or erasing group-specific patterns.

[3]We borrow the *community* terminology from graph theory, as further explained in the next section.

Measuring the Distortion of Preferences

We select metrics that evaluate whether preference inference preserves the landscape of preferences along its two main dimensions: propositions and participants. On the proposition side, we use *Balanced Accuracy* to measure whether individual missing votes are predicted correctly despite label imbalance, and *Approval Rate Hit Ratio* to assess whether the resulting patterns of support and opposition across propositions remain statistically credible. On the participant side, we measure *Consensus Distortion* to evaluate whether inference alters the approval patterns of particular participant groups, and *Community Stability* to assess whether groups of participants with similar voting behavior remain represented after matrix completion. We complement these measures with fairness metrics for socio-demographic groups: *minimum group Balanced Accuracy* to capture the performance experienced by the worst-served group, and *Equalized Odds Ratio* to measure disparities in error rates across groups. Together, these metrics provide a comprehensive assessment of whether inference preserves both how propositions are positioned and how participants (including behavioral communities and socio-demographic groups) are represented in the preference landscape. We define each metric below.

Approval Rate Hit Ratio (ARHR) ARHR measures whether a PI model $\mathcal{M}$ provides statistically credible approval rates *at the proposition-level*. For each proposition t_j, the observed approval rate $\mathrm{AR}(t_j|\mathbf{V}_{\mathrm{obs}})$ is treated as an estimate of the approval rate in the underlying full matrix $\mathbf{V}$. Considering votes on t_j as Bernoulli trials (where success is seen as supporting a proposition), we compute a 95% Wilson confidence interval[4] (Wilson 1927) from the observed votes in $\mathbf{V}_{\mathrm{obs}}$, denoted $\mathrm{CI}_{0.95}(t_j|\mathbf{V}_{\mathrm{obs}})$. The Approval Rate Hit Ratio is then, for a given PI model $\mathcal{M}$:

$$\frac{1}{|\mathcal{T}|}\sum_{t_j\in\mathcal{T}}\mathbf{1}\left[\mathrm{AR}(t_j \mid \hat{\mathbf{V}}^{\mathcal{M}}) \in \mathrm{CI}_{0.95}(t_j \mid \mathbf{V}_{\mathrm{obs}})\right] \quad (2)$$

A system with a low ARHR outputs statistically unlikely approval rates, meaning it is likely to distort the actual approval from the participant.

Consensus Distortion (CD) We conceptualize consensus on a proposition as the distribution of approval across relevant participant groups, following work that evaluates common ground through cross-group agreement (Small et al. 2021; Konya et al. 2025; Hafid, Berriche, and Cointet 2026). Under this view, a proposition is seen as consensual only if its support is represented accurately across groups: substantially overestimating or underestimating the approval of any group may create a misleading picture of shared support or disagreement. CD therefore measures how much Preference Inference alters the *group-level approval* profile on which judgments of consensus are based. Let $\mathcal{C}$ denote a partition of $\mathcal{U}$ where each user u_i is assigned to a cluster $c_k \in \mathcal{C}$. $\mathcal{C}$ can be defined based on protected attributes, such as age or sex, or be inferred from the observed voting data. We denote $\mathrm{AR}(c_k, t_j|\mathbf{V})$ as the approval rate of the user-group c_k on t_j. We define the group-level approval-rate distortion as:

$$\Delta(c_k, t_j, \mathcal{M}) = \left|\mathrm{AR}(t_j \mid c_k, \hat{\mathbf{V}}^{\mathcal{M}}) - \mathrm{AR}(t_j \mid c_k, \mathbf{V})\right| \quad (3)$$

Because a large distortion for even one group can produce a misleading impression of cross-group consensus, the proposition-level consensus distortion is defined as the maximum group-level distortion:

$$\mathrm{CD}(t_j, \mathcal{M}) = \max_{c\in\mathcal{C}_j} \Delta(c, t_j, , \mathcal{M}) \quad (4)$$

Finally, the democratic-process-level consensus distortion $\mathrm{CD}(\mathcal{D}, \mathcal{M})$ is the average of the proposition-level maxima over the eligible set of propositions $\mathcal{T}'$.

Unlike Approval Rate Hit Ratio, which compares inferred approval rates to confidence intervals estimated from $\mathbf{V}_{\mathrm{obs}}$, Consensus Distortion is defined against the underlying matrix $\mathbf{V}$. In empirical evaluations, $\mathbf{V}$ is approximated by held-out votes. The two metrics also differ in their level of analysis: ARHR evaluates whether *proposition-level* aggregate approval rates are preserved, whereas CD evaluates whether *group-level* approval rates are preserved within each proposition before aggregating across propositions.

Community Stability (CS) We previously defined *communities* as groups of participants who display similar voting . Hypothesizing that such dynamics can be found in $\mathbf{V}_{\mathrm{obs}}$, these communities should not disappear after inference. Voting matrices can be viewed as *Signed Bipartite Graphs* (SBG)—graphs with two distinct sets of nodes (participants and propositions) and edges existing only between the two sets (preferences). A *Community* in graph theory refers to a cluster of nodes strongly interconnected and weakly connected to others. In the *PI* setup, it corresponds to participants with similar voting behaviors. The *Community Stability* metric evaluates how these communities evolve after inferring the unobserved preferences of their population. More specifically, we only consider the subset of users who expressed at least $N_{\min} = 10$ opinions on propositions, and apply agglomerative clustering[5] (Müllner 2011) on the signed bipartite spectral representation (Kunegis et al. 2010) of the sub-graph to find the optimal number of clusters k^* maximizing the silhouette score (Rousseeuw 1987). We then apply the same method with k^* clusters on the same user subset, but on the fully inferred matrix $\hat{\mathbf{V}}^{\mathcal{M}}$. We compare the resulting clustering using the Adjusted Rand Index (Hubert and Arabie 1985). A higher score indicates community stability, while a low score hints at their disappearance.

Other Metrics We compute additional standard metrics. First, we use *Balanced Accuracy* (**B-Acc**)—the average of the True Positive Rate (aka recall) and True Negative Rate—as it is well-suited for imbalanced datasets (Brodersen et al. 2010). Moreover, when protected attributes are available, we

[4] We use the Wilson score interval as it correctly handles cases with small sample size and extreme probability values. More details are given in the Supplementary Material.

[5] Agglomerative clustering is favored over other methods as it has no stochastic component.

Platform	Dataset	Topic	# Votes	# Participants	# Propositions	# Attributes	Lang
Polis	austria-climate	Energy	133,048	1,756	1,039	–	DE
	bowling-green	BGKY Improvement	226,136	2,031	896	–	EN
	15-per-hour-seattle	Seattle Minimum Wage Impact	2,995	339	54	–	EN
	march-on	Operation Marching Orders	536,923	6,289	2,160	–	EN
	bank-farmland	Land Use and Conservation	33,530	404	293	–	EN
	vtaiwan.uberx	Taiwan Uber Regulation	49,996	1,921	197	–	EN
	bg2050	Bowling Green Future	14,341	126	371	–	EN
Remesh	right-to-assemble	Right to Assemble	7,547	307	1,296	4	EN
	foreign-intervention	US Foreign Relations	4,520	308	767	–	EN
	campus-protests	University Campus Protests	6,066	309	1,020	7	EN
Anonymized	steuer-debate	Taxation	16,204	1,050	25	4	DE
	gc-sante	Healthcare	226,010	25,058	1,331	–	FR
	ingerence	Foreign Digital Interference	103,671	8,968	559	–	FR
	eurhope	Future of Europe	892,273	54,421	4,366	–	22

Table 1: Datasets statistics. The following datasets comprise protected socio-demographic attributes as metadata: *right-to-assemble*: age, ethnicity, gender, politics; *campus-protests*: age, education, gender, income, living-area, politics, religion; *steuer-debate*: age, gender, living-area, politics.

compute fairness-aware metrics: **Min. B-Acc** is simply the *Minimum Balanced Accuracy* across group subsets (Weerts et al. 2023), reflecting a Rawlsian principle of evaluating outcomes for the worst-served group (Rawls 1971). *Equalized Odds Ratio* (**EOR**; Hardt, Price, and Srebro, 2016) is defined as the minimum of the True Positive Rate and False Positive Rate ratios across groups, where each ratio measures the disparity between the largest and smallest rates.

$$\min\left(\frac{\min_{c\in\mathcal{C}}\mathrm{TPR}(c)}{\max_{c\in\mathcal{C}}\mathrm{TPR}(c)}, \frac{\min_{c\in\mathcal{C}}\mathrm{FPR}(c)}{\max_{c\in\mathcal{C}}\mathrm{FPR}(c)}\right) \tag{5}$$

An EOR of 1 indicates that all groups achieve identical TP, TN, FP, and FN rates.

4 Experiments and Results

Datasets

We select datasets from two distinct online citizen consultation platforms—namely, Polis[6] and Remesh.[7] To these, we add four newly compiled datasets sourced from a large-scale international civic technology platform (cf. Table 1 for details on all used datasets). We contribute this novel corpus alongside our work to address limitations in the existing datasets. First, it provides non-English evaluation data (French, German, and 22 languages in *eurhope*[8]), countering the English-heavy bias of existing resources and enabling future multilingual analyses. Second, it captures national and transnational policy debates—such as taxation and foreign digital interference—at a generally larger scale than previous localized datasets. Finally, it offers diverse structural properties for evaluation, ranging from highly dense consultation matrices (e.g., over 50% interaction density in *steuer-debate*) to massive, sparse environments (*eurhope* alone comprises over 892k votes and 54k participants).

[6]https://github.com/compdemocracy/openData

[7]https://github.com/akonya/polarized-issues-data

[8]bg, cs, da, de, el, en, es, et, fi, fr, hr, hu, it, lt, lv, nl, pl, pt, ro, sk, sl, and sv (ISO 639-1)

To unify the datasets across the different platforms, we restrict our analysis to *agree* and *disagree* votes, discarding *neutral* votes (unavailable in Remesh) and pairwise preferences (exclusive to Remesh). For protected socio-demographic attributes, to ensure that the computed statistics are reliable, we only consider groups with $|c| \geq 25$. Statistics of the post-processed datasets can be found in the Supplementary Material.

Models

We evaluate a comprehensive set of Preference Inference models spanning the main methodological families identified in the related work.

Collaborative Filtering (CF) CF predicts preferences using historical interaction patterns rather than explicit metadata. User-based models (**CF-user**) infer unseen votes by identifying individuals with highly correlated voting histories. Proposition-based models (**CF-prop**) calculate similarity between the propositions themselves based on overlapping ratings. The algorithm then aggregates these proximity metrics across the interaction matrix to infer the preferences.

Nearest Neighbor (NN) NN uses the embeddings of propositions from pretrained encoders. It relies on the text embedding $\mathbf{e}$ of the propositions by assigning to $\hat{v}_{ij}$ the opinion v_{ij*} of the closest proposition in the embedding space for which the user already expressed a preference:

$$j* = \mathrm{argmax}_{j'\neq j}\mathbf{e}_{t_j}\cdot\mathbf{e}_{t_{j'}} \quad \forall t \in \mathcal{T}_{u_i} \tag{6}$$

where $\mathcal{T}_{u_i}$ is the set of propositions on which u_i voted.

Factorization Machines (FMs) FMs learn representations of users and propositions from existing preferences to infer new ones. Standard FMs initialize both user and proposition embeddings randomly in $\mathbb{R}^P$ and optimize them such that the utility for a pair (u_i, t_j), mapped via the sigmoid function $\sigma(u_i \cdot t_j)$, approximates 1 if user u_i agrees with proposition t_j, and 0 otherwise.

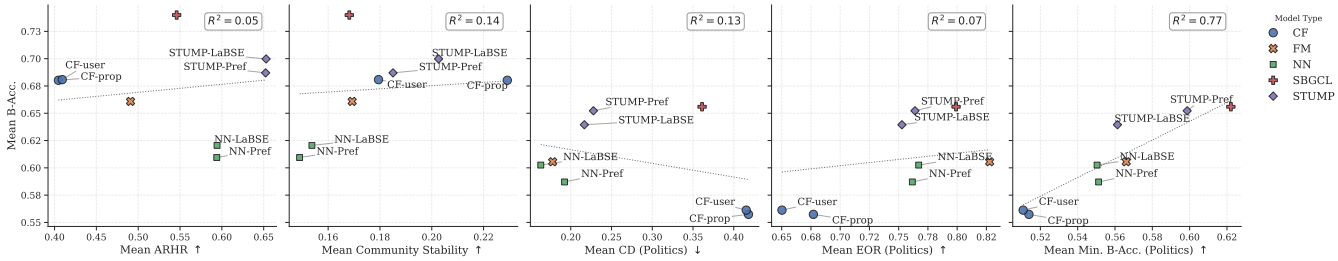

Figure 2: B-Acc vs. ARHR, CS, CD, EOR and Min. B-Acc for the *politics* protected attribute. CD, EOR and Min. B-Acc, are averaged over the three datasets with protected attributes and the other metrics are averaged over all datasets for each model. Regressions are computed over all (model, dataset) pairs, and not the displayed averages.

Additionally, we implement **STUMP** (Konya et al. 2022), an FM variant tailored for preference inference. STUMP initializes user representations $u_i \in \mathbb{R}^P$ randomly[9] and uses pretrained proposition embeddings $\mathbf{e}_{t_j} \in \mathbb{R}^L$ as initialization. It then learns user embeddings alongside a linear projection matrix $\mathbf{W}_t \in \mathbb{R}^{L \times P}$ for propositions, and computes the utility of (u_i, t_j) as $\hat{v}_{ij} = \sigma(u_i \cdot \mathbf{W}_t^T \mathbf{e}_{t_j})$, where σ denotes the sigmoid function.[10]

SBG link-prediction As stated previously, consultations can be viewed as Signed Bipartite Graphs (SBG). Recent works have explored machine learning for link prediction in SBG. We implement **SBGCL** (Zhang et al. 2023), a Graph Neural Network which uses a two-level view of the graph (the explicit inter-sets and implicit intra-set relation of nodes) and applies augmentation techniques and contrastive learning to learn representations of the nodes.

LLM-based methods Modern recommendation approaches leverage in-context learning and LLMs (Brown et al. 2020). In this framework, an LLM is prompted with a simple instruction (cf. Supplementary Material) containing the user's observed positive and negative votes and asked to predict whether the user would approve of a new proposition (returning 1 for approval, 0 otherwise). The final prediction corresponds to the most likely token between the two[11]:

$$\hat{v}_{ij} = \text{argmax}_{v \in \{0,1\}} \left(\hat{P}(v \mid t_j, \mathcal{T}_{u_i}) \right) \tag{7}$$

Experimental Protocol

For each consultation, votes ($\mathbf{V}_{\text{obs}}$) are randomly partitioned into a 70-15-15 train-validation-test split. For learned models, we conduct an extensive grid-search over learning rates and model-specific parameters (see Supplementary Material for the search space and optimal configurations). We select the model checkpoint that maximizes *B-Acc* on the validation set after a maximum of 400 training epochs, and evaluate its performance on the corresponding test set. For B-Acc, EOR, Min. B-Acc, and CD, we report results on the test set, a subset of $\mathbf{V}_{\text{obs}}$. ARHR and CS are computed on the full $\mathbf{V}^{\mathcal{M}}$.

We compare two text-embedding models: (i) **LaBSE** (Feng et al. 2022), a 110M-parameter multilingual model; and (ii) **Pref** (Blair, Procaccia, and Tambe 2026), a 1.2B-parameter model based on Sentence-T5 (Ni et al. 2022) that is specifically trained for preference representation. For LLM recommendation, we use the open-weight post-trained model Qwen3-8B (Team 2025).

Results

We present per-model averages throughout this section[12] (detailed breakdowns are available in the Supplementary Material). Our evaluation addresses two central questions: (i) Does higher balanced accuracy necessarily imply a less distorted landscape of preferences? (ii) Do existing PI methods preserve opinion structures adequately, or is further improvement necessary?

Accuracy vs. Opinion Distortion Figure 2 presents B-Acc compared to distortion and fairness metrics per model. Although SBGCL achieves the highest B-Acc overall, it falls short on our proposed metrics, ranking 5th on ARHR and CS and 6th on CD. More generally, we find little to no correlation between B-Acc and these metrics ($R^2 = 0.05$ for ARHR, $R^2 = 0.14$ for CS, $R^2 = 0.13$ for CD on the political attribute)[13], with the exception of Min. B-Acc, which is naturally strongly correlated. Figure 4 displays model performance on the attribute-based metrics (EOR, Min. B-Acc and CD) across the three common attributes (Gender, Age, Politics), further supporting the finding that SBGCL lags behind other models despite its high B-Acc.

Are Existing PI methods Sufficient? No model noticeably outperforms the others across the board. NN models achieve strong CD and ARHR results, but exhibit limited B-Acc and score the lowest on CS. Both collaborative filtering models show high B-Acc. scores—particularly impressive

[9]We experimented with initializing user representations from the embeddings of propositions they voted on, but this did not improve performance.

[10]STUMP also learns a context-specific projection for consultations. As these contexts are only available on the Remesh datasets, we discard them and remove the projection.

[11]Because LLM recommendation is not scalable to large consultations (e.g. eurhope necessitate 250M inference), we only report its results on the three datasets with protected attributes.

[12]All results are scaled by $\times 10^2$.

[13]We provide spearman correlation between all metrics used in this work in the supplementary materials.

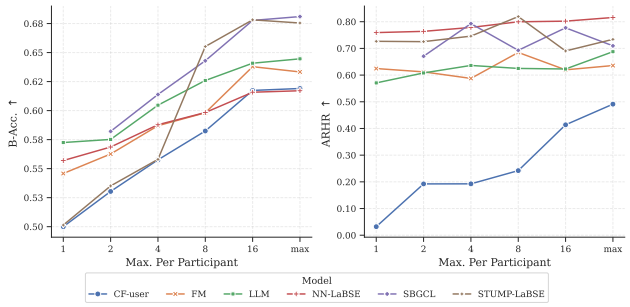

Figure 3: B-Acc. and ARHR averaged over three datasets across varying numbers of examples per user.

given that they are not learned—yet obtain the poorest results on almost all distortion and fairness metrics. Overall, STUMP-based models provide the best trade-off, showing accuracy results slightly below SBGCL, but at the same time distorting the preference landscape significantly less. In general, current results regarding preference landscape distortion fall short of acceptable standards for democratic processes. No model produces statistically plausible approval rates (at a 95% confidence interval); specifically, no model achieves an ARHR score above 0.95 on any dataset (cf. Supplementary Material). Similarly, the lowest average CD is 21.7 (obtained by FM) which is arguably too high for democratic processes. Finally, CS remains consistently low, with the best performing model (CF) achieving an average of only 22.9. Additionally, a per-language analysis on *Eurhope* (reported in the Supplementary Material) shows disparities in treatment. Ultimately, these findings reveal a critical limitation in current preference inference methods, underscoring the need for improved approaches that prioritize faithful representation and preservation of the underlying preference landscape alongside predictive accuracy.

Data Efficiency Following Konya et al. (2022), we evaluate the effect of preference sparsity by limiting the number of observed preferences per participant (1, 2, 4, 8, 16, or all available)[14] on the three datasets with socio-demographic attributes. Figure 3 reports B-Acc and ARHR (results for the other metrics are provided in the Supplementary Material). B-Acc, Min. B-Acc, and CD consistently improve as more preferences become available, an important result not observed when evaluating using only accuracy (Konya et al. (2022)).[15] This further highlights the need for collective-centric metrics beyond accuracy. In contrast, ARHR, EOR, and CS remain largely unchanged, with the latter exhibiting noisy but uniformly low values (0–0.18). These findings suggest that several limitations of current preference inference models are structural rather than a consequence of sparsity.

[14] SBGCL requires at least two observed preferences per participant and is therefore not evaluated with one.

[15] Konya et al. (2022) report no improvement in accuracy above five preferences per participant, but do not evaluate using collective-centric metrics.

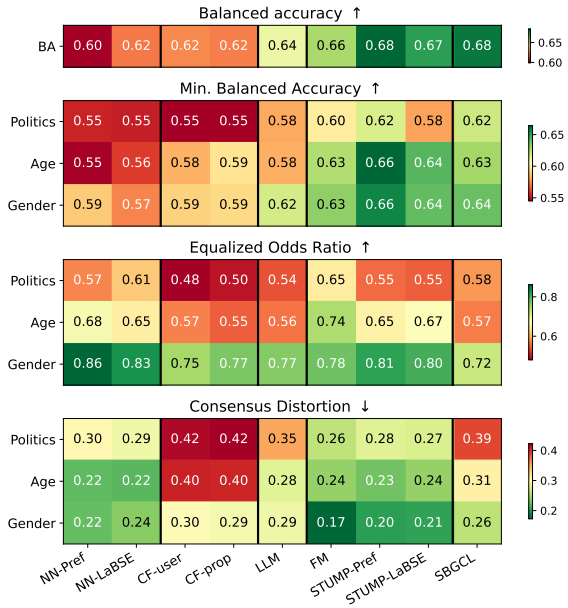

Figure 4: Equalized Odds Ratio (↑), Min. B-Acc (↑) and CD (↓) averaged over campus-protests, right-to-assemble, and steuer-debate. B-Acc is shown for reference.

5 Conclusion

Participatory democracy platforms enable collective decision-making among thousands of participants, yet the resulting voting data are often highly sparse. Preference inference is therefore essential for reconstructing the preference landscape and supporting downstream decision-making.

We introduce a collective-centric evaluation framework, grounded in deliberative democratic theory, to assess whether preference inference systems faithfully preserve the collective structure of expressed preferences rather than merely predicting individual votes accurately. Our findings demonstrate that preference inference is not a neutral computational step: models with high predictive accuracy can substantially distort patterns of consensus and the preferences of particular groups. Concerningly, across all evaluated consultations, no existing method faithfully preserves the underlying preference landscape. Alongside this evaluation framework, we release a new multilingual benchmark spanning four consultations, nearly 90,000 participants, one million votes, and 22 languages. In particular, the Eurhope consultation constitutes the largest publicly available consultation dataset to date, comprising eight times more participants and twice as many propositions as the largest previously available benchmark, while being the first multilingual.

This work establishes collective-centric evaluation as a standard objective for future preference inference systems, encouraging models that optimize not only predictive accuracy but also faithful collective representation.

Limitations and Ethical Considerations

Although our framework provides principles for evaluating preference inference systems, it should not be interpreted as

legitimizing preference inference as a substitute for genuine democratic participation. Preference inference can only complement citizen consultations whose design already ensures broad, representative, and meaningful participation. Moreover, our evaluation framework operationalizes a specific set of democratic principles grounded in deliberative theory, but these principles are not exhaustive. Alternative conceptions of deliberation and democratic representation may motivate additional evaluation criteria.

Acknowledgments

This research was funded by BPI-France under the project AI For Democracy - Democratic Commons, one of seven winners of BPI-France's 'Digital Commons for Generative AI' call for projects, conducted as part of the France 2030 investment plan. We thank the members of the *Democratic Commons* project, especially Manon Berriche, for their help and careful proof-reading. This work was performed using HPC resources from GENCI–IDRIS (Grant 2026-A0201017543).

References

Allaway, E.; and McKeown, K. 2020. Zero-Shot Stance Detection: A Dataset and Model Using Generalized Topic Representations. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8913–8931. Online: Association for Computational Linguistics.

Barriere, V.; and Balahur, A. 2023. Multilingual multi-target stance recognition in online public consultations. *Mathematics*, 11(9): 2161.

Bilich, J.; Varga, M.; Masood, D.; and Konya, A. 2023. Faster peace via inclusivity: An efficient paradigm to understand populations in conflict zones. *arXiv preprint arXiv:2311.00816*.

Blair, C.; Procaccia, A. D.; and Tambe, M. 2026. Embeddings for Preferences, Not Semantics. arXiv:2605.08360.

Brodersen, K. H.; Ong, C. S.; Stephan, K. E.; and Buhmann, J. M. 2010. The Balanced Accuracy and Its Posterior Distribution. In *2010 20th International Conference on Pattern Recognition*, 3121–3124.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models Are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Converse, P. E. 2006. The nature of belief systems in mass publics (1964). *Critical review*, 18(1-3): 1–74.

Deldjoo, Y.; Jannach, D.; Bellogin, A.; Difonzo, A.; and Zanzonelli, D. 2024. Fairness in recommender systems: research landscape and future directions. *User Modeling and User-Adapted Interaction*, 34(1): 59–108.

Dryzek, J. S. 2012. *Foundations and frontiers of deliberative governance*. Oxford University Press.

Dryzek, J. S.; and List, C. 2003. Social choice theory and deliberative democracy: A reconciliation. *British journal of political science*, 33(1): 1–28.

Feng, F.; Yang, Y.; Cer, D.; Arivazhagan, N.; and Wang, W. 2022. Language-agnostic BERT Sentence Embedding. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 878–891. Dublin, Ireland: Association for Computational Linguistics.

Fish, S.; Gölz, P.; Parkes, D.; Procaccia, A.; Rusak, G.; Shapira, I.; and Wuthrich, M. 2026. Generative social choice. *Journal of the ACM*, 73(2): 1–52.

Fishkin, J.; Garg, N.; Gelauff, L.; Goel, A.; Munagala, K.; Sakshuwong, S.; Siu, A.; and Yandamuri, S. 2019. Deliberative democracy with the online deliberation platform. In *The 7th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019)*, 1–2.

Ge, M.; Delgado-Battenfeld, C.; and Jannach, D. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*, 257–260.

Green, P. E.; and Srinivasan, V. 1978. Conjoint analysis in consumer research: issues and outlook. *Journal of consumer research*, 5(2): 103–123.

Habermas, J. 2015. *Between facts and norms: Contributions to a discourse theory of law and democracy*. John Wiley & Sons.

Hafid, S.; Berriche, M.; and Cointet, J.-P. 2026. Algorithmic Approaches to Opinion Selection for Online Deliberation: A Comparative Study. In *Pluralistic Alignment Workshop at ICML 2026*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Hubert, L.; and Arabie, P. 1985. Comparing partitions. *Journal of classification*, 2(1): 193–218.

Innerarity, D. 2023. Predicting the past: a philosophical critique of predictive analytics. *IDP: revista de Internet, derecho y política= revista d'Internet, dret i política*, (39): 2.

Johnson, C. C.; et al. 2014. Logistic matrix factorization for implicit feedback data. *Advances in Neural Information Processing Systems*, 27(78): 1–9.

Konya, A.; Qiu, Y. L.; Varga, M. P.; and Ovadya, A. 2022. Elicitation Inference Optimization for Multi-Principal-Agent Alignment. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Konya, A.; Thorburn, L.; Almasri, W.; Leshem, O. A.; Procaccia, A.; Schirch, L.; and Bakker, M. 2025. Using collective dialogues and AI to find common ground between Israeli and Palestinian peacebuilders. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 312–333.

Konya, A.; Turan, D.; Ovadya, A.; Qui, L.; Masood, D.; Devine, F.; Schirch, L.; Roberts, I.; and Forum, D. A. 2023. Deliberative Technology for Alignment. *arXiv preprint arXiv:2312.03893*.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8): 30–37.

Kunegis, J.; Schmidt, S.; Lommatzsch, A.; Lerner, J.; De Luca, E. W.; and Albayrak, S. 2010. Spectral analysis of signed graphs for clustering, prediction and visualization. In *Proceedings of the 2010 SIAM international conference on data mining*, 559–570. SIAM.

Landemore, H. 2020. Open democracy: Reinventing popular rule for the twenty-first century.

Lee, H.; Im, J.; Jang, S.; Cho, H.; and Chung, S. 2019. Melu: Meta-learned user preference estimator for cold-start recommendation. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 1073–1082.

Mason, L. 2022. *Uncivil agreement: How politics became our identity*. University of Chicago Press.

Müllner, D. 2011. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*.

Ni, J.; Hernandez Abrego, G.; Constant, N.; Ma, J.; Hall, K.; Cer, D.; and Yang, Y. 2022. Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 1864–1874. Dublin, Ireland: Association for Computational Linguistics.

Ovadya, A. 2023. 'Generative CI'through Collective Response Systems. *arXiv preprint arXiv:2302.00672*.

Rawls, J. 1971. A Theory of Justice: Original Edition.

Rawls, J. 1997. The idea of public reason revisited. *The university of Chicago law review*, 64(3): 765–807.

Revel, M.; Milli, S.; Lu, T.; Watson-Daniels, J.; and Nickel, M. 2025. Representative ranking for deliberation in the public sphere. *arXiv preprint arXiv:2503.18962*.

Revel, M.; and Pénigaud, T. 2025. AI-Facilitated Collective Judgements. arXiv:2503.05830.

Revel, M.; and Pénigaud, T. 2026. AI-Enhanced Deliberative Democracy and the Future of the Collective Will. *Philosophy & Technology*, 39(2): 81.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20: 53–65.

Salganik, M. J.; and Levy, K. E. 2015. Wiki surveys: Open and quantifiable social data collection. *PloS one*, 10(5): e0123483.

Small, C.; Bjorkegren, M.; Erkkilä, T.; Shaw, L.; and Megill, C. 2021. Polis: Scaling deliberation by mapping high dimensional opinion spaces. *Recerca: revista de pensament i anàlisi*, 26(2).

Sorensen, T.; Mishra, P.; Patel, R.; Tessler, M. H.; Bakker, M. A.; Evans, G.; Gabriel, I.; Goodman, N.; and Rieser, V. 2025. Value profiles for encoding human variation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2047–2095.

Team, Q. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Thurstone, L. 1927. A law of comparative judgment. *Psychological Review*, 34(4).

Weerts, H.; Dudík, M.; Edgar, R.; Jalali, A.; Lutz, R.; and Madaio, M. 2023. Fairlearn: Assessing and Improving Fairness of AI Systems. *Journal of Machine Learning Research*, 24.

Wilson, E. B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158): 209–212.

Zhang, Z.; Liu, J.; Zhao, K.; Yang, S.; Zheng, X.; and Wang, Y. 2023. Contrastive Learning for Signed Bipartite Graphs. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, 1629–1638. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9408-6.

Zhao, Y.; Wang, Y.; Liu, Y.; Cheng, X.; Aggarwal, C. C.; and Derr, T. 2025. Fairness and diversity in recommender systems: a survey. *ACM Transactions on Intelligent Systems and Technology*, 16(1): 1–28.

Zhu, S.; Yang, S.; Bakker, M. A.; Pentland, A.; and Pei, J. 2025. Can AI Truly Represent Your Voice in Deliberations? A Comprehensive Study of Large-Scale Opinion Aggregation with LLMs. *arXiv preprint arXiv:2510.05154*.