PoliSpace: Political Opinion Prediction using Shared Latent Space

Johanna D'Ciofalo Khodaverdian[1,2], Frédéric Berdoz[2], Roger Wattenhofer[2]

[1]KTH Royal Institute of Technology, [2]ETH Zürich

jdk@kth.se, fberdoz@ethz.ch, wattenhofer@ethz.ch

Abstract

Modeling and predicting political opinions remains a challenging task for AI systems. We introduce POLISPACE, a framework that embeds individuals and political questions into a shared latent space, where vector alignment directly reflects agreement. Individuals are represented as learnable vectors, and questions are encoded via a contrastively fine-tuned language model with individual-specific affine transformations to capture subjective interpretation. This design yields an embedding space that is robust to question rephrasing and sensitive to negation. Evaluated on real-world voting data from Swiss national elections, POLISPACE achieves an 89.81% F1 score on seen opinion prediction for rephrased questions, outperforming prior embedding-based approaches (66.17% F1). We further introduce retrieval and confidence-aware thresholding mechanisms that reach up to 100% F1 at reduced coverage, demonstrating how the learned embedding space can be leveraged to improve prediction reliability. Ablation studies confirm the importance of both individual-specific interpretation modeling and negation-aware encoder fine-tuning. The learned individual embedding space also exhibits clear clustering by political party affiliation, despite no party-label supervision during training, capturing the underlying ideological structure of the Swiss political landscape.

1 Introduction

While AI has reached human-level competence in several areas, a task that remains challenging is modeling and predicting political opinions. Reliable opinion prediction holds promise for addressing long-standing challenges in democracy, including the representation gap and limited political knowledge (Achen and Bartels, 2017).

Recent advancements in LLMs have opened new opportunities for computational opinion modeling, with approaches ranging from prompting and personalized LLMs to learned representations of individuals and beliefs. In some approaches, however, the LLM itself ultimately produces the opinion prediction, making the mechanism underlying a particular prediction difficult to interpret. Representation learning can be used to offer a different way of approaching this problem, by making the structure used for prediction explicit. Recent work has demonstrated that beliefs can be represented in latent spaces that capture meaningful relationships between beliefs and support downstream opinion prediction (Lee et al., 2025). In this work, we focus specifically on learning a representation space for opinion prediction, where predictions can be derived directly from the geometry of the space.

We introduce POLISPACE, a framework that embeds individuals and political questions in a shared latent space. POLISPACE represents individuals as learnable vectors and utilizes LLM-based question encodings. By jointly optimizing individual and question representations so that alignment directly reflects agreement, the framework provides a transparent and efficient prediction mechanism. To tailor the representation space to opinion prediction, POLISPACE learns individual-specific transformations to account for differences in question interpretation and a contrastively fine-tuned text encoder to handle linguistic variations in question formulations and negations.

Our primary contributions are summarized as follows:

- **We introduce an interpretable representation-learning framework that learns reusable embeddings of individuals and political questions for opinion prediction.** Agreement is directly represented by the alignment between individual and question embeddings. We evaluate our model on rephrased and original questions from real-

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 5447–5463

November 6-10, 2026 ©2026 Association for Computational Linguistics

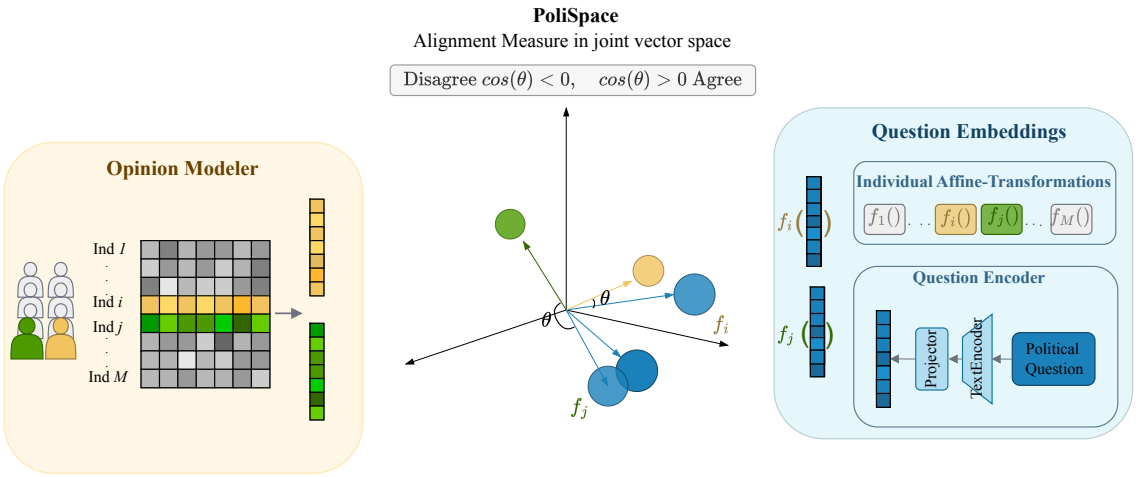


Figure 1: **Overview of POLISPACE architecture.** POLISPACE embeds individuals and political questions into a z-dimensional space. Questions are encoded by a fine-tuned Flan-T5-Base (Chung et al., 2024) followed by a three-layer MLP projector. For each individual, a learned affine transformation ($W \in \mathbb{R}^{z \times z}$, $b \in \mathbb{R}^z$) is applied to the encoder output to produce a personalized question embedding. Individuals are represented as z-dimensional learnable vectors, optimized so that agreement corresponds to positive cosine similarity with the question embedding. At inference, the model predicts *Agree* when cosine similarity exceeds zero.

world Swiss election data from candidates in the canton of Zürich, achieving an 89.81% F1 score on seen opinion prediction for indirectly seen questions. This outperforms prior work by Lee et al. (2025), which exhibits limited robustness to question rephrasing.

- **We propose two additional capabilities enabled by our embedding structure:** retrieval and confidence-aware thresholding. These mechanisms allow POLISPACE to achieve up to 100% F1 score on seen opinion prediction for indirectly seen questions at reduced coverage, providing a principled trade-off between prediction accuracy and coverage.

- **We show that individual embeddings recover the Swiss political ideological landscape** without using political party labels during training, and that question embeddings reflect answer-correlation structure in the data.

2 Related work

Prior work has addressed opinion prediction using different approaches, ranging from traditional predictive models to LLM-based personalization and representation learning.

Legislative and survey-based prediction. Albuquerque and Cajueiro (2025) predict deputies' votes by modeling legislative voting as a time-series process. They incorporate 219 features including a proposition's theme, party affiliation, and past voting history to train ensemble classifiers to predict deputies' votes on new plenary rolls, reporting an accuracy of ∼92%.

LLM personalization and steering. A subset of methods are prompt-based, conditioning LLMs on user-specific demographic or ideological attributes to simulate responses to political questions (Yang et al., 2024; Zhang et al., 2024; Sanders et al., 2023). Alternatively, Li et al. (2024) propose a method to learn virtual tokens from individuals' survey responses via prefix-tuning, steering an LLM toward specific personal views to predict answers to unseen questions with an accuracy of 60.78%. However, as these methods rely on LLMs for prediction, the underlying prediction mechanism remains difficult to interpret.

Representation learning for beliefs. Lee et al. (2025) tackle opinion prediction through a representation learning approach, with the broader goal of understanding human beliefs and their interplay. The authors fine-tune an LLM using triplet loss, where belief statements that frequently co-occur form positive pairs and non-co-occurring statements form negative pairs. They also explore the task of opinion prediction, representing an individual i as the average of their belief embeddings.

Given a belief $\mathbf{b}$, their stance is predicted as

$$\mathbf{b}' = \arg \min_{\mathbf{b} \in \{\mathbf{b}_{PRO}, \mathbf{b}_{CON}\}} d(\mathbf{i}, \mathbf{b}) \quad (1)$$

where $d(\cdot, \cdot)$ denotes a distance metric. They report an F1 score of 59% on unseen opinion prediction.

Our aim is to develop an *interpretable* framework for opinion prediction. Hence, we use a representation learning-based approach similar to Lee et al. (2025). In contrast to Lee et al. (2025), our method is developed focusing solely on the task of predicting opinions. Rather than representing individuals as averages of belief embeddings, POLISPACE learns individual embeddings such that alignment with a question embedding indicates agreement. Our representation space additionally accounts for subjective question interpretation, as well as question rephrasings and negations, such that the geometry of the space consistently reflects agreement across different formulations of a question. Unlike Lee et al. (2025), we also do not use any frequency matching between beliefs to construct the embedding space.

3 Methodology

This section introduces POLISPACE, a framework for embedding political opinions and questions in a shared vector space.

3.1 Problem Formulation

Given a dataset consisting of M individuals $\mathcal{I}$ and a set of N political questions $Q = \{q_n\}_{n=1}^{N}$, each individual $i \in \mathcal{I}$ has an answer vector $\mathbf{a}_i = (a_i^1, a_i^2, \ldots, a_i^N)$ where $a_i^n \in \{0, 1\}$ denotes a binary response to question q_n: 1 indicates agreement and 0 indicates disagreement. Answers may be either observed or unobserved. Our goal is to embed individuals and questions in a shared z-dimensional vector space such that an individual's answer to a given question can be predicted via the cosine similarity between their respective embeddings.

3.2 PoliSpace

The overall framework of POLISPACE is illustrated in Figure 1.

Question embeddings: Questions are encoded in a two-stage manner. A Question Encoder, consisting of a text encoder and an MLP Projector, generates initial question embeddings $\mathbf{e_{q_n}}$. The text encoder has been fine-tuned using a contrastive objective (Hadsell et al., 2006; Gao et al., 2021) with rephrased and negated questions to ensure the embedding space exhibits semantic diversity and negation sensitivity. This ensures different questions occupy different regions of the space while forcing a question and its negation into different representations, a distinction standard encoders often fail to maintain (Kassner and Schütze, 2020). These two properties are essential for correctly modeling individual opinions using our proposed measure of alignment. Second, to model subjective interpretations, POLISPACE learns an individual-specific affine transformation

$$f_i(\mathbf{e_{q_n}}) = W_i \, \mathbf{e_{q_n}} + \mathbf{b_i}, \quad \forall \, i \in \mathcal{I}, \quad (2)$$

applied to the initial question embedding $\mathbf{e_{q_n}}$, yielding a personalized question representation $\mathbf{e^i_{q_n}}$.

To mitigate the high parameter counts of full-rank transformation matrices $W_i \in \mathbb{R}^{z \times z}$ as z increases, we also explore a parameter-efficient variant. We factorize the weight matrix using low-rank factorization:

$$W_i = U_i V_i^\top + I_z, \quad (3)$$

where $U_i, V_i \in \mathbb{R}^{z \times r}$ are low-rank matrices with rank $r \ll z$, and I_z represents a residual identity connection for training stability. This reduces the parameter footprint per individual from $z^2 + z$ to $2zr + z$.

Opinion modeler: Each individual $i \in \mathcal{I}$ is represented as a learnable vector $\mathbf{e_i}$ stored in an embedding lookup table, initialized with Xavier uniform initialization. These embeddings are optimized on observed voting history such that the alignment between an individual's vector and a personalized question embedding directly measures agreement. At inference, the model predicts *Agree* or *Disagree* based on the cosine similarity between the individual and personalized question embeddings:

$$\hat{y}(q_n, i) = \begin{cases} 1, & \text{if } \cos(\mathbf{e_i}, \mathbf{e^i_{q_n}}) > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

4 Experimental Setup

4.1 Dataset

We evaluate POLISPACE on data from the 2023 Swiss National Council elections, collected by smartvote (Politools, 2025), a Swiss Voting Advice Application (VAA). smartvote provides voting

recommendations by asking candidates and voters to complete a questionnaire of political questions; responses are used to rank candidates by proximity to the voter's positions.

We use the 60 Standard questions from the smartvote questionnaire as our question set $Q_O = \{q_n\}_{n=1}^{60}$ (English formulation). These are free-form yes/no policy questions answered on a 4-point Likert scale (0 = No, 25 = Rather no, 75 = Rather yes, 100 = Yes), which we binarize by mapping values above 50 to *Agree* (1) and values at or below 50 to *Disagree* (0). We focus on the M=1,029 candidates from the canton of Zürich, who are required to answer all questions.

In addition to the Standard questions, we use GPT-4 (OpenAI et al., 2024) with few-shot prompting to construct additional question sets supporting text encoder fine-tuning and evaluation (prompts provided in Appendix C). To fine-tune the text encoder with a contrastive objective (Hadsell et al., 2006; Gao et al., 2021), we generate two variants of Q_O: a rephrased set Q_R^{Train}, where sentence structure and vocabulary are modified while preserving semantic content, and a negated set Q_N, where key directional terms (e.g., "increase", "agree") are inverted to express the opposite stance. For evaluation on unseen rephrasings, we construct a held-out rephrased set Q_R^{Eval} using the same procedure as Q_R^{Train}. All generated variants were manually reviewed for correctness.

4.2 Evaluation Scenarios

From the 60 questions in Q_O, 8 are held out entirely for testing; the remaining 52 are split per-individual into training and validation sets (90/10 split). The same train-validation split is applied to Q_R^{Eval} to ensure consistency across question variants. We train PoliSpace on the training split of Q_O, while evaluation is conducted on the remaining sets, corresponding to four scenarios:

1. **Scenario 1: Directly Seen Questions - Unobserved Answers**: predicting answers to questions seen during training where the individual's stance is unknown.

2. **Scenario 2: Indirectly Seen Questions - Observed Answers**: predicting answers to rephrased versions of questions where the individual's stance is known.

3. **Scenario 3: Indirectly Seen Questions - Unobserved Answers**: predicting answers to rephrased versions of questions seen during training where the individual's stance is unknown.

4. **Scenario 4: Unseen Questions - Unobserved Answers**: predicting opinions on entirely new questions.

4.3 Baselines

We compare PoliSpace against five baseline approaches: **Random**: Predicting *Agree* or *Disagree* uniformly at random; **Majority Class**: Predicting the majority class per individual, based on their answer distribution observed during training; **In-Context Learning**: We evaluate three LLMs: the closed-source GPT-4 Turbo (snapshot 2024-04-09) (OpenAI et al., 2024) and GPT-5.5 (OpenAI, 2026) models, and the open-source Qwen3-8B model (Yang et al., 2025). For each model, we provide the individual's training pairs as context and prompt it to predict whether the individual would *Agree* or *Disagree* to a new question. This baseline is evaluated on one randomly selected individual from each of the 11 political parties present in the set of individuals; **Belief-Embedding**: The method introduced by Lee et al. (2025), implemented using the authors' original code repository and applied to our dataset. We trained models using both cosine and Euclidean distance metrics, with Euclidean distance yielding the best performance (see Appendix E); **Collaborative Filtering** (Goldberg et al., 1992; Herlocker et al., 2017): We identify the nearest neighbor in response space using cosine similarity over the 52 training questions and use their responses to predict the target individual's responses on the 8 test questions.

4.4 Text Encoder Fine-tuning

The text encoder is fine-tuned with LoRA (Hu et al., 2021) guided by the contrastive learning objective proposed by Hadsell et al. (2006). To construct training pairs, we use the rephrased and negated question datasets Q_R^{Train} and Q_N, forming a dataset $D = Q_O \cup Q_R^{Train} \cup Q_N$. A question pair $(d_i, d_j) \in D \times D$ with $i \neq j$ is labeled positive if one question is a rephrased version of the other, and negative otherwise. Positive pairs are assigned label $Y = 0$ and negative pairs $Y = 1$.

Model	Directly Seen Q Unobserved A	Indirectly Seen Q Observed A	Indirectly Seen Q Unobserved A	Unseen Q Unobserved A
Random	51.72±0.46	52.38±0.07	51.50±0.44	51.92±0.85
Majority class	52.53±0.09	55.96±0.13	52.48±0.11	55.76±0.13
ICL GPT-4 Turbo	67.90±0.78	**100.00±0.00**	67.74±1.70	65.47±1.04
ICL GPT-5.5	75.52 ± 2.90	**100.00±0.00**	73.78 ± 3.28	65.12 ± 0.95
ICL Qwen3-8B	57.49 ± 2.20	99.94 ± 0.11	55.39 ± 0.48	51.06 ± 6.24
Belief-Embedding	69.08±1.07	66.17±0.50	63.78±0.09	59.45 ± 0.19
Coll. Filtering	-	-	-	**77.08**
PoliSpace	81.30±0.34	89.81±1.99	**79.76±0.67**	59.8±3.43
Baseline	**83.30±0.05**	-	-	-
Retrieval	-	98.15±0.89	-	54.36±2.86

Table 1: **Performance comparison of PoliSpace and baselines on opinion prediction** across four evaluation scenarios. Results are mean F1 (±SD) over three runs; **bold** indicates best performance, underline indicates second best. Baselines: **Random** (uniform binary); **Majority Class** (individual's observed majority answer from training); **In-Context Learning (ICL)** (GPT-4 Turbo (OpenAI et al., 2024), GPT-5.5 (OpenAI, 2026), and Qwen3-8B (Yang et al., 2025) prompted with individual training pairs as context, evaluated on one randomly selected individual from each of the 11 political party categories); **Belief-Embedding** (Lee et al., 2025); **Collaborative Filtering** (Goldberg et al., 1992; Herlocker et al., 2017) (predicting answers using the most similar individual's response, determined via cosine similarity over shared training questions). PoliSpace variants: **Baseline** learns question embeddings from scratch (upper bound for Scenario 1, cannot generalize to new questions); **Retrieval** predicts answers by retrieving the response of the most semantically similar training question via cosine similarity.

The contrastive loss is defined as:

$$\ell = \frac{1}{2}(1-Y)\, D_E(d_i, d_j)^2 + Y \cdot \frac{1}{2} \max\big(0,\, m - D_E(d_i, d_j)\big)^2 \quad (5)$$

where we use a margin of $m = 0.7$. The embedding distance metric $D_E(d_i, d_j)$ is computed as the cosine distance between the question embeddings generated by the text encoder E, defined as:

$$D_E(d_i, d_j) = 1 - \frac{E(d_i) \cdot E(d_j)}{\|E(d_i)\| \|E(d_j)\|} \quad (6)$$

Our design choice is motivated by the results presented in Table 4 and Table 3 of Appendix B.

4.5 Training Details

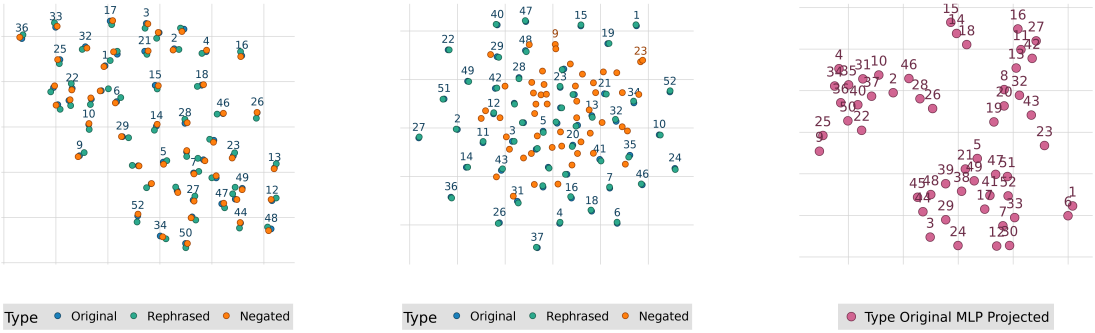
We embed individuals and questions into a shared z=256-dimensional space. Questions are initially encoded using a contrastively fine-tuned Flan-T5-Base model (Chung et al., 2024) to obtain 768-dimensional representations, which are subsequently projected into the joint 256-dimensional embedding space via a Multi-Layer Perceptron (MLP). The MLP contains hidden layers of sizes 512 and 256, with ReLU activation functions. The individual embeddings and individual-specific affine transformations used to personalize question embeddings are implemented as learnable `nn.Embedding` parameters.

The model is optimized using the Adam optimizer with a learning rate of 10^{-3} and a batch size of 32. Training runs for up to 50 epochs with early stopping (patience 10), with training taking approximately 20 minutes on an NVIDIA RTX A6000. We additionally employ a `ReduceLROnPlateau` learning rate scheduler (factor 0.1, patience 5, minimum 10^{-6}) driven by the validation loss. The training objective is the Cosine Embedding Loss with margin m=−1:

$$\ell(\mathbf{e}^{\mathbf{i}}_{\mathbf{q_n}}, \mathbf{e_i}, a_i^n) = \begin{cases} 1 - c_{ni} & \text{if } a_i^n = 1 \\ \max(0,\, c_{ni} - m) & \text{if } a_i^n = 0 \end{cases} \quad (7)$$

where $c_{ni} = \cos(\mathbf{e}_i, \mathbf{e}^i_{q_n})$ denotes the cosine similarity between the individual embedding and the personalized question embedding. EFor the parameter-efficient variant incorporating the low-rank approximation of the weight matrix W_i, we apply discriminative (per-parameter-group) learning rates within the Adam optimizer: 10^{-3} for the MLP projector and bias vectors, 5×10^{-3} for the individual embeddings, and 5×10^{-4} for the low-rank factors U_i and V_i.

All final evaluation metrics are reported as the mean F1 score ± standard deviation computed across three random seeds.



(a) Pre-trained `Flan-T5-Base` (b) Fine-tuned `Flan-T5-Base` (c) MLP projections

Figure 2: **t-SNE of question embeddings (training set).** (a) Pre-trained vs. (b) contrastively fine-tuned `Flan-T5-Base`. We show original, rephrased, and negated variants. (c) Final embeddings obtained by projecting contrastively fine-tuned representations through the learned MLP projector. Fine-tuning clusters semantically equivalent variants while separating negated variants. The MLP-projected space shows clearer structure, with questions eliciting similar response patterns forming tighter local clusters, indicating that the learned representation captures latent political opinion structure beyond surface-level semantics.

5 Results

5.1 Opinion Prediction

POLISPACE demonstrates strong predictive capacity across the four evaluation scenarios, outperforming the Belief-Embedding baseline of Lee et al. (2025) on seen and rephrased questions (Table 1).

The advantage of POLISPACE is most evident in scenario 2 (Indirectly seen questions - Observed answers), where it achieves 89.81% F1 compared to 66.17% for Belief-Embedding (Lee et al., 2025). While in-context learning with GPT-4 Turbo (OpenAI et al., 2024) and GPT-5.5 (OpenAI, 2026) achieves perfect performance on Scenario 2, with Qwen3-8B (Yang et al., 2025) achieving a similarly high F1 score of 99.94%, this setting is particularly favorable to in-context learning given the models' strong semantic reasoning capabilities and access to the original question, and consequently the individual's response to the underlying issue in the prompt. However, POLISPACE achieves competitive performance that can be further enhanced through the retrieval mechanisms presented in Section 5.2. Moreover, POLISPACE outperforms all in-context learning baselines on Scenarios 1 and 3, where the model must generalize to previously unseen individual-question pairs. Compared to the strongest in-context learning baseline, GPT-5.5 (OpenAI, 2026), POLISPACE achieves F1 scores of 81.30% vs. 75.52% and 79.76% vs. 73.78%, respectively, while also exhibiting slightly lower variance.

For the most challenging setting, scenario 4 (Unseen questions - Unobserved answers), POLISPACE achieves 59.8% F1, comparable to Belief-Embedding (59.45%). Both methods remain close to the majority-class baseline, reflecting the fundamental difficulty of zero-shot opinion prediction on entirely new topics, further amplified by the design of the smartvote questionnaire, which covers diverse policy topics with limited thematic overlap between questions. Collaborative filtering achieves the highest F1 (77.08%); however, this baseline relies on existing answer patterns from other individuals and cannot generalize to questions for which no prior responses exist across the population. In-context learning with GPT-4 Turbo (OpenAI et al., 2024) performs second best (65.47% F1). While POLISPACE does not outperform GPT-4 Turbo in this setting, the two methods offer different approaches to opinion prediction. POLISPACE provides an interpretable framework for opinion prediction and learns reusable representations of individuals and questions, eliminating the need to re-specify personal information in each prompt. We discuss these differences further in Section 6.

5.2 Retrieval and Thresholding

This section evaluates retrieval and confidence-aware thresholding as mechanisms for improving prediction reliability. The scenario assumes access to an individual's answers for all 52 training questions, and a new rephrased question is presented

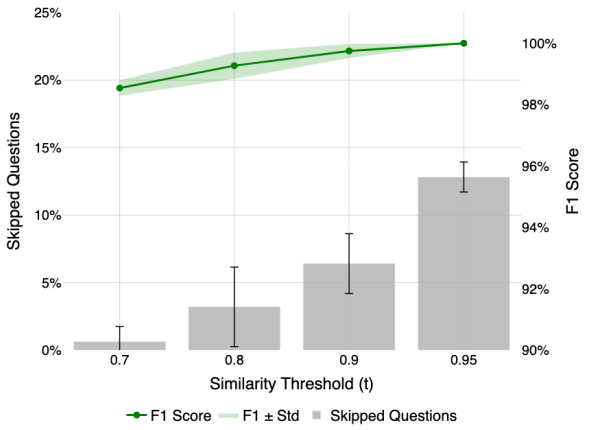


Figure 3: **Thresholded retrieval: F1 vs. coverage trade-off.** The model uses an individual's previously answered questions as a reference set. For a given question, the most semantically similar reference question is identified via cosine similarity. A prediction is retrieved only when this similarity exceeds a threshold t; otherwise the model abstains. Higher thresholds yield near-perfect F1 at the cost of reduced coverage, reaching 100% F1 when $\sim$20% of predictions are skipped.

at inference. To predict the stance, we compute the cosine similarity between the new question embedding and each of the 52 previously answered question embeddings. The most similar question is identified, and its corresponding answer is retrieved as the prediction. When a similarity threshold t is additionally applied, the model abstains from predicting when the similarity falls below t, avoiding unreliable predictions and responding only when semantically similar context exists in the known answer set.

As shown in Table 1, answer retrieval alone raises the F1 score on Scenario 2 to 98.15%, approaching the in-context learning baselines while remaining interpretable. Thresholding further improves performance: as shown in Figure 3, the model achieves 99.75% F1 while abstaining on only 6.41% of questions, and perfect 100% F1 is achievable when the model answers only high-confidence queries (approximately 20% abstention rate). This offers a principled way to balance coverage and reliability depending on application requirements.

5.3 Ablation 1: Individual-Specific Question Interpretations

We assess the importance of modeling individual interpretations of political questions by training a variant of our architecture that omits the individual-specific affine transformations. In this configuration, the model uses the same question representation $\mathbf{e_{q_n}}$ across all individuals, which is derived from the contrastively fine-tuned text encoder and the subsequent MLP projection. As shown in the first column of Table 2, incorporating personalized question interpretations yields consistent performance gains across all evaluated scenarios. This improvement is most noticeable in scenario 2 (Indirectly Seen Q - Observed A), where the F1 score increases substantially from 73.89% to 89.81% (+15.92 pp).

5.4 Ablation 2: Text Encoder Fine-tuning

We investigate the impact of fine-tuning the text encoder to yield embeddings that better capture semantic relationships. Prior to fine-tuning, the encoder fails to distinguish between a question and its negated variant (Figure 2). Post-fine-tuning, the t-SNE plots show a greater overlap of semantically aligned questions (the original version and its rephrased counterpart) and clear separation of their negated counterparts.

We train a model using a frozen, pretrained Flan-T5-Base encoder (Chung et al., 2024). The results presented in the text encoder fine-tuning ablation in Table 2 demonstrate that the model utilizing a fine-tuned text encoder outperforms the one using the frozen pre-trained one in all scenarios. These results underscore the importance of aligning encoder representations with task-specific semantic distinctions.

5.5 Ablation 3: Low-rank approximation

For an embedding dimension of $z = 256$ and a rank of $r = 4$, parameterizing the weight matrix with a low-rank factorization drastically reduces the per-individual parameter count from 65,792 parameters down to 2,304. While this constraint maintains relatively stable performance on directly and indirectly seen questions (Scenarios 1–3), it causes a noticeable drop on entirely unseen questions (Scenario 4: -10.92 pp). This suggests that full-rank parameterization improves the ability of the transformation layers to extrapolate latent political structures to entirely unobserved issues.

5.6 Structure of the Learned Embedding Space

Figure 4 shows a PCA visualization of the learned individual embeddings for all M=1,029 individuals, colored by political party affiliation. The em-

Model Configuration	Directly Seen Q Unobserved A	Indirectly Seen Q Observed A	Indirectly Seen Q Unobserved A	Unseen Q Unobserved A
POLISPACE (Full Model)	**81.30±0.34**	**89.81±1.99**	**79.76±0.67**	**59.80±3.43**
I. Question Interpretation				
POLISPACE$_{\text{w/o Interpretation}}$	74.97±0.08	73.89±0.67	73.03±0.60	57.95±0.01
II. Text Encoder Fine-tuning				
POLISPACE$_{\text{w/o FT}}$	80.68±0.16	84.36±4.12	76.83±1.63	53.61±4.83
III. Low-rank approx.				
POLISPACE$_{\text{low-rank}}$	80.72±0.26	85.78±2.32	76.98±0.37	48.88±2.64

Table 2: **Comprehensive Ablation Studies.** Performance is reported as mean F1 score (±SD) over three runs across the four evaluation scenarios. We evaluate the impact of individual question interpretation modeling, text encoder fine-tuning and low-rank approximation against our full POLISPACE framework. Modeling personal interpretations consistently improves performance, with the largest gain on Scenario 2 (+15.9 pp). Fine-tuning the encoder on rephrased and negated questions for negation sensitivity and semantic diversity improves performance across all scenarios, most notably on unseen questions (Scenario 4: +6.2 pp). Low-rank approximation ($r = 4$) drastically reduces the individual-specific parameter footprint by over 96%, maintaining relatively stable performance across Scenarios 1–3 but suffering a major drop on entirely unseen questions (Scenario 4: -10.92 pp).

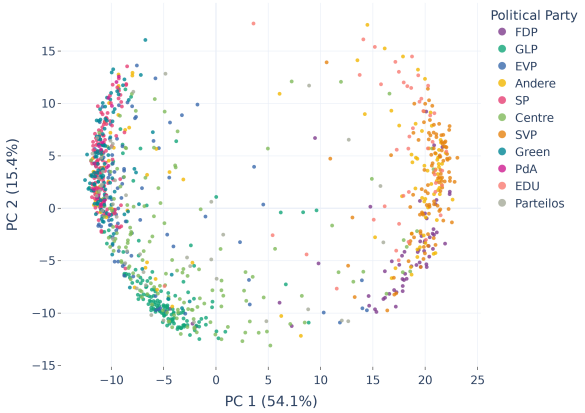


Figure 4: **PCA visualization of learned individual latent representations**. Individual embeddings (z=256) for M=1,029 individuals, colored by political party affiliation. Individuals cluster according to political party affiliation, with ideologically similar parties occupying nearby regions of the embedding space.

beddings exhibit clear clustering by party, with a separation between the SVP (right-wing) and the SP and Green parties (left-wing) along the primary axis. This structure emerges without any explicit supervision on party labels, indicating that the model has successfully captured the underlying ideological landscape of the Swiss political system directly from voting patterns.

The MLP projector produces question embeddings that exhibit clearer clustering than the fine-tuned Flan-T5-Base encoder outputs (see Figure 2). Further analysis reveals that the model appears to group questions according to the correlation structure of their answer distributions: questions that tend to be answered similarly across individuals are mapped to nearby regions of the embedding space. For instance, Figure 5 shows that question 16 is highly correlated with questions 11, 13, and 42, for which the projected embeddings are closer together than the corresponding fine-tuned encoder embeddings. This suggests that the joint training objective implicitly encourages the question representations to reflect the underlying political opinion structure of the data beyond what is captured by surface-level question semantics.

6 Discussion

POLISPACE achieves strong performance on Scenarios 1–3, demonstrating that a shared embedding space with individual-specific affine transformations is a competitive approach for opinion prediction when some overlap exists between training and target questions (through rephrasing or direct reuse). The emergent party-level clustering in the individual embedding space (Figure 4) and the correlation-reflecting structure of the question embeddings (Figure 2 and Figure 5) provide qualitative evidence that the model captures meaningful political structure beyond surface-level pattern matching.

While in-context learning achieves perfect performance on Scenario 2 and GPT models outperform POLISPACE in Scenario 4, the two approaches offer different trade-offs. Most importantly, POLISPACE provides an explicit prediction mechanism: predictions follow directly from the

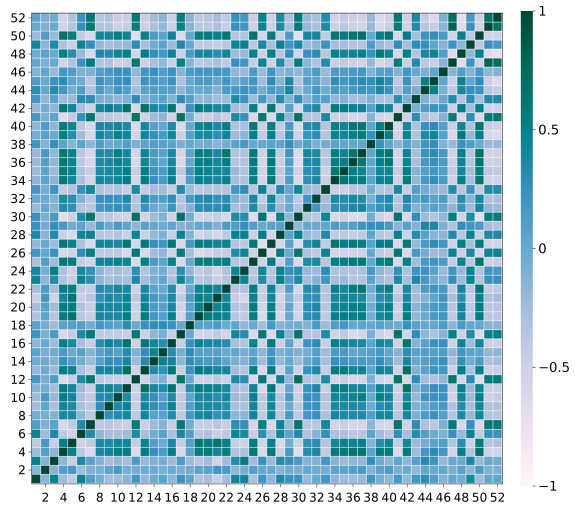


Figure 5: **Pearson correlation matrix of respondent answers.** The plot displays the correlation structure across all 52 political questions based on the responses of the 1,029 individuals. Darker green blocks indicate strong positive correlations between question pairs, while lighter shades denote weaker relationships.

geometric alignment between individual and question representations, whereas the mechanism underlying an LLM's prediction remains implicit. Its learned representations are also persistent and reusable across predictions and enable additional tasks such as answer retrieval and confidence-based abstention. In contrast, in-context learning can readily condition on new individuals but requires their response histories to be repeatedly provided through model- and context-dependent prompts, while POLISPACE requires learning individual-specific parameters.

The primary limitation of POLISPACE is its performance on Scenario 4 (Unseen Questions – Unobserved Answers, 59.8% F1), in which our model performs comparable to the Belief-Embedding baseline (59.45%). This reflects the difficulty in zero-shot opinion prediction. Predicting an individual's opinion on an entirely new political topic from their answers to a disjoint set of questions requires cross-topic generalization that the available training data do not adequately support. The smartvote questionnaire is designed for broad thematic coverage with minimal overlap between questions, which limits the transfer of learned associations. The result that both POLISPACE and Belief-Embedding fail to substantially exceed the majority-class baseline in this scenario suggests that the signal available in 52 questions is insufficient for zero-shot generalization to arbitrary new topics. Notably, collaborative filtering achieves 77.08% F1 on Scenario 4, but this is not a directly comparable result as it relies on existing answer patterns from other individuals and cannot generalize to questions for which no such prior responses exist. Addressing the zero-shot limitation likely requires either richer question coverage with greater thematic overlap or incorporating external political knowledge beyond the questionnaire.

7 Conclusion

In this paper, we propose an interpretable framework specifically designed for opinion modeling and prediction, in which individual and question representations are jointly optimized so that alignment directly reflects agreement.

Our evaluation on real-world Swiss election data from smartvote demonstrates that POLISPACE outperforms prior representation-learning-based approaches on rephrased questions (89.81% vs. 66.17% F1) and approaches the in-context learning baselines with retrieval (98.15%). Ablation studies suggest that negation sensitivity and modeling personal interpretations help improve opinion prediction.

Performance on entirely unseen questions remains a central open challenge, shared with existing approaches, and likely requires richer training signals or external political knowledge to address. While the task remains challenging, POLISPACE improves the consistency of predictions across varied question formulations.

8 Ethical considerations

While POLISPACE aims to improve the accuracy and interpretability of political opinion modeling, automated modeling of political preferences raises significant ethical concerns. The framework could be misused for democratic manipulation, such as gerrymandering or targeted influence campaigns, and regulatory safeguards are necessary to prevent such misuse. Furthermore, deploying opinion modeling in politically sensitive settings raises questions of bias, transparency, and accountability. In scenarios where model predictions influence democratic outcomes, it remains unclear who bears responsibility when predictions are incorrect or misaligned with individual intent. We release this work with the expectation that it is used to support, rather than undermine, democratic processes.

9 Limitations

- **Single-country evaluation.** Our experiments are limited to Swiss national election data from smartvote. The framework's generalizability to other political systems, cultures, and languages has not been evaluated.
- **Binary opinion modeling.** We binarize responses into agree/disagree, discarding the intensity and nuance of the original Likert-scale answers.
- **Fixed individual population.** POLISPACE requires retraining to incorporate new individuals, as each individual is a learned embedding. The framework does not support zero-shot generalization to unseen individuals.
- **Limited unseen-question performance.** Predicting opinions on entirely new questions (Scenario 4) remains challenging, with performance near baseline levels, reflecting the inherent difficulty of zero-shot belief prediction.
- **Questionnaire design dependency.** The smartvote questionnaire covers diverse policy topics with limited thematic overlap, which may restrict the transfer of learned associations between questions.
- **Static opinions.** The model captures a snapshot of individual positions at a single point in time and does not account for opinion change or temporal dynamics.
- **Canton-level bias.** Our analysis is restricted to individuals from the canton of Zürich, which is generally regarded as more left-leaning. This may result in an overrepresentation of left-leaning perspectives. A more comprehensive evaluation would apply the framework to all candidates across Switzerland.
- **Candidate vs. voter responses.** We restricted our analysis to candidates rather than voters. Candidates typically provide more coherent and strategically aligned responses, which may inflate model performance. Future work should evaluate the framework on voter data to better assess real-world applicability.

References

Christopher H. Achen and Larry M. Bartels. 2017. *Democracy for Realists: Why Elections Do Not Produce Responsive Government*, revised edition. Princeton University Press, Princeton, NJ.

Pedro Caiua Campelo Albuquerque and Daniel Oliveira Cajueiro. 2025. Forecasting political voting: A high dimensional machine learning approach. *Machine Learning with Applications*, 22:100739.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. 1992. Using Collaborative Filtering to Weave an Information Tapestry. *Communications of the ACM*, 35(12):61–70.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Jon Herlocker, Al Borchers, and John Riedl. 2017. An algorithmic framework for performing collaborative filtering. *ACM SIGIR Forum*, 51:227–234.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Byunghwee Lee, Rachith Aiyappa, Yong-Yeol Ahn, Haewoon Kwak, and Jisun An. 2025. A semantic embedding space based on large language models for modelling human beliefs. *Nature Human Behaviour*, 9(9):1928–1940.

Junyi Li, Charith Peris, Ninareh Mehrabi, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and

Rahul Gupta. 2024. The steerability of large language models toward data-driven personas. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7290–7305, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI. 2026. GPT-5.5 System Card.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Politools. 2025. Smartvote. Accessed March 2026.

Nathan E. Sanders, Alex Ulinich, and Bruce Schneier. 2023. Demonstrations of the potential of ai-based political issue polling. *Preprint*, arXiv:2307.04781.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Joshua C. Yang, Damian Dailisan, Marcin Korecki, Carina I. Hausladen, and Dirk Helbing. 2024. Llm voting: Human choices and ai collective decision-making. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1696–1708.

Xinnong Zhang, Jiayu Lin, Libo Sun, Weihong Qi, Yihang Yang, Yue Chen, Hanjia Lyu, Xinyi Mou, Siming Chen, Jiebo Luo, Xuanjing Huang, Shiping Tang, and Zhongyu Wei. 2024. Electionsim: Massive population election simulation powered by large language model driven agents. *Preprint*, arXiv:2410.20746.

A Ethical Statement

Politools anonymized and collected the dataset in accordance with all applicable data protection laws, including the revised Swiss Federal Act on Data Protection (nFADP) and the Telecommunications Act (TCA). We adhere to the platform's terms of service, ensuring that no outcomes reported in this work can be traced back to individual users. In line with these research usage terms, the dataset remains private. We apply standard protocols for handling sensitive data responsibly. Researchers may be able to request access directly from Politools.

B Fine-tuning Methods

`Flan-T5-Base` (Chung et al., 2024) was fine-tuned using LoRA (Hu et al., 2021) with a rank of 8. The training was conducted over 100 epochs with early stopping (patience = 10). The batch size was set to 32, and the learning rate scheduler had a patience of 5 and a reduction factor of 0.1. Optimization was performed using the Adam optimizer. The initial learning rate was selected using PyTorch Lightning's learning rate finder.

Two objective functions were explored: triplet loss and contrastive loss (Hadsell et al., 2006), both employing cosine distance as the similarity metric. Samples were constructed from the datasets Q_O, Q_R^{Train} and Q_N. For the triplet loss, triplets were constructed in the following way: for each question $q_n \in Q_O$, the positive example was its rephrased version $\sim q_n \in Q_R^{Train}$, and the negative examples included:

- its negated version $q_n \in Q_N$.
- all other original questions $q_j \in Q_O$, where $j \neq n$.
- all other negated questions $q_j \in Q_N$, where $j \neq n$.
- all other rephrased versions $q_j \in Q_R^{Train}$, where $j \neq n$.

For both loss functions, margin values $m = \{0.5, 0.7, 0.9, 1.1, 1.3, 1.5\}$ were tested. Evaluation was based on cosine similarity between:

- The original question $q_n \in Q_O$ and its rephrased version $\sim q_n \in Q_R^{Train}$.
- The original question $q_n \in Q_O$ and its negated version $\neg q_n \in Q_N$.

Results are shown in Table 3. Initially, increasing the margin value m improves the model's ability to distinguish between a question and its negated version. However, beyond a certain point, further increases in m lead to a collapse in the representation space, where the encoder fails to separate the two. This collapse is likely due to the margin becoming too large relative to the number of samples, making it difficult for the model to learn embeddings that effectively separate all examples in the representation space.

We trained POLISPACE models using the encoders that best distinguished original questions $q_n \in Q_O$ from their negated counterparts $\neg q_n \in Q_N$. The encoder fine-tuned using contrastive loss with a margin of $m = 0.7$ resulted in the best-performing POLISPACE model across all evaluation scenarios (see Table 4).

C Data Generation Prompts

We report the prompts used to generate the rephrased and negated question datasets with GPT-4 (OpenAI et al., 2024). `X` is replaced by the desired number of variants.

Rephrasing prompt

> I will provide you with a sentence that contains a political statement or question. Your task is to create X rephrased versions of the political statement or question that the sentence contains, while preserving any introductory or closing text. Some of the inputs will contain some words before the statement/question, such as 'What is your position the following statement', these words should remain unchanged and you should only rephrase the actual political statement / question. When rephrasing the statement/question, follow these guidelines:
>
> (1) If there are any words prior to the statement/question, you should append them to the rephrased statement/question; (2) You are only supposed to rephrase the actual statement/question, everything else should be left unchanged; (3) Similarly, some statements/questions contain words after the statement/question, such as '[BePart question]'. This should also be included in the rephrased version and kept unchanged; (4) Vary sentence structure: Alter the order of sentence elements where natural and grammatically correct. For example, consider switching subject and object or reordering clauses; (5) Different vocabulary: Use different vocabulary for the rephrased questions, such as 'increase' instead of 'raise', or 'remove' instead of 'eliminate'; (6) Ensure natural phrasing: The rephrased versions should read smoothly and sound like authentic, fluent language. It should not contain overly complicated words; (7) Preserve response equivalence: It is important that the answer form remains unchanged. If a person answered agree to

FT Method	Margin m	Training set		Test set	
		$(\mathbf{Q}, \mathbf{Q_R})$	$(\mathbf{Q}, \mathbf{Q_N})$	$(\mathbf{Q}, \mathbf{Q_R})$	$(\mathbf{Q}, \mathbf{Q_N})$
Triplet	0.5	0.887±0.036	0.205±0.106	0.802±0.140	0.274±0.131
	0.7	0.931±0.0214	**0.111±0.171**	0.770±0.251	**0.197±0.261**
	0.9	0.981±0.006	0.235±0.378	0.868±0.099	0.274±0.415
	1.1	**0.995±0.002**	0.995±0.014	0.923±0.072	0.996±0.002
	1.3	0.994±0.002	0.996±0.012	0.931±0.047	0.997±0.002
	1.5	0.994±0.002	0.996±0.011	**0.932±0.041**	0.997±0.002
Contrastive	0.5	**0.957±0.014**	0.350±0.099	**0.919±0.028**	0.425±0.171
	0.7	0.950±0.029	0.183±0.079	0.858±0.096	0.352±0.162
	0.9	0.777 ± 0.100	**0.110±0.090**	0.566±0.105	0.362±0.133
	1.1	0.548±0.195	0.216±0.136	0.513±0.188	**0.239±0.110**
	1.3	0.316±0.135	0.461±0.157	0.227±0.097	0.487±0.149
	1.5	0.500±0.171	0.821±0.104	0.487±0.184	0.835±0.071

Table 3: **Average cosine similarity between original questions and rephrased/negated variants after fine-tuning with triplet or contrastive loss, across margin values m.** Results are reported separately for the training and test sets. Higher similarity with rephrased questions and lower similarity with negated questions indicate better semantic discrimination. Bold indicates best-performing model.

Encoder	Directly Seen Q Unobserved A	Indirectly Seen Q Observed A	Indirectly Seen Q Unobserved A	Unseen Q Unobserved A
Contrastive $m = 0.7$	**81.63**	**89.17**	**80.44**	**59.11**
Contrastive $m = 0.9$	81.13	82.10	75.82	56.73
Triplet $m = 0.7$	81.58	88.27	78.26	52.82

Table 4: **POLISPACE performance across text encoder fine-tuning configurations.** `Flan-T5-Base` (Chung et al., 2024) is fine-tuned with either contrastive or triplet loss across different margin values m and used as the text encoder for POLISPACE. F1 scores are reported across four evaluation scenarios. Seed 0 was used for all runs. Bold indicates best-performing configuration.

the original question, they should also agree to the rephrased question, and vice versa; (8) If asked to produce numerous rephrased version, make sure the different variant are different from each other. They should have different sentence structures.

Create exactly `X` distinct rephrasings of the political statement/question. Output them as a numbered list (1., 2., etc.).

Negation prompt

I will provide you with a sentence that contains a political statement or question. Your task is to create `X` negated / opposite versions of the political statement or question that the sentence contains, while preserving any introductory or closing text. Some of the inputs will contain some words before the statement/question, such as 'What is your position on the following statement', these words should remain unchanged and you should only negate the actual political statement / question. Some questions contain contextual information before the question, this contextual information should remain unchanged and only the sentence containing the question should be negated. When negating the statement/question, follow these guidelines:

(1) If there are any words prior to the statement/question, you should append them to the negated statement/question; (2) You are only supposed to negate the actual statement/question, everything else should be left unchanged; (3) Similarly, some statements/questions contain words after the statement/question, such as '[BePart question]'. This should also be included in the negated version and kept unchanged; (4) Try keeping all other words in the statement the same to the greatest extent possible, that is try only switching words such as agree/disagree, increase/decrease, that would make the statement/question express the opposite belief; (5) The sentence structure should remain the same. Do not shift around the order of the other words unless necessary. Perform the negation by flipping key words that signal stance or direction, such as agree/disagree, increase/decrease; (6) It is crucial that the negated statement/question states the opposite of the original statement/question. You need to ensure that the semantic meaning is the opposite; (7) Do not preserve response equivalence: It is important that the answer form is changed. If a person answered agree to the original question, they should answer disagree to the negated question,

and vice versa; (8) If asked to produce numerous negated versions, make sure the different variants are different from each other. They should have different words indicating agree/disagree or increase/decrease.

Create exactly X distinct negations of the political statement/question. Output them as a numbered list (1., 2., etc.).

D Cross-Encoder Baseline

We additionally evaluate an open-source BERT-style cross-encoder baseline using ModernBERT-base (Warner et al., 2024). For each prediction, the model receives the individual's observed question–answer pairs followed by the target question and predicts *Agree* or *Disagree* using a two-class classification head. We perform full-parameter fine-tuning using leave-one-out examples, uniformly sampling five target questions per individual, resulting in 5,145 training examples. The model is trained with cross-entropy loss and a learning rate of 2×10^{-5}, with the checkpoint selected based on validation macro-F1. Table 5 reports performance across the four evaluation scenarios.

Model	S1	S2	S3	S4
ModernBERT-base	78.70	71.33	65.73	60.86

Table 5: F1 scores of the ModernBERT-base cross-encoder across the four opinion prediction scenarios.

E Additional Results

This appendix provides supplementary material, the four evaluation scenarios (Figure 6), the Belief-Embedding baseline configurations (Table 6, Figure 7), k-fold cross-validation (Table 7), and t-SNE embedding visualizations (Figure 8–Figure 9).

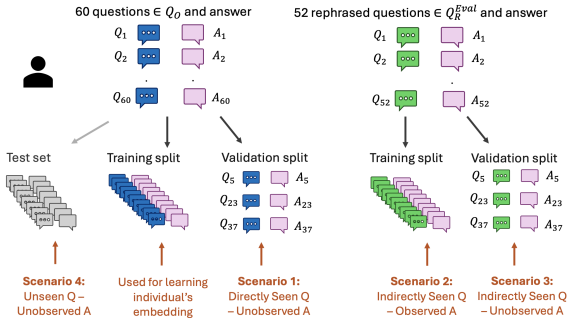


Figure 6: **Four evaluation scenarios of increasing difficulty.** From 60 smartvote questions, 8 are held out entirely for testing. The remaining 52 are used to generate (individual, question, answer) triplets, split into per-individual training and validation sets (the split differs across individuals). Rephrased versions of the questions are constructed with the same split applied. **1) Directly Seen Q, Unobserved A**: validation split of original questions; the model has seen these questions during training (for other individuals) but not the current individual's answer. **2) Indirectly Seen Q, Observed A**: training split of rephrased questions; the individual's answer to the original version is known. **3) Indirectly Seen Q, Unobserved A**: validation split of rephrased questions; the original question was seen but the individual's answer was not. **4) Unseen Q, Unobserved A**: the 8 held-out questions, never encountered during training.

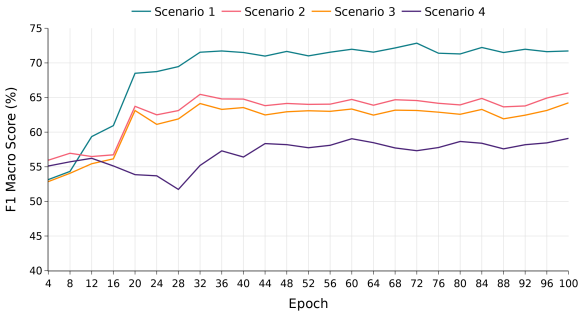


Figure 7: **Belief-Embedding training convergence.** F1 scores across 100 epochs (Euclidean distance, m=20) for the four evaluation scenarios. Performance stabilizes after approximately 40 epochs.

Metric	Margin	Directly Seen Q Unobserved A	Indirectly Seen Q Observed A	Indirectly Seen Q Unobserved A	Unseen Q Unobserved A
Euclidean	w/o FT	50.57	54.98	52.37	54.93
	5	67.74	65.52	61.23	49.66
	7	68.31	63.22	61.72	50.10
	15	72.54	64.69	63.23	57.95
	20	71.54	65.46	64.14	55.19
	30	67.24	68.19	66.23	55.54
	40	69.42	68.18	65.91	59.43
Cosine Sim.	0.5	71.45	65.11	61.44	58.55
	0.7	70.62	65.76	63.21	58.86

Table 6: **Belief-Embedding baseline across training configurations.** S-BERT models were fine-tuned and evaluated following Lee et al. (2025). smartvote questions are converted into belief statements ("I agree/disagree with: S"); individual embeddings are the mean of their training-set belief vectors. S-BERT is fine-tuned for 50 epochs with triplet loss, varying margin and distance metric (Euclidean or cosine). At inference, the individual's stance is determined by comparing distances to the two opposing belief embeddings ($\mathbf{s}_{PRO}$ and $\mathbf{s}_{CON}$). The same train–validation–test split used for POLISPACE is applied. The best-performing model for each configuration is reported using F1 scores across four scenarios. Best results use Euclidean distance with margin 15–40, though no configuration matches POLISPACE on rephrased questions.

Fold	Directly Seen Q Unobserved A	Indirectly Seen Q Observed A	Indirectly Seen Q Unobserved A	Unseen Q Unobserved A
1	81.16	92.13	79.24	63.43
2	81.07	94.24	79.93	53.86
3	80.71	89.92	80.32	59.82
4	81.39	93.23	80.35	48.55
5	80.79	89.45	80.05	52.73
6	81.56	84.67	79.91	49.29
7	82.34	92.63	80.91	54.43
Mean $\pm$ Std Dev	81.29 $\pm$ 0.55	90.90 $\pm$ 3.24	80.10 $\pm$ 0.51	54.59 $\pm$ 5.39

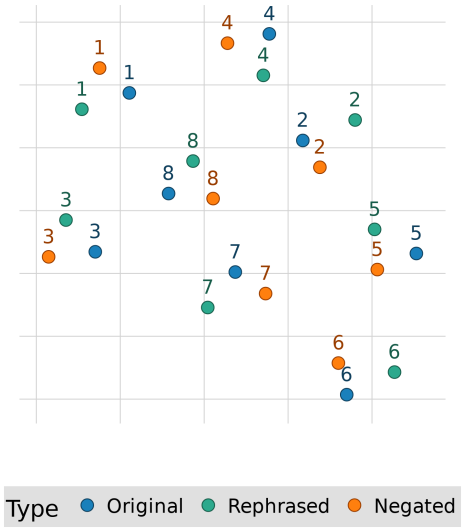
Table 7: **k-fold cross-validation results for POLISPACE.** F1 scores per fold (each holding out 8 test questions) across four scenarios. Low standard deviations (0.51—5.39 pp) confirm stable performance across different question splits.



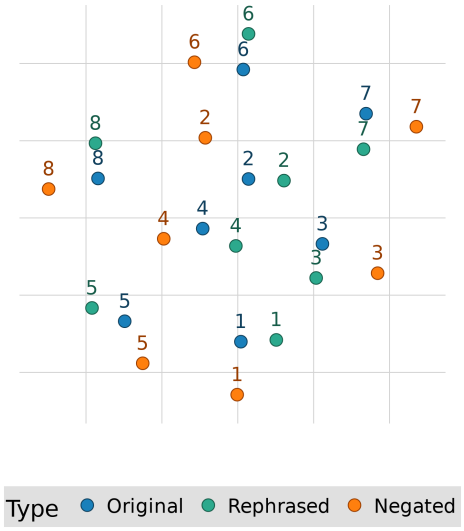
(a) Pre-trained S-BERT

(b) Fine-tuned S-BERT

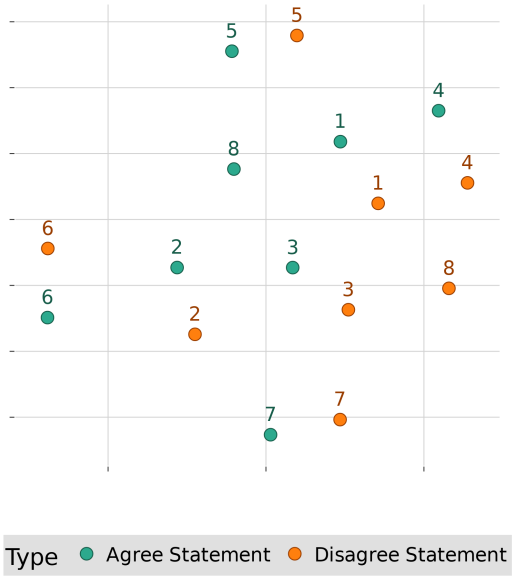
Figure 8: **t-SNE of belief-statement embeddings (training set).** (a) Pre-trained vs. (b) fine-tuned S-BERT (Lee et al., 2025). Each point pair represents "agree" and "disagree" variants. Triplet-loss fine-tuning collapses co-occurring beliefs into tighter clusters.



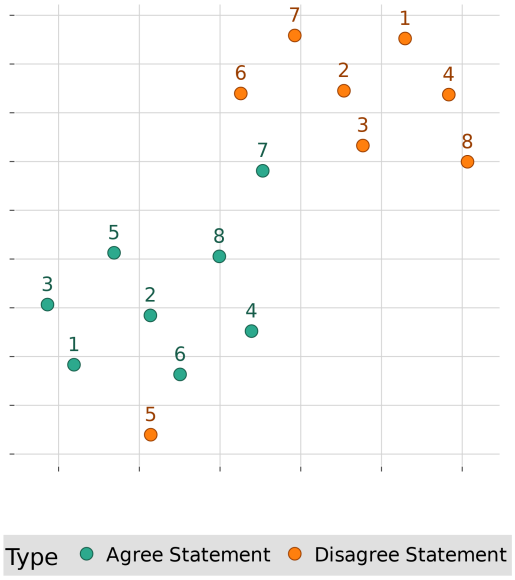
(a) Pre-trained Flan-T5-Base



(b) Fine-tuned Flan-T5-Base



(c) Pre-trained S-BERT



(d) Fine-tuned S-BERT

Figure 9: **t-SNE of test-set embeddings.** (a, b) Flan-T5-Base before and after contrastive fine-tuning on question variants; (c, d) S-BERT before and after Belief-Embedding fine-tuning (Lee et al., 2025). Fine-tuning improves clustering in both approaches, with POLISPACE showing clearer separation between question groups.