# *LMDA-Radar*: Ideological Discourse Profiling with Lexical Multi-Dimensional Analysis and AI

**Anonymous ACL submission**


## Abstract

AI-generated texts are increasingly being circulated in society, conveying the potential to influence people's worldviews by reinforcing or reshaping ideologies. In this work, we investigate the capacity of Large Language Models (LLMs) to reinforce or reverse ideological discourses. To identify LLM leaning towards certain ideology, we propose LMDA-Radar, a framework that leverages Lexical Multi-Dimensional Analysis (LMDA) to profile ideological discourse and systematically classify textual leaning. We evaluate our framework using Claude Haiku 4.5, GPT-4o-mini and GPT-5-mini. We rely on GPT-4o responses to ideological priming when tasked to rewrite 800 news articles about immigration. The generated corpus comprises three distinct conditions, namely *convergent prompting*, where the model preserves the original ideological position; *divergent prompting*, where it reverses the position; and *internal knowledge generation*, where no explicit ideological guidance is provided. Results demonstrate that while ideological steering is generally successful across all prompting conditions, the model displays no inherent systemic preference toward a specific ideology, validating the capability of LMDA-Radar for dynamic model auditing and linguistic profiling.[1]


## 1 Introduction

Large Language Models (LLMs) are increasingly embedded in public discourse, playing a growing role in journalism, social media, education, and political communication (Shalevska, 2024). This raises concerns about the large-scale manipulation of ideological narratives through automated text generation. At the same time, these phenomena also motivate an important research question:

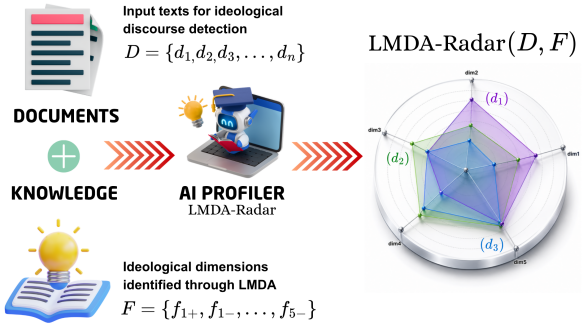



Figure 1: Overview of the LMDA-Radar framework for ideological discourse profiling. The system receives a set of documents $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ together with ideological dimensions extracted through Lexical Multi-Dimensional Analysis (LMDA). Each dimension is modeled as an axis composed of two opposing poles (e.g., $f_{1+}$ vs. $f_{1-} = dim1$) derived from correlated lexical patterns identified through factor analysis.

whether LLMs themselves can serve as computational tools for modeling, analyzing, and auditing ideological discourse at scale (Ziems et al., 2024). Previous research has shown that LLMs can absorb ideological tendencies from relatively small amounts of data and generalize these shifts across topics or tasks (Chen et al., 2024; Zhang, 2025). Such findings suggest that ideological manipulation in LLMs is not limited to isolated prompts or surface-level lexical imitation, but may reorganize broader discourse patterns. Immigration discourse is particularly relevant in this context due to its strong political polarization and multidimensional ideological structure. For this reason, we investigate how LLMs respond to systematic ideological steering under controlled discourse configurations. To go beyond single-axis political leaning scales (e.g., "left" vs. "right"), we employ Lexical Multidimensional Analysis (LMDA) to model ideological movement as shifts across multiple discourse dimensions derived from recurring lexical patterns. Rather than treating ideological orienta-

[1] Supplementary material available at: https://osf.io/r9gqe/overview?view_only=9459b2007d0549cdb8d21f8c831c3dfc

tions as fixed categories, this approach characterizes how they emerge from combinations of linguistic signals distributed across discourse.

This perspective aligns with broader theories of ideology that understand political orientations not as isolated opinions, but as structured systems of social attitudes reflected through recurring discourse practices (van Dijk, 1993). In texts, these ideologies manifest through discourse patterns, including lexical choices, framing strategies, and rhetorical preferences. Because LLMs increasingly generate and disseminate public discourse, monitoring such patterns computationally becomes important not only for identifying ideological tendencies, but also for evaluating how robust these systems are to ideological manipulation. In this sense, mapping texts to ideological orientations allows researchers to investigate whether LLM-generated discourse remains stable under steering conditions or undergoes systematic ideological reconfiguration.

In NLP, discussions about politically sensitive language are often framed in terms of bias, typically focusing on harmful statistical associations or model outputs (Kabir et al., 2025; Rettenberger et al., 2025; Tan and Lee, 2025; Lin et al., 2025; Dai et al., 2025; Prama and Islam, 2025; Nadeem et al., 2025; Bang et al., 2024; Jenny et al., 2024; Lunardi et al., 2024; Wan et al., 2023; Feng et al., 2023; Sheng et al., 2019; Zhao et al., 2019; Kurita et al., 2019; Basta et al., 2019). While this perspective is important, Blodgett et al. (2020) argue that treating these phenomena exclusively as "bias" may overlook the broader ideological structures that shape discourse itself.

By approaching the problem through the concept of ideology, we move beyond isolated linguistic artifacts, such as individual keywords, sentiment polarity, stereotypical lexical associations, or demographic word biases, and toward the systems of values, beliefs, and power relations that organize discourse patterns in society. At the same time, such artifacts remain important as observable linguistic signals through which broader ideological structures can be identified. In this sense, ideological analysis allows bias to be interpreted not only as a computational irregularity, but also as the manifestation of socially situated discursive structures reflected in language use.

We therefore investigate whether LLMs preserve, reverse, or intensify ideological stances under controlled prompting configurations. More specifically, we analyze how `GPT-4o` responds to explicit ideological priming when rewriting immigration news articles. We do not assume that political ideology can be reduced to a single left–right axis; instead, this binary distinction serves only as a starting point for the extraction of finer-grained ideological discourses through LMDA scoring.

Raw LMDA dimension scores, however, remain dependent on the lexical distributions associated with each factor, making direct comparisons across dimensions less straightforward. To address this, we introduce LMDA-Radar, a normalized multidimensional ideological profiling approach designed to preserve relative document positioning while transforming dimension scores into comparable scales. We then apply this framework to measure ideological steering under three generation conditions: convergent prompting, in which the model is expected to maintain the original ideological profile; divergent prompting, in which the model is expected to reverse it; and internal generation, in which no explicit ideological guidance is provided.

Results show that ideological prompting steers model outputs in distinct ways. Convergent prompting amplifies ideological discourse, making generated texts more distinguishable from human-authored baselines. Divergent prompting produces systematic ideological reversal, whereas internal generation attenuates ideological positions in the absence of explicit priming. We hypothesize that this framework can contribute to the development of more ideologically robust LLM systems and provide finer-grained methods for auditing ideological manipulation.

## 2 Framework Construction

### 2.1 Corpus Design and Annotation

We designed a corpus that consists of 800 news articles about immigration collected from news media outlets in the United States and Great Britain. The articles span five consecutive years (2016–2020) and were annotated by a discourse analyst according to political orientation (left-wing or right-wing) based on the known political alignment of the publication and the stances expressed. The corpus follows a balanced cell-based design. Each cell corresponds to a unique combination of country, year, and partisan orientation, resulting in 20 cells overall (2 countries × 5 years × 2 partisan orientations). Each cell contains 40 articles, resulting in a corpus of 800 documents divided into two political subsets. We present annotation validation results

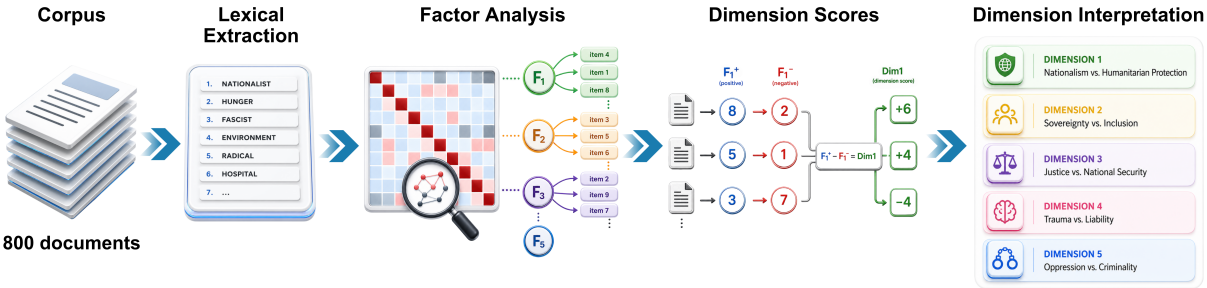



Figure 2: Lexical Multi-Dimensional Analysis (LMDA) Pipeline

in Appendix B.2. Complete annotated corpus is available in the supplementary material.

After labeling the documents according to their political orientation, we employ Lexical Multi-Dimensional Analysis (LMDA) to expand this binary categorization into a multidimensional ideological space. Although left and right may themselves be understood as ideological orientations, they manifest in texts through discourse, that is, through recurring lexical and rhetorical patterns. These rhetorical patterns refer to recurring discursive strategies used to frame social reality, such as emphasizing threat, appealing to morality or authority, assigning blame, positively or negatively representing groups, using emotional or technical language, and constructing narratives (van Dijk, 1993).

### 2.2 Lexical Multi-Dimensional Analysis

Lexical Multi-Dimensional Analysis (LMDA) is a corpus-based framework for modeling latent discourse dimensions through the statistical co-occurrence patterns of lexical items. At its core, LMDA relies on factor analysis to identify clusters of correlated lexical features that tend to occur together in similar communicative contexts. The method derives dimensions directly from the distributional structure of the corpus, allowing recurrent discourse patterns to emerge empirically from the data (Fig. 2).

Given a corpus of $n$ documents and a vocabulary $\mathcal{K}$ of size $m$, the procedure begins with a term-document matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$, where each entry $X_{i,v}$ represents the frequency of lexical item $v$ in document $i$. In this study, monograms were adopted as the primary lexical unit.

The dimensions extracted through LMDA are interpreted through factor loadings. Lexical items with positive loadings and lexical items with negative loadings on the same factor form two opposite poles of that dimension. Importantly, the terms *positive* and *negative* do not refer to semantic polarity, sentiment, affect, moral value, or ideological desirability. They refer exclusively to the mathematical direction of the correlations estimated during factor analysis. In practice, positive and negative poles correspond simply to two internally coherent clusters of lexical items that systematically co-occur in contrasting distributional patterns along the corpus. Next, we provide an overview of the LMDA pipeline adopted in this study, summarized in Appendix B.1, Alg. 1.

**Lexical Extraction** Not all lexical items carry meaningful discourse signal. Therefore, we first apply TreeTagger[2] as a POS filter to retain only content words, namely nouns, verbs, adjectives, and adverbs. Keywords are then extracted through log-likelihood ($G^2$) by contrasting the two political subsets against each other. This procedure yields a vocabulary $\mathcal{K}$. To reduce noise, a Pearson correlation matrix $\mathbf{R} \in \mathbb{R}^{m \times m}$ is computed over the columns of $\mathbf{X}$. Lexical items are then selected according to their strongest positive (top 400) and negative (bottom 400) inter-correlations, producing a reduced subset $\mathbf{X}_{\text{sel}} \subset \mathbf{X}$.

**Polychoric Correlation Matrix** Because lexical frequencies behave as ordinal variables in sparse linguistic data, a polychoric correlation matrix $\mathbf{R}_{\text{poly}}$ is computed from $\mathbf{X}_{\text{sel}}$. The polychoric estimator assumes that each observed ordinal variable is a discretized realization of an underlying continuous latent variable. This step refines the estimation of lexical relationships before factor extraction.

[2]We used the default pretrained English TreeTagger model (Schmid, 1995) without additional parameter tuning.

**Factor Analysis** Principal Axis Factoring (PAF) is applied to $\mathbf{R}_{\text{poly}}$ to extract an initial set of latent factors. This procedure produces an initial loading matrix $\mathbf{L}^{(0)}$. To improve interpretability, Promax oblique rotation is applied to $\mathbf{L}^{(0)}$, generating the final loading matrix $\mathbf{L} \in \mathbb{R}^{m \times q}$. In PAF, factors are extracted sequentially according to the amount of shared variance they explain in the data. As a result, earlier factors tend to represent broader or more dominant latent discourse patterns, whereas later factors capture more specific or residual structures. This ordering does not imply importance in a normative sense, but rather reflects the relative contribution of each factor to the overall correlation structure of the corpus. Unlike orthogonal rotations, Promax allows correlations among factors, which is particularly suitable for ideological discourse analysis, since discourse dimensions often overlap and interact. The number of factors $q$ is determined jointly through the eigenvalue criterion and scree plot inspection, that is, by identifying the point at which the explained variance curve stabilizes. In this study, $q = 5$.

**Factor Loadings** Each lexical item $v$ is assigned to its primary factor $j^*$ according to the factor with the highest absolute loading: $j^*(v) = \arg\max |L_{v,j}|$, where $j \in \{1, \ldots, q\}$. Only lexical items with absolute loadings higher than $0.3$ are retained. Those with positive loadings form the positive pole $\mathcal{V}_j^+$ of factor $j$, while lexical items with negative loadings form the negative pole $\mathcal{V}_j^-$. Again, these poles should not be interpreted semantically or affectively. They merely represent opposing clusters of correlated lexical behavior estimated through factor analysis.

**Factor Scores and Dimension Scores** Each document, then, receives *factor scores*. Factor scores represent the cumulative frequency of lexical items associated with each pole within a given document:

$$F_{i,j}^+ = \sum_{v \in \mathcal{V}_j^+} X_{i,v} \qquad F_{i,j}^- = \sum_{v \in \mathcal{V}_j^-} X_{i,v} \quad (1)$$

Those scores are used to compute *dimension scores*. For each document $i$ and factor $j$, a unit-weighted dimension score $S_{i,j}$ is computed as: $S_{i,j} = F_{i,j}^+ - F_{i,j}^-$. A positive score indicates greater lexical alignment with the positive pole of the dimension, whereas a negative score indicates greater alignment with the negative pole. Differences in vocabulary distribution and factor composition can produce uneven LMDA score ranges across dimensions, making direct cross-dimensional comparisons and multidimensional profile visualization less straightforward when relying solely on raw scores.

**Microanalysis** Although LMDA extracts dimensions statistically, the interpretation of those dimensions requires qualitative examination. For this reason, a microanalysis is conducted over the documents located at the extremities of each dimension. The microanalysis is conducted manually by a discourse analyst, who examines how lexical patterns materialize in discourse practices, rhetorical strategies, framing devices, and ideological positioning. First, the highest-loading lexical items of each pole are examined for recurrent semantic patterns. Second, texts scoring highest and lowest (top and bottom 20) on each dimension are retrieved and manually analyzed. Finally, poles are formalized as ideological dimensions. Each dimension is documented through five fields: *Factor* identifies the LMDA dimension being activated; *Polarity* specifies whether that is the positive or negative pole of that dimension; *Dimension Label & Wing* provides a concise ideological characterization of the pole; *Dimension Description* summarizes the dominant discourse patterns associated with that pole; and *Lexical Items* correspond to the keywords with the strongest loadings on the pole. These structured descriptions were later incorporated into the prompting strategy used in the AI rewriting experiments.

### 2.3 Ideological Dimension Interpretation

Having outlined the theoretical and methodological principles of LMDA, we now describe its application to the corpus. After the standard LMDA filtering procedure, factor analysis was applied to the keyword correlation matrix, resulting in five ideological dimensions composed of 400 lexical items associated with distinct discourse patterns surrounding immigration. Extreme-scoring texts revealed a clear tendency for left- and right-wing documents to cluster at opposite poles, allowing each pole to be associated with the political orientation most strongly represented among its texts. The resulting interpretation is therefore grounded simultaneously in statistical evidence and discourse analysis principles, including the critical perspective proposed in Critical Discourse Analysis (van Dijk,

1993). A summary of the dimensions is presented in Appendix B.1, whereas complete descriptions are available in the supplementary material.

## 3 Ideological Steering Experiment

The steering experiment was conducted using the original corpus as input for controlled ideological rewriting. Each text was submitted to prompting configurations designed to guide the generation toward specific ideological dimensions extracted through LMDA. These prompts incorporated LMDA-derived ideological cues, including representative lexical items and the qualitative descriptions associated with each dimension pole.

Three rewriting conditions were evaluated: convergent rewriting, divergent rewriting, and internal rewriting. Table 1 displays all experiment conditions and configurations. The prompt templates for each rewriting condition are shown in Appendix A.1.

Table 1: Experimental conditions and configurations.

| Wing | Condition | Configurations |
|---|---|---|
| Left | Human | `hlw` |
| | Internal | `ilw` |
| | Convergent | `clwf1neg`, `clwf2neg`, `clwf3pos`, `clwf4pos`, `clwf5pos` |
| | Divergent | `drwf1pos`, `drwf2pos`, `drwf3neg`, `drwf4neg`, `drwf5neg` |
| Right | Human | `hrw` |
| | Internal | `irw` |
| | Convergent | `crwf1pos`, `crwf2pos`, `crwf3neg`, `crwf4neg`, `crwf5neg` |
| | Divergent | `dlwf1neg`, `dlwf2neg`, `dlwf3pos`, `dlwf4pos`, `dlwf5pos` |

*Note.* `h` = human, `i` = internal knowledge generation, `c` = convergent prompting, `d` = divergent prompting, `lw` = left-wing, `rw` = right-wing, `f` = factor, `pos` = positive, and `neg` = negative.

In *convergent rewriting*, the model receives LMDA-derived ideological cues aligned with the original political orientation of the source text, evaluating whether aligned prompts intensify the original ideological profile. In *divergent rewriting*, the model receives cues associated with the opposite political orientation, assessing whether the ideological profile shifts toward an opposing discursive direction. Finally, in *internal rewriting*, the model rewrites the text without explicit ideological guidance, relying only on its internal representations and prior knowledge. This condition serves as one of the baselines, along with the original corpus, for comparison against the guided steering configurations.

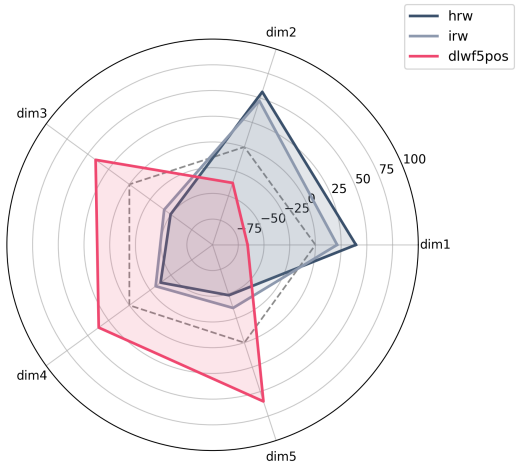



Figure 3: Radar profiles comparing mean scores of original human-written texts (`hrw`), internally generated rewritings without explicit ideological guidance (`irw`), and an LMDA-guided steering configuration (`dlwf5pos`).

## 4 LMDA-Radar

### 4.1 LMDA-Radar Scoring Framework

We propose LMDA-Radar, a normalized multidimensional ideological profiling approach designed to preserve the relative positioning of documents within each dimension while transforming the scores into comparable scales. LMDA-Radar preserves the general scoring logic adopted in LMDA, in which factor scores associated with the positive and negative poles of a dimension are first computed and then combined into a single dimension score. In LMDA, these scores are obtained from the cumulative frequency of lexical items associated with each pole (Eq. 1), where $F^{+}_{i,j}$ and $F^{-}_{i,j}$ correspond to the positive and negative pole factor scores for document $d_i$ in dimension $j$. The final dimension score is then computed as the difference between those scores.

LMDA-Radar follows a similar rationale, but replaces frequency-based accumulation with LLM-based ideological scoring. Instead of directly counting lexical occurrences, the model receives ideological knowledge extracted through LMDA, including both the representative lexical items associated with each pole and the qualitative descriptions obtained during the microanalysis stage. The prompt was therefore designed to combine lexical and interpretive information so that the model could evaluate the degree to which a text aligns with each ideological pole (prompt template is featured in Appendix A.1).

For each dimension, the model independently assigns scores to the positive and negative poles using a bounded range from 0 to 100: $R^+_{i,j}, R^-_{i,j} \in [0, 100]$. The final LMDA-Radar dimension score is then obtained through the same difference-based logic used in LMDA: $\hat{S}_{i,j} = R^+_{i,j} - R^-_{i,j}$. This procedure preserves the relational structure between opposing ideological poles while transforming the dimensions into comparable bounded scales. As a result, documents can be represented as multidimensional ideological profiles: $d_i \rightarrow (\hat{S}_{i,1}, \hat{S}_{i,2}, \ldots, \hat{S}_{i,q})$. Radar plots of normalized scores allow the visualization of multidimensional ideological movement. Figure 3 suggests that internally generated rewritings remain relatively close to the original right-wing human texts, whereas divergent prompting produces broader multidimensional shifts extending beyond the targeted dimension, affecting correlated ideological dimensions as well.

### 4.2 Validation of LMDA-Radar

Because LMDA-Radar transforms raw LMDA scores into bounded dimension scores, it is necessary to verify whether this transformation preserves the ordering produced by the original LMDA procedure. The goal of the validation is not to reproduce the raw score values, since these remain scale-dependent, but to assess whether documents that receive higher or lower scores in LMDA occupy comparable relative positions after normalization. Although LMDA-Radar is grounded on LMDA-derived lexical patterns, the framework also incorporates qualitative dimension descriptions, allowing models to capture broader rhetorical and semantic structures. As a result, perfect alignment between LMDA and LMDA-Radar rankings is neither expected nor necessarily desirable.

To evaluate this, we ranked all 800 documents within each ideological dimension using both the original LMDA scores and the LMDA-Radar scores with three LLMs-as-judges: `Claude Haiku 4.5`, `GPT-5-mini`, and `GPT-4o-mini`. Since these are proprietary systems with undisclosed parameter counts, comparisons were conducted at the behavioral level. We selected compact variants of each model family to reduce computational costs. We then computed Spearman's rank correlation coefficient between the two rankings: $\rho_j = \mathrm{corr}_{\mathrm{Spearman}}(\mathrm{rank}(\mathbf{S}_{:,j}), \mathrm{rank}(\hat{\mathbf{S}}_{:,j}))$, where $\mathbf{S}_{:,j}$ denotes the original LMDA scores for dimension $j$, and $\hat{\mathbf{S}}_{:,j}$ denotes the corresponding normalized LMDA-Radar. Spearman's $\rho$ is appropriate here because the validation concerns rank preservation rather than equality of score magnitudes.

## 5 Ideological Steering Evaluation using LMDA-Radar

To evaluate ideological steering under controlled conditions, we constructed a stratified 10% sample of the corpus. The original dataset consists of 800 texts distributed into 20 balanced cells defined by publication year, country, and political orientation. From each cell, 4 texts were selected, resulting in 80 original documents in total, equally divided between left-wing and right-wing texts (40 each).

Using `GPT-4o`, each subset of 40 texts was then submitted to 11 rewriting configurations: one internal rewriting condition, five convergent rewriting conditions, and five divergent rewriting conditions. This resulted in 440 rewritten texts per ideological subset, totaling 880 rewritten texts. When combined, the final evaluation dataset comprises 960 documents. Finally, both original and rewritten texts were scored using LMDA-Radar in a single run.

### 5.1 Steering Success Criteria

Steering success was evaluated by comparing the LMDA-Radar dimension score of each rewritten text against its original human-written baseline. A rewriting attempt was considered successful when two conditions were simultaneously met: (i) the rewritten text occupied the targeted ideological pole, and (ii) the rewritten score moved further toward that pole relative to the original text. Formally, if the target pole was positive, success required $S^r_{i,j} > S^h_{i,j}$; if the target pole was negative, success required $S^r_{i,j} < S^h_{i,j}$, where $S^h_{i,j}$ and $S^r_{i,j}$ denote the human and rewritten scores for document $i$ in dimension $j$, respectively. Rewritten scores equal to zero were not considered valid realizations of a target pole, since they may emerge from compensation effects between opposing lexical poles rather than from a clear ideological orientation.

## 6 Results & Discussion

### 6.1 LMDA-Radar Scoring Validation Results

Table 2 presents the Spearman correlations between the original LMDA rankings and the rankings produced by LMDA-Radar. The top/bottom 20 evaluation traces back to the microanalysis stage, since these texts were originally used to interpret each

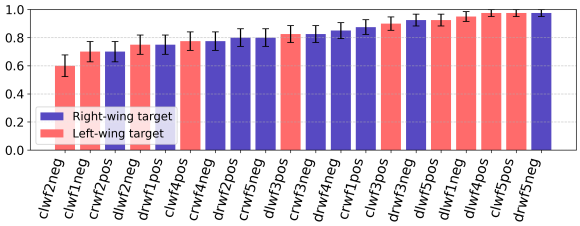



Figure 4: Proportional dimension steering success by configuration. A steering attempt was considered successful when the rewritten score $S_r$ belonged to the targeted ideological pole and, if so, was more salient than the the corresponding human baseline $S_h$.

ideological dimension. Overall, the results indicate that LMDA-Radar was able to preserve the relative ordering of documents with stronger ideological signal within dimensions.

Table 2: Spearman correlations ($\rho$) between the original LMDA rankings and the LMDA-Radar rankings considering (i) the full corpus ($n = 800$) and (ii) the top and bottom 20 scoring texts of each dimension ($n = 40$).

| | Full Corpus | | | Top/Bottom 20 Texts | | |
|---|---|---|---|---|---|---|
| **Dim** | **(1)** | **(2)** | **(3)** | **(1)** | **(2)** | **(3)** |
| 1 | 0.394 | **0.459** | 0.383 | 0.626 | **0.769** | 0.618 |
| 2 | 0.465 | **0.486** | 0.199 | 0.713 | **0.752** | 0.496* |
| 3 | **0.491** | 0.448 | 0.329 | 0.843 | **0.859** | 0.805 |
| 4 | 0.598 | **0.635** | 0.474 | 0.778 | 0.773 | **0.790** |
| 5 | 0.375 | **0.408** | 0.294 | **0.827** | 0.823 | 0.746 |
| Mean | 0.464 | **0.487** | 0.336 | 0.758 | **0.795** | 0.691 |

*Note.* (1) Claude Haiku 4.5, (2) GPT-5-mini, (3) GPT-4o-mini. All coefficients reached statistical significance at $p < 0.001$, except the coefficient marked with (*), which reached significance at $p = 0.01$.

Among the evaluated models, `GPT-5-mini` obtained the highest overall agreement with the original LMDA rankings, both when considering the full dataset ($\bar{\rho} = 0.487$) and when restricting the analysis to the top and bottom 20 scoring texts of each dimension ($\bar{\rho} = 0.795$). Therefore, in the following experiments, we report steering results using only this model to reduce unnecessary computational costs.

## 6.2 Ideological Steering Results

**Steering success** Figure 4 indicates that LMDA-guided ideological steering was generally effective, with most conditions achieving high success rates. Although performance varied, both convergent and divergent rewriting conditions were able to move texts toward the targeted ideological poles. Configurations targeting Dimensions 4 and 5 had the highest success rates overall, whereas those that targeted Dimensions 1 and 2 appeared comparatively less effective.

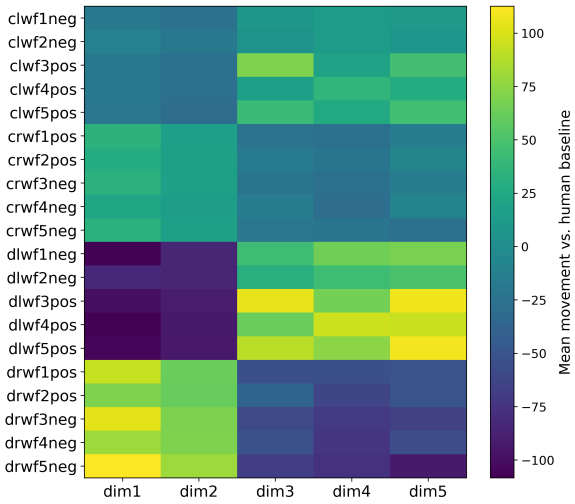



Figure 5: Mean cross-dimensional movement by steering configuration shows how far rewritten text moved from original human-written baseline.

**Cross-dimensional movement** As Fig. 5 shows, steering toward a given ideological pole tends to propagate to correlated discourse structures. This confirms that the LMDA dimensions are not independent, but organized into broader ideological cluster. For example, Dimensions 3 and 5 are closely related in topic (both cover historical aspects of immigration), hence, steering one inevitably steers the other.

Convergent and divergent conditions have distinct movement profiles. Convergent steering generally produced more localized and moderate shifts concentrated around the targeted dimension. On the other hand, divergent steering produced substantially larger multidimensional displacement, which is expected, since reversing the ideological orientation of a text requires broader restructuring of its discourse profile than merely amplifying an existing orientation. For instance, `dlw` configurations yielded some of the largest shifts overall, even in dimensions that were not targeted by the configurations. However, these larger movements should not be interpreted as evidence of preference toward a particular political orientation; rather, they suggest that some source texts were initially further from the targeted ideological poles.

**Human vs. Internal baselines** When the model was prompted to rewrite texts using its internal knowledge, ideological shift is subtle, but different

in size depending on the political orientation of the source texts. Tab. 3 reports a pattern in the internal rewriting condition: internally generated rewritings tended to shift the ideological profiles away from the original human-written baselines and toward the opposite ideological direction. This tendency was consistent for originally right-wing texts. Conversely, originally left-wing texts showed weaker and less systematic movement in the opposite direction. Although these results do not, by themselves, demonstrate ideological preference, they suggest that internally generated rewriting may not preserve the original ideological configuration.

Table 3: Mean LMDA-Radar dimensional movement of internally generated texts (`ilw`, `irw`) relative to the human-written baseline (`hlw`, `hrw`).

| Conf. | n | Dim1 | Dim2 | Dim3 | Dim4 | Dim5 |
|---|---|---|---|---|---|---|
| `ilw` | 40 | 1.475 | 5.825 | 1.425 | -5.200 | 0.400 |
| `irw` | 40 | -18.275 | -9.125 | 7.675 | 5.925 | 13.000 |

## 7 Related Work

Previous research has demonstrated that language models can reflect political and social biases. Following the detection of human ideological biases in surveys, toxic language datasets, and news events (Yoder et al., 2022; Sahoo et al., 2022; Nakshatri et al., 2025), recent research has turned its attention to understanding similar biases reproduced by language models. Above all, social and political biases have received significant attention given their impact on public discourse and society (Blodgett et al., 2020; Rettenberger et al., 2025; Sheng et al., 2019; Santurkar et al., 2023; Durmus et al., 2024; Wan et al., 2023; Tan and Lee, 2025; Lin et al., 2025; Dai et al., 2025; Bang et al., 2024; Prama and Islam, 2025; Jenny et al., 2024; Lunardi et al., 2024). Studies have shown that language models reproduce social biases related to gender, race, and religion, which emerge from their training data and learned representations (Feng et al., 2023; Zhao et al., 2019; Kurita et al., 2019; Basta et al., 2019; Bender et al., 2021), with noticeable left-leaning bias in several LLMs (Hartmann et al., 2023; Rozado, 2023, 2024; Motoki et al., 2024; Hagendorff, 2025; Faulborn et al., 2025). In this sense, the present study moved beyond traditional bias detection approaches by employing LMDA as a multidimensional discourse framework rather than treating ideology as a binary or single-axis phenomenon. Whereas much previous work evaluates political bias through discrete labels, we presented interacting lexical-discursive dimensions derived from corpus structure.

Model outputs are susceptible to modulation through prompt engineering and finetuning strategies. The problem of prompt brittleness, where LLMs present substantial variations in generated outputs due to non-semantic changes in prompt formatting, has been investigated (Ngweta et al., 2025; Li et al., 2025; Yang, 2025), including its effect on political leaning (Ceron et al., 2024). Research on the effect of deliberate poisoning and ideological priming via prompt designing or via finetuning shows that models can be steered toward specific ideological position or behavior (Chen et al., 2024; Shu et al., 2023; Agiza et al., 2024; Zhou et al., 2023; Bernardelle et al., 2025; Stammbach et al., 2024; Wright et al., 2024). Our findings align with previous research showing that ideological profiles are highly steerable through prompting. However, the observed effects suggest asymmetric multidimensional steerability rather than isolated bias toward a single political orientation, with some asymmetries potentially reflecting interactions between prompt structure and corpus organization.

## 8 Conclusion

This study provides systematic evidence that LLM ideological positioning can be influenced through targeted prompting in political news writing contexts. Our results show that ideological steering can restructure multidimensional discourse profiles, highlighting the ethical challenges involved in deploying highly steerable generative systems at scale. Although instruction-following is a core property of LLMs, the present findings raise concerns about the potential misuse of these systems for ideological manipulation. The ability of LLMs to reproduce complex discourse structures does not imply that they should uncritically generate socially damaging narratives when prompted to do so. Future research should further investigate which lexical and discourse signals are incorporated during the rewriting process, as well as which signals are prioritized by evaluation models when assigning ideological classifications.

## 9 Limitations

**Anglocentric and U.S.-centered perspective** The experiments were conducted exclusively in English and relied primarily on political discourses

grounded in Western, particularly United-Statesian and British, sociopolitical contexts. As ideological structures, political categories, and discourse practices vary substantially across languages and societies, the findings may not generalize to non-Anglophone or non-Western contexts.

**Domain and corpus dependency** The analysis focused exclusively on immigration discourse, and the observed steering behaviors may differ in other ideological domains. In addition, because LMDA derives dimensions from corpus-specific lexical co-occurrence patterns, the extracted ideological configurations reflect the statistical organization of the present dataset. Different domains or corpora may therefore yield different ideological dimensions and steering dynamics.

**Factor interpretation** Our framework depends on the interpretive stage of LMDA. While factor extraction is statistically grounded, the labeling and qualitative interpretation of dimensions involve analytical decisions made during microanalysis. Although these interpretations were anchored in representative texts and high-loading lexical items, alternative readings of the dimensions remain possible.

**Prompt brittleness** Prior research has shown that LLM outputs may vary substantially under small prompt modifications. Although the present prompts were grounded in LMDA-derived ideological cues, different prompt formulations may produce distinct behaviors.

## 10 Ethical considerations

This study investigates ideological steering in LLM-generated immigration discourse, a topic that may include politically sensitive, polarizing, or potentially offensive content. Because the corpus contains real-world political narratives from both left-wing and right-wing perspectives, some texts may reproduce harmful stereotypes, hostile rhetoric, discriminatory framing, or exclusionary language related to immigration and national identity. Such content is analyzed strictly for research purposes and does not reflect the authors' views.

The goal of this work is not to endorse ideological manipulation, but to better understand how LLMs respond to ideological steering and how such systems may be exploited in real-world settings. By examining these behaviors under controlled experimental conditions, we aim to contribute to the development of more transparent and ideologically robust language technologies. The generated data should not be used outside research contexts, particularly for political persuasion, targeted propaganda, or the dissemination of harmful ideological content.

**Use of AI Tools** We used AI for language editing, paraphrasing, and the creation of illustrative visual elements in some figures. These AI-generated elements were limited to conceptual or decorative components and did not include experimental plots, quantitative results, or analytical visualizations. All analyses, interpretations, and final revisions were performed and verified by the authors.

## 11 Computational budget

All experiments were conducted through commercial API access using a standard personal computer for orchestration and data processing. Costs were primarily associated with remote inference usage and API calls to proprietary LLM services.

## References


Ahmed Agiza, Mohamed Mostagir, and Sherief Reda. 2024. PoliTune: Analyzing the Impact of Data Selection and Fine-Tuning on Economic and Political Biases in Large Language Models. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):2–12.

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring Political Bias in Large Language Models: What Is Said and How It Is Said. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11142–11159, Bangkok, Thailand. Association for Computational Linguistics.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Pietro Bernardelle, Stefano Civelli, Leon Fröhling, Riccardo Lunardi, Kevin Roitero, and Gianluca Demartini. 2025. Political Ideology Shifts in Large Language Models. *arXiv preprint*. ArXiv:2508.16013 [cs].

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. Beyond Prompt Brittleness: Evaluating the Reliability and Consistency of Political Worldviews in LLMs. *Transactions of the Association for Computational Linguistics*, 12:1378–1400.

Kai Chen, Zihao He, Jun Yan, Taiwei Shi, and Kristina Lerman. 2024. How Susceptible are Large Language Models to Ideological Manipulation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17140–17161, Miami, Florida, USA. Association for Computational Linguistics.

Sunhao Dai, Zhanshuo Cao, Wenjie Wang, Liang Pang, Jun Xu, See-Kiong Ng, and Tat-Seng Chua. 2025. Media Source Matters More Than Content: Unveiling Political Bias in LLM-Generated Citations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17256–17276, Suzhou, China. Association for Computational Linguistics.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards Measuring the Representation of Subjective Global Opinions in Language Models. *arXiv preprint*. ArXiv:2306.16388 [cs].

Mats Faulborn, Indira Sen, Max Pellert, Andreas Spitz, and David Garcia. 2025. Only a Little to the Left: A Theory-grounded Measure of Political Bias in Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31684–31704, Vienna, Austria. Association for Computational Linguistics.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Thilo Hagendorff. 2025. On the Inevitability of Left-Leaning Political Bias in Aligned Language Models. *arXiv preprint*. ArXiv:2507.15328 [cs].

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint*. ArXiv:2301.01768 [cs].

David F. Jenny, Yann Billeter, Bernhard Schölkopf, and Zhijing Jin. 2024. Exploring the Jungle of Bias: Political Bias Attribution in Language Models via Dependency Analysis. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 152–178, Miami, Florida, USA. Association for Computational Linguistics.

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025. Break the Checkbox: Challenging Closed-Style Evaluations of Cultural Alignment in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24–51, Suzhou, China. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Jiahui Li, Sean Papay, and Roman Klinger. 2025. Are Humans as Brittle as Large Language Models? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2130–2155, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2025. Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10634–10649, Abu Dhabi, UAE. Association for Computational Linguistics.

Riccardo Lunardi, David La Barbera, and Kevin Roitero. 2024. The Elusiveness of Detecting Political Bias in Language Models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pages 3922–3926, New York, NY, USA. Association for Computing Machinery.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring ChatGPT political bias. *Public Choice*, 198(1):3–23.

Afrozah Nadeem, Mark Dras, and Usman Naseem. 2025. Steering Towards Fairness: Mitigating Political Stance Bias in LLMs. In *Proceedings of the 8th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Texts*, pages 52–61, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Nishanth Sridhar Nakshatri, Nikhil Mehta, Siyi Liu, Sihao Chen, Daniel Hopkins, Dan Roth, and Dan Goldwasser. 2025. Talking Point based Ideological Discourse Analysis in News Events. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 575–594, Vienna, Austria. Association for Computational Linguistics.

Lilian Ngweta, Kiran Kate, Jason Tsay, and Yara Rizk. 2025. Towards LLMs Robustness to Changes in Prompt Format Styles. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 529–537, Albuquerque, USA. Association for Computational Linguistics.

Tabia Tanzin Prama and Md. Saiful Islam. 2025. Evaluating Credibility and Political Bias in LLMs for News Outlets in Bangladesh. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 665–677, Vienna, Austria. Association for Computational Linguistics.

Luca Rettenberger, Markus Reischl, and Mark Schutera. 2025. Assessing political bias in large language models. *Journal of Computational Social Science*, 8(2):42.

David Rozado. 2023. The Political Biases of ChatGPT. *Social Sciences*, 12(3):148.

David Rozado. 2024. The political preferences of LLMs. *PLOS ONE*, 19(7):e0306621.

Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. 2022. Detecting Unintended Social Bias in Toxic Language Datasets. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 132–143, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? In *Proceedings of the 40th International Conference on Machine Learning*, pages 29971–30004. PMLR. ISSN: 2640-3498.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT Workshop*, Dublin, Ireland.

Elena Shalevska. 2024. The Future of Political Discourse: AI and Media Literacy Education. *Journal of Legal and Political Education*, 1(1):50–61.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Manli Shu, Jiongxiao Wang, Chen Zhu, Jonas Geiping, Chaowei Xiao, and Tom Goldstein. 2023. On the exploitability of instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 61836–61856. Curran Associates, Inc.

Dominik Stammbach, Philine Widmer, Eunjung Cho, Caglar Gulcehre, and Elliott Ash. 2024. Aligning Large Language Models with Diverse Political Viewpoints. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7257–7267, Miami, Florida, USA. Association for Computational Linguistics.

Bryan Chen Zhengyu Tan and Roy Ka-Wei Lee. 2025. Unmasking Implicit Bias: Evaluating Persona-Prompted LLM Responses in Power-Disparate Social Scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1075–1108, Albuquerque, New Mexico. Association for Computational Linguistics.

Teun A. van Dijk. 1993. Principles of Critical Discourse Analysis. *Discourse & Society*, 4(2):249–283.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. LLM Tropes: Revealing Fine-Grained Values and Opinions in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17085–17112, Miami, Florida, USA. Association for Computational Linguistics.

Sewon Yang. 2025. PRISM: Overcoming Brittleness and Scalability in Large Language Models.

Michael Yoder, Lynnette Ng, David West Brown, and Kathleen Carley. 2022. How Hate Speech Varies by Target Identity: A Computational Analysis. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 27–39, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tianyi Zhang. 2025. Probing Political Ideology in Large Language Models: How Latent Political Representations Generalize Across Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23349–23360, Suzhou, China. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaotian Zhou, Qian Wang, Xiaofeng Wang, Haixu Tang, and Xiaozhong Liu. 2023. Large Language Model Soft Ideologization via AI-Self-Consciousness. *arXiv preprint*. ArXiv:2309.16167 [cs].

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1):237–291.

# A Prompts

## A.1 Rewriting prompts

Figure 6 shows the rewriting templates per condition. The internal condition relies exclusively on the model's internal knowledge, whereas the convergent and divergent conditions incorporate LMDA-derived ideological cues.

## A.2 LMDA-Radar prompt

The prompt template used in LMDA-Radar is shown in Fig. 7.

**Prompt Templates for Rewriting**

**Internal Rewriting**

```
You are a news writer. Rewrite the text below to reflect your own internal knowledge on immigration.
DON'T WRITE META-COMMENTS. JUST OUTPUT THE TEXT.
Text to rewrite: {Text}
```

**Guided Ideological Rewriting**

```
You are a news writer. Your news agency is a {target_ideology} organization. Rewrite the text below to reflect a {target_ideology} discourse perspective on immigration.
DON'T WRITE META-COMMENTS. JUST OUTPUT THE TEXT.
Ideological cues:
{Factor}
{Polarity}
{Dimension Label & Wing}
{Dimension Description}
{Lexical Items}
Text to rewrite: {Text}
```

*Convergent rewriting occurs when {target_ideology} matches the ideological orientation of the original text. Divergent rewriting occurs when {target_ideology} corresponds to the opposite ideological orientation.*

Figure 6: Prompt templates used in the ideological rewriting experiment. `{target_ideology}` indicates the ideological orientation assigned to the rewritten text (left-wing or right-wing). `{Factor}` identifies the LMDA dimension associated with the steering condition, whereas `{Polarity}` specifies whether the positive or negative pole is targeted. `{Dimension Label & Wing}` corresponds to the interpretive ideological label assigned during microanalysis. `{Dimension Description}` summarizes the dominant discourse associated with the pole, and `{Lexical Items}` contains representative lexical items with high factor loadings. `{Text}` corresponds to the original news article submitted for rewriting.

**LMDA-Radar Scoring Prompt Template**

**Role.** You are a discourse analyst specializing in political language.

**Task.** Score the input text on the ideological dimension poles extracted through LMDA.

**Instructions.**

- Each pole is defined by a dimension label, a short description, and characteristic lexical items.
- Lexical items are treated as the primary evidence for scoring.
- Rhetorical framing and argumentative stance may be used as secondary evidence.
- Each pole receives a score from 0 to 100.

**Text to analyze:** `{TEXT_TO_ANALYZE}`

**Dimension poles:**

**F1_Positive:** `{dimension label}`
Description: `{dimension description}`
Typical words: `{lexical items}`
**F1_Negative:** `{dimension label}`
Description: `{dimension description}`
Typical words: `{lexical items}`
...
**F5_Negative:** `{dimension label}`
Description: `{dimension description}`
Typical words: `{lexical items}`

**Required output format:**

```
{
  "F1_Positive": {"score": X},
  "F1_Negative": {"score": X},
  . . .
  "F5_Positive": {"score": X},
  "F5_Negative": {"score": X}
}
```

Figure 7: Template of the LMDA-Radar scoring prompt. The prompt provides the model with the text to be analyzed, the LMDA-derived pole descriptions, and the characteristic lexical items associated with each pole. The model is instructed to return only a JSON object containing scores from 0 to 100 for each ideological pole.

# B Additional results

## B.1 LMDA

Alg. 1 presents an overview of the LMDA pipeline adopted in this study. Starting from a corpus of texts, the procedure extracts keywords, computes lexical correlations, and identifies latent ideological dimensions through factor analysis. Keywords are then separated into opposing poles according to their factor loadings, allowing the computation

of dimension scores for each document. Finally, representative texts and lexical patterns are examined through microanalysis to interpret and label the ideological dimensions. Although qualitative, this procedure adds interpretative depth to the statistical model by connecting lexical distributions to concrete discourse realizations. In this sense, the manual examination does not weaken the framework; rather, it complements the quantitative stage with linguistic interpretation, allowing the dimensions to be described in theoretically meaningful terms. The resulting dimensions are described in Tab. 5.

### B.2 Corpus Annotation Validation

To validate the annotation procedure, following Chen et al. (2024), who reports high inter-annotator agreement for `GPT-4` in right–left political orientation assessment, we measure inter-annotator agreement between the discourse analyst and LLM-based judges, namely `Claude Haiku 4.5`, `GPT-4o-mini`, and `GPT-5-mini`. Due to the computational cost associated with large-scale LLM-based annotation, inter-annotator agreement was evaluated on a stratified subset corresponding to 10% of the corpus consisting of four texts from each cell defined by the combination of country, year, and partisan orientation. Then, each model received the following instruction: "*You are an expert political discourse analyst. Classify the following news article on immigration according to its political alignment. Return only one label: left or right. Do not explain your answer*." Table 4 displays the raw proportion of identical labels between the expert and each model, Cohen's $\kappa$, and Krippendorff's $\alpha$. The high agreement between expert annotations and LLM-based judges suggests that the political leanings in the corpus are consistently identifiable, supporting the validity of the annotation procedure.

### B.3 LMDA-Radar Ideological Profiles

Figure 8 presents the average LMDA-Rdar profiles of divergent and convergent configurations against human and internal baselines. Figure 9 show all steering configurations separately.

### B.4 Factor Steering Success rate

Because dimension scores may approach zero through compensation between positive and negative poles, factor-pole scores were additionally inspected to verify whether a target ideological pole was effectively activated. Factor-level steering success was evaluated by comparing the LMDA-Radar factor-pole scores of rewritten texts against their

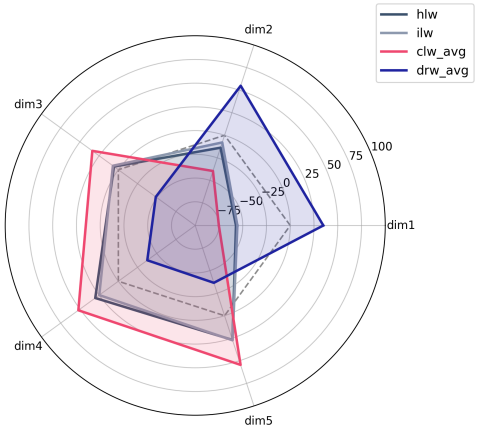
**Algorithm 1:** LMDA

**Input:** Corpus $\mathcal{D} = \{d_1, \ldots, d_n\}$
**Output:** Dimensions $\mathcal{F}$

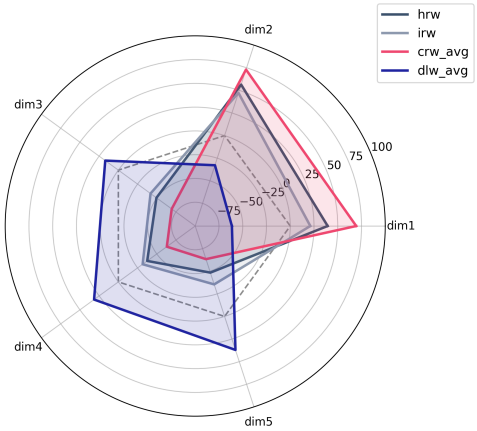
// Keep only content words
$\mathcal{V} \leftarrow \mathrm{POSFilter}(\mathcal{D})$
// Extract keywords
$\mathcal{K} \leftarrow \mathrm{Keyness}_{G^2}(\mathcal{V})$
// Build the term-document matrix
$\mathbf{X} \in \mathbb{R}^{n \times |\mathcal{K}|} \leftarrow \mathrm{TDM}(\mathcal{D}, \mathcal{K})$
// Correlate keywords
$\mathbf{R} \leftarrow \mathrm{corr}_{\mathrm{Pearson}}(\mathbf{X})$
// Select the top and bottom 400 most correlated keywords
$\mathcal{K}_{sel} \leftarrow \mathrm{Top}^{+}_{400}(\mathbf{R}) \cup \mathrm{Top}^{-}_{400}(\mathbf{R})$
$\mathbf{X}_{sel} \leftarrow \mathbf{X}_{[:, \mathcal{K}_{sel}]}$
// Upgrade to polychoric correlations
$\mathbf{R}_{poly} \leftarrow \mathrm{corr}_{polychoric}(\mathbf{X}_{sel})$
// Extract latent factors
$\mathbf{L}^{(0)} \leftarrow \mathrm{PAF}(\mathbf{R}_{poly})$
// Decide how many factors to keep
$q \leftarrow \mathrm{Eigenvalue}(\mathbf{R}_{poly}) + \mathrm{ScreePlot}(\mathbf{R}_{poly})$
// Rotate obliquely
$\mathbf{L} \leftarrow \mathrm{Promax}(\mathbf{L}^{(0)}, q)$
// Assign each keyword to a factor $j$
**for each** $v \in \mathcal{K}_{sel}$ **do**
  $j^*(v) \leftarrow \arg\max_{j \in \{1, \ldots, q\}} |L_{v,j}|$
// Split keywords into positive and negative factors
**for each** $j$ **do**
  $\mathcal{V}_j^{+} \leftarrow \{v : j^*(v) = j \ \wedge \ L_{v,j} > 0.3\}$
  $\mathcal{V}_j^{-} \leftarrow \{v : j^*(v) = j \ \wedge \ L_{v,j} < -0.3\}$
// Score each text to obtain dimension scores matrix $\mathbf{S}$
**for each** $d_i \in \mathcal{D},\ j \in \{1, \ldots, q\}$ **do**
  $S_{i,j} \leftarrow \sum_{v \in \mathcal{V}_j^{+}} X_{i,v} - \sum_{v \in \mathcal{V}_j^{-}} X_{i,v}$
// Select top and bottom 20 scoring texts per dimension
$\mathcal{E}_j \leftarrow \mathrm{TopBottom}_{20}(S_{:,j})$
// Interpret and label dimensions
$\mathcal{F} \leftarrow \mathrm{Microanalysis}(\mathcal{D}, \mathcal{E}_j, \mathcal{V}_j^{+}, \mathcal{V}_j^{-})$
**return** $\mathcal{F} = \{f_{1+}, f_{2-}, \ldots, f_{j+}, f_{j-}\}$

Table 4: Inter-annotator agreement between expert labels and LLM-based judges on a stratified 10% sample of the corpus ($n = 80$).

| Model | Agreement | $\kappa$ | $\alpha$ |
|---|---|---|---|
| Claude Haiku 4.5 | 0.9250 | 0.850 | 0.8509 |
| GPT-4o-mini | **0.9875** | 0.975 | 0.9752 |
| GPT-5-mini | 0.8125 | 0.625 | 0.6202 |




(a) Originally left-wing source texts.




(b) Originally right-wing source texts.

Figure 8: Average LMDA-Radar ideological profiles. Mean human-written texts (`hlw`, `hrw`) scores are compared against average internally generated rewritings (`ilw`, `irw`) and the mean convergent/divergent steering configurations associated with each ideological orientation.

corresponding human-written baselines. In convergent conditions, a rewriting attempt was considered successful when the targeted factor pole increased relative to the original human text. In divergent conditions, success was evaluated through pole dominance: the targeted factor-pole score had to become larger than the opposite pole score within the same factor.

### B.5 Internal Baseline: Cosine Similarity

Since each text is represented as a five-dimensional LMDA-Radar vector, cosine similarity was used to compare the multidimensional ideological profiles of steering configurations against internally generated texts.

Figure 10 indicates that LMDA-guided steering systematically reorganized the multidimensional ideological profiles toward the corresponding internal baseline orientation. This tendency was particularly consistent in right-wing-targeting configurations (`crw`, `drw`), whose similarity to the `irw` baseline remained relatively stable regardless of the targeted factor pole.

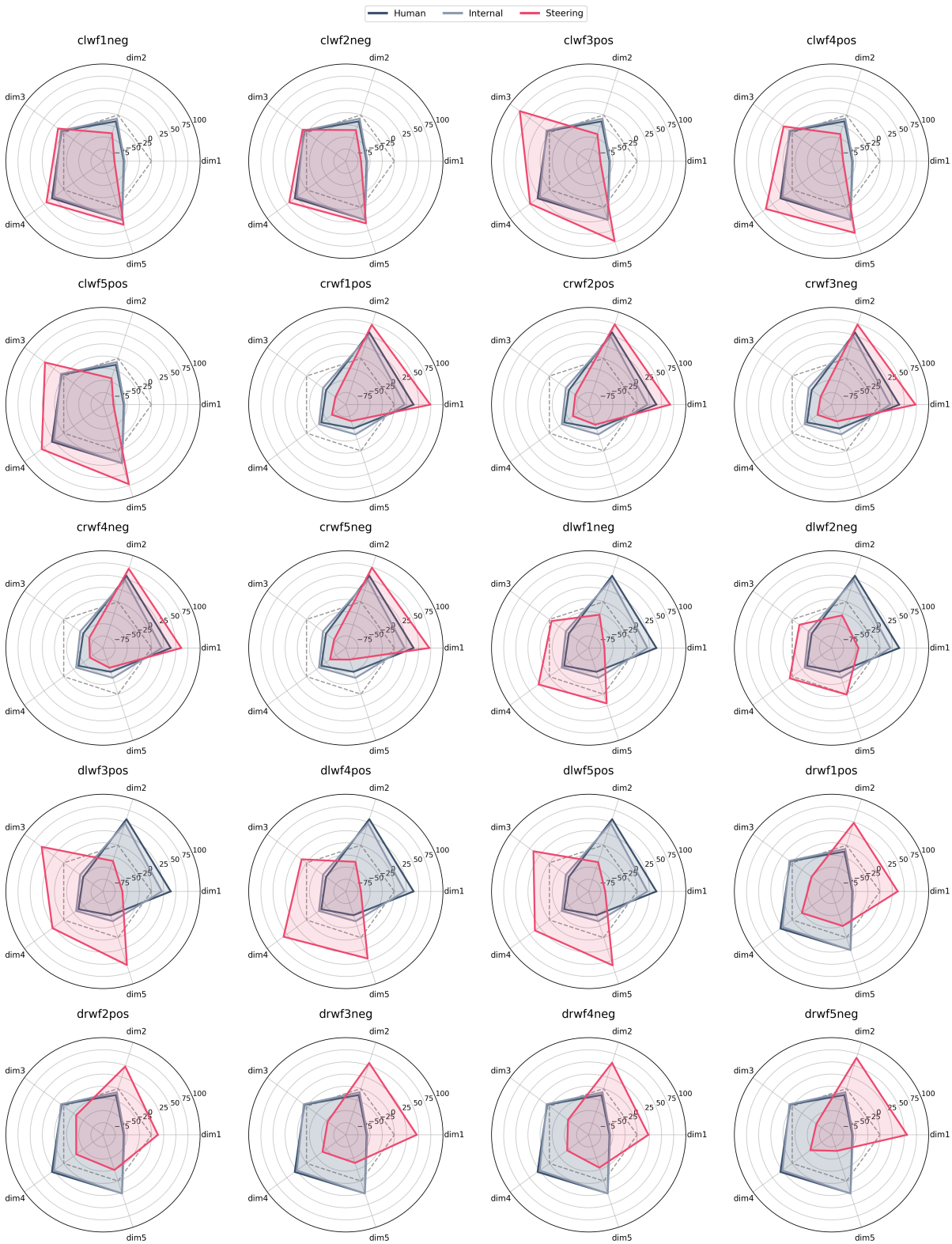



Figure 9: Average LMDA-Radar profiles for each convergent and divergent configuration compared against human and internal baselines.

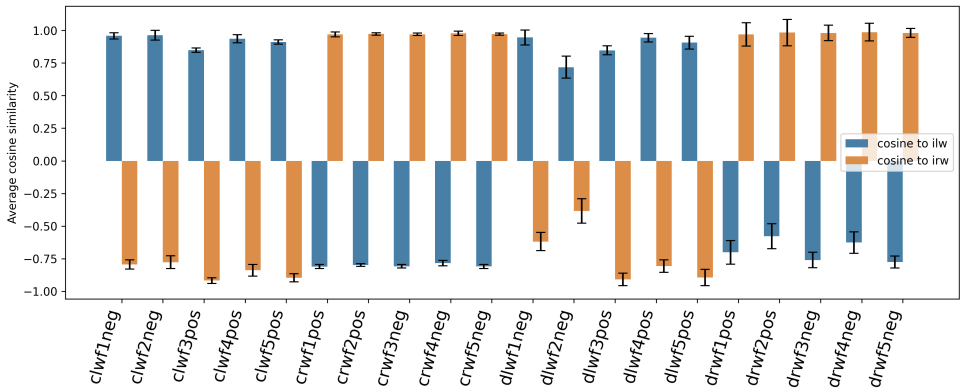



Figure 10: Average cosine similarity (Steering configurations to Internal configurations).

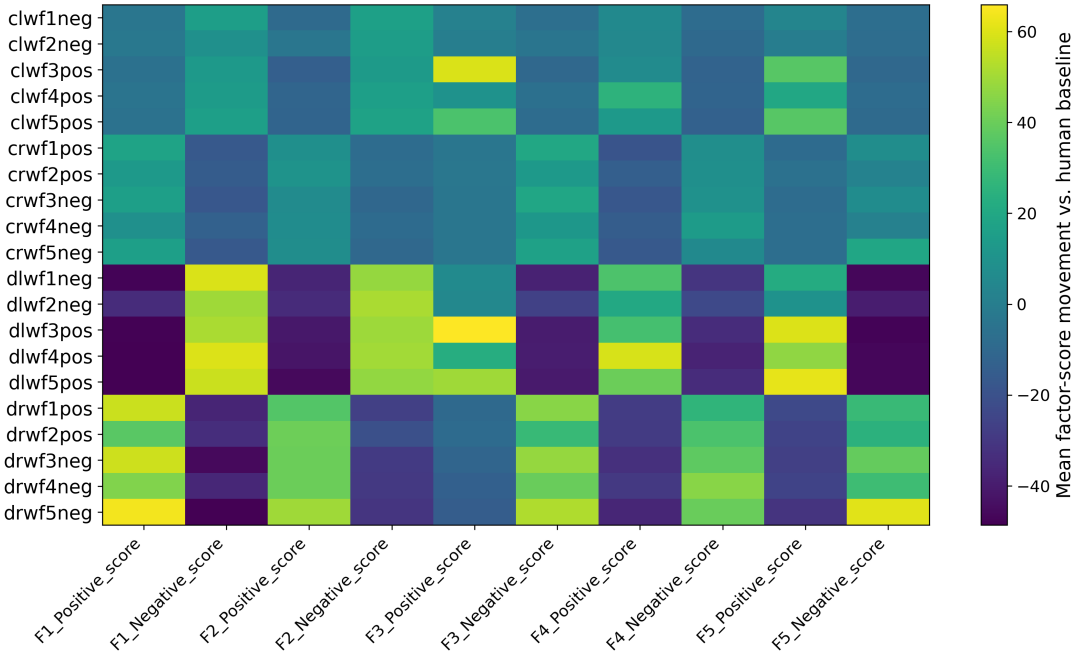



Figure 11: Mean cross-factor movement produced by each steering configuration relative to the original human-written baseline. Convergent configurations generally produced localized amplification around the targeted factor poles, whereas divergent configurations resulted in broader multidimensional restructuring.

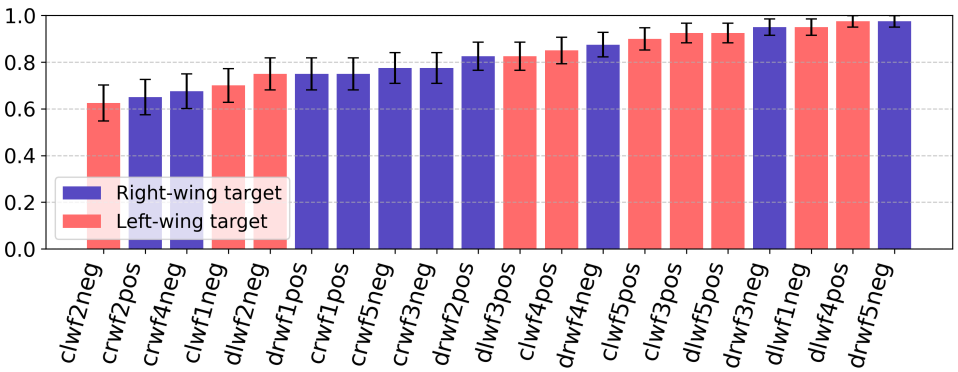



Figure 12: Proportional factor-level steering success by configuration. In convergent conditions, success indicates that the targeted factor pole became more salient relative to the original human-written baseline. In divergent conditions, success indicates that the targeted factor pole became dominant over its opposite pole within the same factor.

Table 5: Ideological Dimensions Identified Through LMDA

| Factor | Positive Pole (+) | Negative Pole (-) |
|---|---|---|
| F1 | **Nationalist and Security-Oriented (Right-Wing)** Immigration as a threat to cultural identity and national stability. Emphasis on border control, terrorism, crime, and populist protectionism. | **Humanitarian Protection (Left-Wing)** Critique of punitive policies. Advocacy for fair treatment, reform, and civil rights. Highlights immigrant contributions and lived experiences. Emphasizes empathy and systemic justice. |
| F2 | **Sovereignty and Political Strategy (Right-Wing)** Immigration as enforcement issue. Emphasis on border security, legal authority, and fiscal discipline. Immigration as political instrument for power consolidation. | **Inclusion and Cultural Support (Left-Wing)** Moral responsibility and cultural inclusion. Emphasizes multiculturalism, integration support, and ethical duty toward displaced populations. |
| F3 | **Post-Colonial Justice (Left-Wing)** Links immigration injustices to colonialism and institutional racism. Cultural identity, resistance, and right to belong. Calls for historical accountability and reform. | **National Security (Right-Wing)** Immigration as threat to national integrity. Confrontational rhetoric emphasizing defense and economic risk. Reactive, control-oriented, nationalist tone. |
| F4 | **Migration Trauma (Left-Wing)** Emotional dimensions: family separation, detention trauma, abuse. Historical parallels to humanitarian disasters. Ethical appeal for empathy and reform. | **Regulatory and Fiscal Liability (Right-Wing)** Technocratic discourse advocating restriction through state policy. Immigrants as risks to manage and costs to contain. Depersonalized governance approach. |
| F5 | **Historical Oppression (Left-Wing)** Immigration within systemic injustice context. Links to colonialism, slavery, racial oppression. Emphasizes psychological toll and demands structural change. | **Criminality (Right-Wing)** Immigration equated with criminality and violence. Binary opposition between citizens and criminal foreigners. Cultural protectionism and demographic anxiety. |

Table 6: Summary of the ideological dimensions extracted through LMDA. Each factor is composed of two opposing poles derived from correlated lexical and discourse patterns identified in the corpus. Positive and negative poles do not represent sentiment polarity, but rather opposite regions of the same latent ideological dimension. The table summarizes the dominant ideological tendencies associated with each pole after the microanalysis stage.