# "My own morality, my own mind": Persona-Conditioned Moral and Cultural Reception on Trump Messages

**Candida Maria Greco[3,1,*], Víctor Yeste[1,2,*], Andrea Tagarelli[3], Paolo Rosso[1,4]**

[1]PRHLT Research Center, Universitat Politècnica de València, Spain
[2]School of Science, Engineering and Design, Universidad Europea de Valencia, Spain
[3]DIMES Department, University of Calabria, Rende, Italy
[4]Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI)
{candida.greco,tagarelli}@dimes.unical.it, vicyesmo@upv.es, prosso@dsic.upv.es

## Abstract

Large language models are increasingly used to simulate social and political audiences, but it remains unclear whether persona-conditioned generations provide coherent reception patterns or align with real public opinion. We study audience reactions to morally salient political messages, using Trump's posts as a relevant case in point, through two theory-grounded human value frameworks, Inglehart–Welzel cultural map and Moral Foundations Theory. Across issue-specific message bundles, we analyze stance together with agreement, legitimacy, offensiveness, and endorsement, asking whether synthetic personas produce systematic reception differences, which persona dimensions explain those differences, and whether aggregate synthetic distributions align with matched Pew public-opinion patterns. The paper examines three related dimensions of synthetic-audience evaluation: internal differentiation, explanatory structure, and external plausibility. It contributes a reproducible study design for political audience simulation and an empirical account of where synthetic moral audiences are informative, and where they should not be treated as calibrated population proxies.

## 1 Introduction

Political messages are not evaluated only by what they say, but also by who receives them. The same message can be read as legitimate, offensive, persuasive, or unacceptable depending on the audience's moral and cultural orientation (Entman, 1993; Chong and Druckman, 2007). NLP research has made substantial progress in identifying moral language and moral frames in political discourse (Johnson and Goldwasser, 2018; Hoover et al., 2020; Roy et al., 2021), but tells us less about how different audiences receive the same morally salient message.

Large language models (LLMs) make audience-conditioned reception studies easier to scale, and have been used to simulate respondents or predict opinion distributions (Argyle et al., 2023; Suh et al., 2025). Yet, persona-steered models can misrepresent public opinion, amplify stereotypes, or respond unevenly to persona specifications (Bisbee et al., 2024; Liu et al., 2024). Evaluation should distinguish internal differentiation, interpretable explanatory structure, and external plausibility.

Within this view, we examine morally salient social-media posts by the polarizing political actor Donald Trump across five issue domains, representing a challenging case of elite political communication (Ott, 2017). By anonymizing the posts and organizing them into topic-specific bundles, we prompt LLM-generated personas to produce structured judgments on stance, agreement, perceived legitimacy, offensiveness, and endorsement. These personas are constructed from Inglehart–Welzel cultural values (Inglehart and Baker, 2000; Inglehart and Welzel, 2005) and Moral Foundations Theory (Graham et al., 2011, 2013). We then compare the resulting aggregated synthetic distributions with weighted Pew public-opinion distributions for corresponding Trump-related issue areas.

**Research questions and contributions.** We ask whether synthetic personas systematically differ in their reception of the same messages (**RQ1**), which persona dimensions explain supportive, rejecting, or ambivalent responses (**RQ2**), and whether synthetic audience distributions align with matched Pew patterns (**RQ3**). By addressing these research questions, our work contributes a controlled synthetic-audience design with structured reaction labels beyond stance, a comparison of cultural-value and moral-foundation persona constructions, and an external Pew-based validation layer.

Overall, this study supports a bounded view of synthetic moral audiences: theory-grounded

[*]These authors contributed equally to this work and are corresponding authors.



*Findings of the Association for Computational Linguistics: EMNLP 2026*, pages 51567–51595
October 24-29, 2026 ©2026 Association for Computational Linguistics

personas reveal interpretable, topic-sensitive differences in political-message reception, especially for morally polarized domains such as `Immigration/border`. At the same time, the Pew comparison shows that external alignment is partial rather than calibrated, underscoring the need to evaluate internal structure and public-opinion validity separately.

**Reproducibility.** Code, rule specifications, and publishable audit artifacts are available in the project repository.[1]

## 2 Related Work

### 2.1 Persona-Conditioned Generation and Synthetic Audiences

LLMs have recently been used as synthetic respondents, simulated publics, and persona-conditioned generators. Argyle et al. (2023) argue that conditioning models on human backstories can recover meaningful variation in survey-like answers, while later NLP work targets distributional public-opinion prediction more directly through survey-data fine-tuning (Suh et al., 2025). Adjacent work shows that language models often reflect uneven social and political defaults rather than neutral opinion distributions (Santurkar et al., 2023), and that persona-steered generation can be less reliable for incongruous or non-stereotypical profiles (Liu et al., 2024). Wang et al. (2025) go further, arguing that models used as substitutes for human participants cannot capture positionality and therefore misportray and flatten identity groups. Prompt interpretation itself is also fragile, which cautions against treating a persona description as a transparent behavioral intervention (Webson and Pavlick, 2022). Our setting uses personas differently: *not as replacements for human respondents*, but as *theory-grounded analytical instruments* for comparing how the same political messages are received across structured audience constructions.

### 2.2 Moral and Cultural Structure in Political Interpretation

Our persona design draws on two established traditions. Moral Foundations Theory models moral judgment through partially distinct intuitive domains and has been used to explain systematic political differences in moral evaluation (Haidt and Joseph, 2004; Graham et al., 2009, 2011, 2013). Inglehart–Welzel value theory provides a separate cultural account organized around traditional versus secular-rational and survival versus self-expression orientations (Inglehart and Baker, 2000; Inglehart and Welzel, 2005). We use these frameworks as complementary coordinates for persona construction, not as exhaustive representations of real individuals; cross-cultural comparability and measurement validity remain contested, especially for broad value dimensions (Alemán and Woods, 2016). Closest to our use of MFT, Smith-Vaniz et al. (2025) administer the MFQ to LLMs directly and report that models do not reproduce human ideological groups, and that stereotyped biographical role-play pushes responses toward more extreme positions. We use MFT in the opposite direction, as a selection variable for constructing contrastive audiences rather than to measure the model's own moral profile.

The message side of our study is connected to a growing literature on moralized political language. NLP work has modeled moral foundations and morality frames in political microblogs (Johnson and Goldwasser, 2018, 2019; Roy et al., 2021), and dedicated corpora support finer-grained study of moral sentiment in social media (Hoover et al., 2020). Political-communication research likewise shows that moral language matters in elite rhetoric and online diffusion (Brady et al., 2017; Wang and Inbar, 2021; Hackenburg et al., 2023). We build on this work by shifting attention from *moral content in messages* to *persona-conditioned moral reception* of those messages.

### 2.3 Validation of LLM-Based Public-Opinion Simulation

LLM-based public-opinion simulation increasingly emphasizes validation beyond face-valid fluency. Prior work shows that synthetic responses can reproduce coarse averages while failing on correlational structure, prompt robustness, or demographic and topic-specific bias (Bisbee et al., 2024; Qu and Wang, 2024; Lee et al., 2024; Boelaert et al., 2025). We follow this cautionary line by treating matched Pew distributions as a *bounded external plausibility check*, not evidence that synthetic personas reproduce real publics.

[1] https://github.com/VictorMYeste/trump-moral-audiences-pipeline

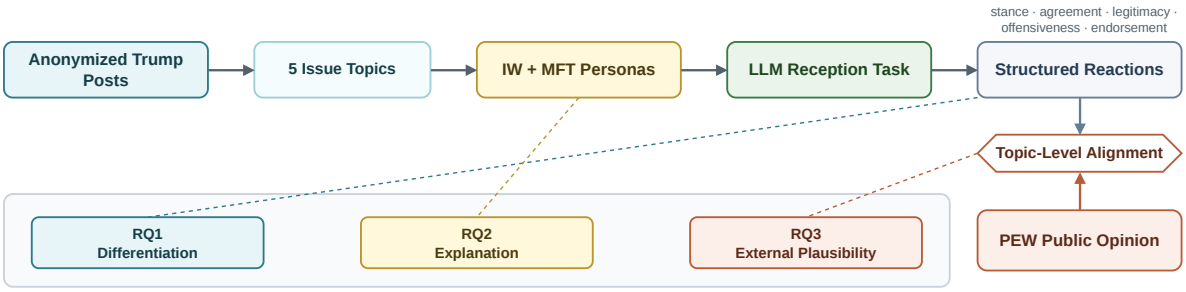


Figure 1: Overview of the proposed methodological flow: topic-filtered Trump posts are paired with IW and MFT personas, evaluated with structured LLM judgments, and compared with matched Pew topic distributions.

## 3 Study Overview

### 3.1 Task Definition

Figure 1 provides an overview of the study pipeline. We model political-message reception as a controlled audience-simulation task. Let $\mathcal{T}$ denote the set of issue topics, and $\mathcal{B}_t$ the set of anonymized message bundles associated with topic $t \in \mathcal{T}$. Each input pairs a topic-specific bundle $b \in \mathcal{B}_t$ with a persona profile drawn from one of two theoretical frameworks: Inglehart–Welzel (IW) cultural values or Moral Foundations Theory (MFT). The model is asked to judge both each post and the bundle as a whole. *For every judgment, we record stance together with agreement, legitimacy, offensiveness, and endorsement*; free-text reasons and emotional reactions are retained as qualitative support but are not the main quantitative outcomes. Appendix B reports the definitions of these response labels.

Each data object target for the evaluation is hence defined as a tuple $(f, t, p, b)$, where $f$ denotes a persona's selection strategy within the IW or MFT framework, $t$ the issue topic, $p$ the selected persona, and $b \in \mathcal{B}_t$ the message bundle. The concrete persona-selection strategies for each framework and the topic-conditioned persona positions are introduced in Section 4.2.

The study covers five issue domains: `COVID/public health`, `Economy/jobs/trade`, `Foreign policy/national security`, `Immigration/border`, and `Judiciary/courts`. These topics are used throughout the synthetic-response analysis and the Pew comparison so that the internal and external evaluations refer to the same issue space. We study *modeled reception of political messages*, not the factual accuracy of the posts, the normative correctness of any response, or the psychological reality of synthetic personas.

### 3.2 Three-Layer Evaluation Design

The paper separates three levels of evidence that are often conflated in synthetic-audience work: *internal differentiation*, whether distinct personas react differently to the same messages; *explanatory structure*, whether response types can be related to interpretable persona dimensions; and *external plausibility*, whether aggregate synthetic distributions correspond directionally to weighted Pew topic distributions.

Table 5 in Appendix A summarizes the data, variables, and analytical procedures associated with each research question, which are described in detail in the following sections.

## 4 Data and Materials

### 4.1 Political Message Corpus

We use two post corpora derived from public Trump social-media archives, spanning the period from 2009 to 2024. The 2009–2021 window comes from the Kaggle Trump Twitter Archive (Heads or Tails, 2021) which itself reproduces The Trump Archive, and the 2022–2024 extension was manually prepared from The Trump Archive (The Trump Archive, 2026). The raw inputs contain 31,366 upstream morally relevant posts: 22,822 from 2009-05-12 to 2021-01-08 and 8,544 from 2022-04-29 to 2024-11-04.

We first applied ME2-BERT (Zangari et al., 2025) to the raw post corpus to identify morally relevant messages. We then applied deterministic filtering for quality, topic assignment, anonymization, and prompt readiness. Topics come from a fixed, auditable registry rather than from a taxonomy induced over the corpus, which keeps the mapping to Pew items reproducible; LLM-based taxonomy induction is the natural alternative when no predefined scheme is available (Brady and Islam, 2025). The full rule table, topic registry,

| Topic | Posts | Pew items |
|---|---|---|
| COVID/public health | 251 | 5 |
| Economy/jobs/trade | 1,116 | 7 |
| Foreign policy/security | 1,112 | 5 |
| Immigration/border | 539 | 4 |
| Judiciary/courts | 300 | 4 |
| Total | 3,318 | 25 |

Table 1: Final issue-topic overlap used for the synthetic-response and Pew-alignment analyses.

moderation-status handling, and anonymization audit are in Appendix C.

After preprocessing, the prompt-ready pool contains 5,270 posts across seven canonical issue topics. For our study, we restrict both posts and Pew items to the five topics with overlap in the external benchmark. This yields 3,318 posts, organized into *topic-specific bundles* so that personas react to issue-level evidence rather than isolated posts. Table 1 reports the final overlap counts.

### 4.2 Synthetic Personas

Our synthetic audience is drawn from the culturally grounded persona resource introduced in (Greco et al., 2026). This resource was constructed by generating LLM-based persona profiles from a controlled configuration space of ten cultural variables derived from the World Values Survey (WVS)[2]: religiosity, child-rearing values, moral acceptability, social trust, political participation, national pride, happiness, gender equality, materialism orientation, and tolerance toward outgroups. Each persona is associated with demographic metadata, a short biographical description, an explicit cultural-variable specification, IW coordinates, and moral-foundation scores. The IW coordinates are derived from persona-conditioned responses to the WVS indicators underlying the Inglehart–Welzel map, whereas the moral-foundation profiles are obtained by eliciting persona-conditioned responses to the 36 items of the Moral Foundations Questionnaire-2 (MFQ-2) (Atari et al., 2023) and aggregating them into six foundation scores. The resulting persona resource was evaluated in Greco et al. (2026) through its organization in the IW space, demographic-level alignment with human WVS-derived response distributions, and the relationship between cultural configurations and MFQ-2 moral profiles.

For each topic $t$, we select personas under three topic-conditioned positions $pos \in \{high, mid, low\}$. These labels denote selection positions within a topic-specific cultural or moral space. *Note that they should not be interpreted as normative or ideological labels, or as direct support for or opposition to the political actor.* Rather, their purpose is to construct contrastive audiences that occupy different positions along the cultural or moral dimensions considered relevant for a given topic, before observing their reactions to the messages.

For each topic, Table 6 in Appendix E specifies the relevant IW dimensions or MFT foundations and provides the corresponding topic-by-topic rationale. This topic-to-dimension mapping is a theory-informed operationalization fixed before persona selection and response generation. Appendix F then formally defines the *high*, *mid*, and *low* positions and the corresponding persona-selection procedures. Persona reaction (e.g., support, neutrality, or rejection) is empirical and not encoded in the *high*/*mid*/*low* labels.

For each $\langle t, pos \rangle$ pair, we build a fixed-size audience of five personas using one of three selection strategies $\mathcal{F} = \{\text{IW-neigh}, \text{IW-contour}, \text{MFT}\}$. The two IW strategies operate in the IW cultural-value space. In IW-neigh, the audience is formed by selecting personas near a topic-specific IW anchor, i.e., the highest-ranked persona for the corresponding topic-position pair. This strategy prioritizes local similarity by selecting personas from the anchor's immediate neighborhood. In IW-contour, the audience is selected to cover a broader region of the corresponding IW contour. Rather than concentrating around the anchor, IW-contour selects high-ranked personas from different regions along the same topic-relevant IW direction, providing broader cultural coverage while preserving the intended topic-conditioned position. The two variants therefore contrast a local audience around the anchor with a less local audience spanning the same topic-relevant region. Appendix F (Section F.1, Figure 5) provides the complete selection procedure and visualization. The MFT strategy instead operates in the six-dimensional MFT space, ranking personas by distance to topic-specific moral target profiles and retaining the closest ones.

For each selection strategy $f \in \mathcal{F}$, topic $t$, and topic-conditioned position $pos$, we denote by $\mathcal{P}_{f,t,pos}$ the set of the top-5 personas used in the experiments.

[2] https://www.worldvaluessurvey.org/

### 4.3 Pew Public-Opinion Benchmark

For external comparison, we use Pew Research Center American Trends Panel (ATP) datasets (Pew Research Center, 2026). From 157 ATP wave folders, we build an inventory of 981 Trump-related candidate rows and retain only Trump-targeted, issue-specific items that use supported response scales and map to exactly one canonical topic. The final Pew subset contains 25 confidence-style items from seven ATP waves in 2020 and 2024. We compare their weighted response distributions with aggregate synthetic topic distributions as an issue-level benchmark. Inclusion rules and excluded-item diagnostics are reported in Appendix D.

## 5 Experimental Setup and Analysis

For all generated contents, we prompt `gpt-oss-20b`, served through `vLLM` via an OpenAI-compatible API; we selected this model for its favorable trade-off between reasoning and generation capabilities, instruction following, and practical deployability, as well as for the transparency and reproducibility enabled by its open weights. We use low-temperature decoding, $T = 0.01$, and set the maximum generation length to 4096 tokens. This setting is intended to reduce stochastic variation while still allowing complete structured JSON outputs.

All experiments were executed on a server with 8× NVIDIA A30 GPUs (24GB VRAM each), 764GB system RAM, and 2× Intel Xeon Gold 6248R processors (96 CPU cores total), running Ubuntu Linux 20.04.6 LTS.

### 5.1 Prompting Procedure and Output Labels

The prompting procedure requires that each request contains a persona profile, the institutional role of the message author (e.g., 'candidate_2024'), the topic-specific anonymized message bundle, and the output JSON schema.

The persona profile varies by framework: IW prompts include the persona biography, cultural grounding, and IW coordinates; MFT prompts include the biography, MFT foundation scores, and item-level MFQ-2 responses. We remove explicit cultural grounding from MFT prompts so that the condition is driven by moral-profile information rather than additional cultural descriptors.

The output JSON schema contains message-level and bundle-level judgments over the same set of labels. In particular, *stance* is a categorical label in $\{support, neutral, reject\}$ and refers to the persona's reception of the underlying topic-related claim or policy direction. The remaining quantitative labels—*agreement*, *legitimacy*, *offensiveness*, and *endorsement*—are valued as 1–5 ratings. The labels operationalize theoretically distinct layers of reception. *Stance* and *agreement* capture substantive reception of the inferred topic-related claim or policy direction (Mohammad et al., 2016; Küçük and Can, 2020); *legitimacy* captures whether that claim is acceptable or justified from the persona's viewpoint, which is distinct from agreeing with it (Tyler, 2006; Gibson et al., 2003); *offensiveness* captures rhetorical rejection of tone and wording (Mutz and Reeves, 2005; Zampieri et al., 2019); and *endorsement* captures willingness to share the message as written, combining substance and style (Brady et al., 2017; Mosleh et al., 2020). A persona may therefore agree with a policy direction yet reject its style, or the reverse. The prompt template, output schema and more details about the operational rules given to the model are reported in Appendix B.

### 5.2 RQ1: Reception Differences Across Personas

We use the *stance* labels as the primary RQ1 outcome, since it directly captures whether the persona supports, rejects, or remains neutral toward the substantive position expressed by the message. We map *stance* labels to ordinal scores, with *support* coded as 1, *neutral* as 0, and *reject* as $-1$.

For each $(f, t, pos)$, we aggregate evaluations for personas $p \in \mathcal{P}_{f,t,pos}$ on bundles $b \in \mathcal{B}_t$. We compute the mean stance score over all the stance labels, both at message level and bundle level.

We perform the same descriptive analysis for the best-profile subset and the top-5 persona sets. The best-profile subset contains the highest-ranked persona for each $(f, t, pos)$; the top-5 analysis aggregates over all personas $p \in \mathcal{P}_{f,t,pos}$. Bundle-level results are used as the main RQ1 presentation layer, with message-level summaries retained as a robustness check.

For the four scalar labels, *agreement*, *legitimacy*, *offensiveness*, and *endorsement*, we compute mean ratings and high-minus-low contrasts within each $(f, t)$ condition; detailed results are reported in Appendix H. Higher values for agreement, legitimacy, and endorsement indicate more favorable substantive reception, whereas higher offensiveness indicates stronger rhetorical or normative rejection; we

therefore interpret offensiveness as a separate style-sensitive dimension rather than as another support score.

### 5.3 RQ2: Persona Dimensions Associated with Response Types

We operationalize RQ2 as a descriptive association analysis between persona-level attributes and bundle-level response types. Since the original persona sets are too small for stable pattern mining, we construct an expanded RQ2-specific set of personas, namely $\widetilde{\mathcal{P}}_{f,t,pos}$. For the IW framework, we expand the neighbor-based method to 100 personas for each $\langle t, pos \rangle$ pair. For the MFT framework, we use 80 personas for each $\langle t, pos \rangle$ pair. The expanded personas are selected to remain close to the original top-5 profiles while increasing demographic variation. Details of this expanded sampling procedure are reported in Appendix G.

The expanded personas are queried similarly as RQ1 but with reduced profiles: only demographic metadata and a short biography are included, while cultural-variable mappings, IW coordinates, MFT scores, and MFQ-2 item responses are omitted. By omitting these fields, we prevent the model from directly conditioning its RQ2 responses on the same cultural and moral variables that were used upstream to select the personas. For each selection strategy $f$, topic $t$, and observed bundle-level stance $s$, we define $G_{f,t,s}$ as the set of unique personas assigned to $f$ and $t$ that produced stance $s$ at least once. This persona counting is carried out at the strategy–topic–persona–stance level, so that repeated bundle records from the same persona do not inflate itemset support. Each of these personas is represented as a transaction of categorical items covering demographics, WVS-derived cultural variables, and discretized theory-specific scores; the demographic attributes used for the expanded RQ2 sample are described in Appendix G. To avoid explaining responses with the same representation used for selection, we add cross-theory items: IW-selected personas receive binned MFT foundation scores, whereas MFT-selected personas receive binned IW coordinates.

We mine frequent itemsets separately within each response group $G_{f,t,s}$, using FP-growth over the transactions $X_p$. The support of an itemset $I$ is the fraction of personas in $G_{f,t,s}$ whose transaction contains all items in $I$. An itemset is frequent if its support $\sigma$ is here regarded at least 0.20. We retain only *closed itemsets* to remove redundant sub-patterns while preserving shared persona attributes. We then apply an adaptive top-support filter within each response group. Let $s_{\max}$ be the maximum support among closed itemsets in $G_{f,t,s}$. We retain only itemsets whose support is at least $\sigma_{\text{top}} = \sigma + \rho(s_{\max} - \sigma)$, with $\rho = 0.50$. Thus, retained patterns must be frequent in the response group and close to the most supported closed patterns in that same group. Finally, we compute *lift* [3] against personas from the same selection strategy and topic that produced a different stance.

### 5.4 RQ3: External Alignment with Pew

RQ3 is an aggregate topic-level comparison between synthetic responses and Pew ATP public-opinion distributions. We retain only Pew items that are Trump-targeted, issue-specific, use supported response scales, and map to exactly one of the five shared topics. Generic approval, broad favorability, trait or affective ratings, knowledge items, and ambiguous multi-topic questions remain only in diagnostic inventories.

For each retained item, respondent answers are weighted, mapped to *positive*, *negative*, *neutral*, or *missing*, and aggregated by topic. Missing and non-substantive answers are excluded from valid denominators. We summarize Pew opinion as $\text{net}(t) = \%\text{positive}_t - \%\text{negative}_t$; because retained Pew items lack true neutral midpoints, Pew neutral mass is zero in the benchmark.

Minimum-temperature RQ1 outputs are mapped into the same response-label space, with *support*, *reject*, and *neutral* as the synthetic counterparts, and we compute topic-level percentages and net scores for bundle- and message-level outputs under all three selection strategies. Alignment is assessed using Pearson and Spearman correlations, mean absolute error, and Pew top-ranked topic recovery.

## 6 Results

### 6.1 RQ1: Theory-Grounded Personas Produce Distinct Reception Profiles

Table 2 reports the main RQ1 results at bundle level. Within each cell, the mean stance of the Top-5 persona group is reported first, followed by that of the corresponding Best persona. Unless otherwise specified, we discuss the Top-5 groups,

[3] Lift is the support of an itemset in the response group divided by its support among same-strategy, same-topic personas with a different stance. Values above 1 indicate that the itemset is more characteristic of the response group.

| | IW-neigh | | | IW-contour | | | MFT | | |
|---|---|---|---|---|---|---|---|---|---|
| Topic $t$ | *high* | *mid* | *low* | *high* | *mid* | *low* | *high* | *mid* | *low* |
| COVID | 0.08/0.65 | 0.08/-0.27 | -0.15/-0.12 | -0.01/0.65 | -0.14/-0.23 | -0.10/-0.04 | 0.61/0.42 | 0.23/0.04 | 0.11/-0.15 |
| Economy | 0.20/0.96 | 0.29/0.75 | -0.56/-0.45 | -0.07/0.97 | 0.56/0.76 | -0.50/-0.53 | 0.82/0.62 | 0.39/0.04 | -0.34/-0.35 |
| Foreign | -0.10/0.05 | -0.26/-0.36 | -0.45/-0.54 | -0.09/0.02 | -0.30/-0.38 | -0.53/-0.51 | 0.52/0.49 | -0.09/-0.30 | -0.33/-0.35 |
| Immigration | -0.08/0.67 | -0.13/-0.48 | -0.74/-1.00 | 0.14/0.72 | -0.27/-0.35 | -0.97/-0.98 | -0.95/-0.94 | -0.10/-0.52 | 0.93/0.80 |
| Judiciary | -0.73/-0.47 | -0.65/-0.63 | -0.69/-0.77 | -0.59/-0.37 | -0.53/-0.63 | -0.67/-0.80 | -0.29/-0.17 | -0.65/-0.87 | -0.81/-0.80 |

Table 2: RQ1 bundle-level mean stance by topic and persona-selection condition. Within each cell, values are reported as Top-5/Best. Topic labels abbreviate `COVID/public health`, `Economy/jobs/trade`, `Foreign policy/security`, `Immigration/border`, and `Judiciary/courts`.

which provide the main audience-level estimate for each topic-conditioned position.

Overall, persona-conditioned reception varies across topic-conditioned positions, but the strength of this separation depends on the topic and the selection strategy. The MFT strategy yields the largest separations. The clearest case is `Immigration/border`: in MFT, the top-5 mean stance shifts from strong rejection for the *high* profile $(-0.95)$ to strong support for the *low* profile $(0.93)$, corresponding to a maximum separation of 1.88 across persona positions.

This is consistent with the MFT mapping (Appendix F, Figure 6): the immigration-*high* target emphasizes Care, whereas the immigration-*low* target emphasizes Loyalty and Authority. The same strategy also separates `Economy/jobs/trade` and `Foreign policy/security`, where *high* profiles are more supportive than *low* profiles. `COVID/public health` shows a weaker but ordered pattern, while `Judiciary/courts` remains broadly negative across positions. Best-persona values generally follow the Top-5 pattern.

The IW strategies produce more conservative separations. IW-neigh shows moderate differences for `Economy/jobs/trade`, `Immigration/border`, and `Foreign policy/security`, but weak or near-zero separation for `COVID/public health` and `Judiciary/courts`. Top-5 groups are often less extreme than their best personas: for example, in `Immigration/border`, the maximum separation in mean stance across persona positions decreases from 1.67 for the Best persona to 0.66 for the Top-5 group. IW-contour often yields a larger maximum separation in mean stance across persona positions than IW-neigh, especially for `Immigration/border` (1.12 vs. 0.66) and `Economy/jobs/trade` (1.06 vs. 0.85).

Appendix H reports the corresponding tweet-level stance diagnostics. The tweet-level stance results are consistent with the bundle-level results, although bundle-level scores are generally more polarized, as expected when the judgment summarizes a coherent topic-matched bundle rather than isolated posts.

### 6.2 RQ2: Persona Dimensions and Response Types

Table 3 reports a sample of frequent closed itemsets associated with bundle-level response types. We focus on non-singleton itemsets since combinations of persona characteristics are more informative than single attributes. A broader version of the table is in Appendix H, Table 14.

The main result is that explanatory structure is sparse and topic-dependent. The strongest multi-item patterns concentrate on `Immigration/border`. Among IW-selected personas, supportive responses combine high Authority and Proportionality, low generalized trust, and tertiary education; the strongest retained pattern has support 0.63 and lift 1.37. Related patterns link high Authority with low social trust, or high Loyalty with high Proportionality, suggesting that support among IW-selected personas also reflects moral orientations toward order, rule enforcement, fairness, and in-group protection.

Among MFT-selected personas, the clearest patterns for the same topic concern rejecting responses: these are associated with urban settlement combined with either postmaterialist orientation or age 30–39, with lift up to 1.17. These cross-theory associations show that, even when personas are selected in moral-foundation space, response types remain partly characterized by cultural and demographic attributes.

Outside `Immigration/border`, associations are weaker. The clearest secondary case is `Judiciary/courts`. Among IW-selected per-

| Strategy $f$ | Topic $t$ | Stance $s$ | Closed frequent itemset $I$ | $\lvert G_{f,t,s}\rvert$ | supp($I$) | Lift |
|---|---|---|---|---|---|---|
| IW-neigh | Immigration | support | education = tertiary; Authority = high; Proportionality = high; social trust = cannot trust people | 200 | 0.63 | 1.37 |
| IW-neigh | Immigration | support | education = tertiary; Authority = high; social trust = cannot trust people | 200 | 0.64 | 1.35 |
| IW-neigh | Immigration | support | Loyalty = high; Proportionality = high | 200 | 0.60 | 1.34 |
| MFT | Immigration | reject | materialism orientation = postmaterialist; settlement = urban | 225 | 0.54 | 1.15 |
| MFT | Immigration | reject | age = 30–39; settlement = urban | 225 | 0.54 | 1.17 |
| IW-neigh | Judiciary | support | education = tertiary; Authority = high | 219 | 0.63 | 1.11 |
| IW-neigh | Judiciary | support | education = tertiary; Authority = high; Proportionality = high | 219 | 0.60 | 1.10 |
| MFT | Judiciary | neutral | education = tertiary; IW Traditional–Secular = secular-rational | 221 | 0.48 | 1.09 |
| MFT | Judiciary | reject | education = tertiary; IW Traditional–Secular = secular-rational | 225 | 0.48 | 1.07 |
| MFT | Judiciary | neutral | education = tertiary; settlement = urban | 221 | 0.58 | 1.06 |
| MFT | COVID | neutral | IW Traditional–Secular = secular-rational; settlement = urban | 220 | 0.51 | 1.02 |
| IW-neigh | Economy | reject | age = 30–39; education = tertiary; Proportionality = high; settlement = urban | 221 | 0.60 | 1.01 |

Table 3: Selected non-singleton closed frequent itemsets associated with RQ2 response types. Support is computed within the ⟨framework strategy, topic, observed response⟩ group. The first block reports the strongest cross-dimensional patterns; subsequent blocks include weaker but interpretable patterns.

sonas, support is associated with tertiary education and high Authority, sometimes also high Proportionality. This is consistent with the institutional character of the topic: support is more characteristic of personas combining formal-education signals with authority-oriented and rule-based moral profiles. Among MFT-selected personas, non-supportive responses are associated with tertiary education and secular-rational IW positioning, with neutral responses also including urban settlement. For `COVID/public health` and `Economy/jobs/trade`, retained non-singleton patterns have lift only slightly above one and should be read as weak diagnostics. No comparable multi-item structure is retained for `Foreign policy/security`.

### 6.3 RQ3: Synthetic Audiences Show Limited but Informative External Alignment

Table 4 reports weak aggregate alignment over the five matched issue domains. The Pew baseline is compressed: only `Economy/jobs/trade` is clearly net-positive ($+8.12$), while `COVID/public health` is the most negative ($-17.02$) and the remaining topics are near balanced. Across synthetic configurations, correlations are near zero and mean absolute errors remain large, at 21.08–29.28 % points. IW-contour is the only condition that recovers Pew's top net-positive topic, `Economy/jobs/trade`, but it still misses the full ranking; MFT at tweet level has the lowest MAE yet not the correct top-ranked topic.

Panel (b) of Table 4 decomposes this result by topic and makes the severity explicit. At message level, MFT reproduces the Pew net score for `Foreign policy/security` to within 0.06 percentage points ($-3.34$ vs. $-3.29$); at bundle level it matches `Immigration/border` to within 3.70 points; and IW-contour preserves the negative direction for `COVID/public health`. `Judiciary/courts`, by contrast, fails severely: Pew is only mildly negative ($-5.71$), whereas every synthetic configuration lies between $-53.16$ and $-68.89$, so even the closest is off by 47.45 points. External alignment is thus a topic-structured failure rather than uniform noise; aggregate correlations average over this heterogeneity.

**(a) Aggregate alignment**

| Strategy | Level | Pearson | Spearman | MAE | Top-1 |
|---|---|---|---|---|---|
| IW-neigh | bundle | -0.003 | -0.100 | 29.14 | no |
| IW-neigh | tweet | -0.089 | -0.100 | 29.28 | no |
| IW-contour | bundle | 0.112 | 0.300 | 26.91 | yes |
| IW-contour | tweet | 0.070 | 0.300 | 26.59 | yes |
| MFT | bundle | 0.015 | -0.100 | 26.33 | no |
| MFT | tweet | -0.022 | -0.100 | 21.08 | no |

**(b) Per-topic net scores**

| Topic $t$ | Pew | Synthetic range | Best $\lvert\Delta\rvert$ |
|---|---|---|---|
| COVID | $-17.02$ | $-8.21$ to $31.54$ | 8.81 |
| Economy | 8.12 | $-4.84$ to $28.38$ | 8.47 |
| Foreign policy | $-3.29$ | $-30.61$ to $3.34$ | 0.06 |
| Immigration | $-0.38$ | $-36.67$ to $-4.08$ | 3.70 |
| Judiciary | $-5.71$ | $-68.89$ to $-53.16$ | 47.45 |

Table 4: Pew–synthetic alignment over topic-level net scores. (a) Aggregate metrics; MAE is in percentage points and Top-1 marks recovery of Pew's most net-positive topic. (b) The underlying per-topic net scores (percentage points): *Synthetic range* spans the six configurations (three selection strategies × bundle and message level), and *Best* $\lvert\Delta\rvert$ is the smallest absolute distance to the Pew net score over those configurations. Table 15 in Appendix H gives the full decomposition.

**Why external alignment fails.** Two mechanisms account for the pattern. The first is a construct gap: the retained Pew items measure confidence in the actor's ability to handle a policy domain, whereas personas judge concrete messages; where the constructs nearly coincide the fit is close, and where they diverge it is not. `Judiciary/courts` is the extreme case, because its bundles are dominated by attacks on courts and judges rather than by judicial policy. The reception labels corroborate this: the topic carries the highest mean offensiveness and the lowest legitimacy and endorsement, yet shows almost no persona separation on any label (Tables 9 and 10), so the response is message-driven rather than persona-driven and cannot recover a mildly negative actor-competence judgment. The second is structural: synthetic outputs place 8–33% of their mass on the neutral category, whereas the retained Pew items have no neutral midpoint. Both are properties of the comparison design rather than of the personas.

## 7 Discussion

### 7.1 What Synthetic Moral Audiences Capture

Culturally grounded synthetic personas capture systematic reception differences, but these differences are not uniform across topics or selection strategies. The clearest case is `Immigration/border`, where care-oriented profiles tend to reject the messages' content, whereas loyalty- and authority-oriented profiles tend to support them. More generally, MFT produces the sharpest separations, while IW-contour often reveals stronger variation than the more local IW-neigh strategy. Conversely, `Judiciary/courts` elicits broadly negative responses across persona groups, suggesting that some message bundles lead otherwise distinct personas toward similar receptions. Where personas do separate, they do so more sharply than the survey baseline, which is consistent with Smith-Vaniz et al. (2025): persona role-play pushes moral responses toward more extreme positions, and in our setting, this surfaces as synthetic net scores far more polarized than Pew.

The strongest patterns are also interpretable, relating to combinations of moral foundations, cultural values, and demographic attributes. Synthetic moral audiences are therefore most useful when the issue itself activates a clear moral or cultural contrast.

### 7.2 What the Pew Benchmark Does and Does Not Validate

The Pew benchmark validates only a limited claim: whether aggregate synthetic reactions preserve some topic-level public-opinion structure across matched issue domains. The benchmark therefore does not validate personas as substitutes for respondents, nor does it equate persona positions with survey groups. Its value is boundary-setting: internal differentiation is not population validity, and synthetic audiences are analytic instruments rather than estimates of the U.S. public.

### 7.3 Implications for Future Audience-Simulation Research

Respondent-to-persona mapping and subgroup-matched benchmarks are usually named as the route to external validity, but they require respondent-level linkage and therefore a different study design. What the present evaluation contributes is the boundary condition such a design must satisfy, together with the three failure modes it has to address: the construct gap between item-level and message-level judgments, the dominance of message tone on topics where personas do not separate, and the neutral-mass mismatch between generated and survey response formats. What remains genuinely open is whether all three can be satisfied at once, and whether the patterns reported here persist across other models and political actors.

## 8 Conclusion

Across Trump message bundles, theory-grounded IW and MFT personas produced systematic but topic-dependent differences in stance and related reception labels, with the strongest separation for `Immigration/border` and sparser explanatory structure elsewhere. The RQ2 analysis shows that some response types can be linked to interpretable combinations of moral, cultural, and demographic attributes, but these patterns are not universal. The Pew benchmark adds complementary boundary evidence: matched topic distributions show a few local correspondences, while also exposing systematic mismatches. Together, the results argue for evaluating synthetic audiences along separate axes of internal differentiation, explanatory structure, and external plausibility.

## Limitations

Persona conditioning should not be interpreted as real respondent identity. The personas provide controlled theoretical profiles for comparing reception patterns, but they are not sampled from a real human-population and cannot substitute for human survey respondents. This caution is consistent with broader concerns about social and representational harms in LLM deployment (Blodgett et al., 2020; Weidinger et al., 2021), and with evidence that models standing in for human participants misportray and flatten identity groups (Wang et al., 2025). Relatedly, our Pew comparison is aggregate and topic-level rather than subgroup-matched. It tests whether synthetic topic distributions resemble weighted public-opinion patterns, not whether a specific synthetic persona corresponds to a specific type of voter or survey respondent.

The corpus of posts used for the evaluation is shaped by upstream and deterministic filtering decisions. We started from posts identified as morally relevant by an MFT scoring model and then applied conservative topic regular expressions, anonymization rules, and prompt-readiness filters. These choices improve auditability and reduce leakage, but they also narrow the set of messages analyzed.

The study focuses on public posts from one political actor, therefore the findings should not be generalized to political communication as a whole without replication on other actors, platforms, and countries. *We also emphasize that the study is about reception, not factual truth or normative correctness.*

We set low-temperature decoding to reduce stochastic instability of the LLM, but it does not eliminate model-specific biases or prompt sensitivity. Since all reception generations are produced with a single LLM, the extent to which the observed patterns generalize across models remains an open question. Different models may exhibit different political defaults, social-desirability tendencies, and degrees of persona adherence; future work should therefore assess the robustness of these findings through cross-model replication. A small number of malformed outputs were skipped, and although the parse-error rate is low, JSON-generation failures remain part of the operational uncertainty of this setup.

RQ2 is descriptive rather than causal. The itemset analysis identifies persona attributes associated with support, rejection, or neutrality in the expanded synthetic sample, but these associations should not be read as universal audience typologies or as causal effects of individual persona dimensions. They are best interpreted as interpretable patterns within this controlled synthetic-audience design.

## Ethical Considerations

The study uses public posts by a prominent political figure and public-opinion data from Pew ATP releases. *We nevertheless anonymized message bundles before prompting by replacing explicit identity, campaign, and user-handle cues where possible.* This is intended to reduce direct speaker-name effects in model judgments and to avoid prompting personas to reason from the hidden identity of the author rather than from message content. Anonymization is not a claim that the original posts are private or impossible to identify; it is a task-control procedure.

The main ethical risk is overinterpreting synthetic persona labels as claims about real groups. Persona labels such as high, mid, and low are topic-conditioned selection positions, not moral evaluations of real people or demographic groups. Claims about public opinion are restricted to the weighted Pew benchmark, and we do not infer the beliefs of demographic or partisan groups from model outputs alone. Moreover, *personas are not proxies for Democrats/Republicans or real voters.*

*We are also committed not to redistribute raw Trump archive files or Pew respondent-level datasets.*

## Acknowledgments

This work was partially supported by the following projects: ELENI: Estrategia Local En Narrativas Inteligente, funded by CDTI (IDI-20260067); Multimodal and Multisensor data-centric AI models (ANNOTATE-MULTI2), grant PID2024-156022OB-C32 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU; Horizon Europe project AI-CODE, GA No. 101135437.

The authors used AI tools for language polishing and development of result-extraction scripts.

## References

José Alemán and Dwayne Woods. 2016. Value orientations from the world values survey: How comparable

are they cross-nationally? *Comparative Political Studies*, 49(8):1039–1067.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T. Stevens, and Morteza Dehghani. 2023. Morality beyond the WEIRD: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 125(5):1157–1188.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.

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.

Julien Boelaert, Samuel Coavoux, Étienne Ollion, Ivaylo Petev, and Patrick Präg. 2025. Machine bias. how do generative language models answer opinion polls? *Sociological Methods & Research*, 54(3):1156–1196.

Alexander Brady and Tunazzina Islam. 2025. Iterative topic taxonomy induction with LLMs: A case study of electoral advertising. *Preprint*, arXiv:2510.15125.

William J. Brady, Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

Dennis Chong and James N. Druckman. 2007. Framing theory. *Annual Review of Political Science*, 10:103–126.

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

James L. Gibson, Gregory A. Caldeira, and Lester Kenyatta Spence. 2003. Measuring attitudes toward the United States Supreme Court. *American Journal of Political Science*, 47(2):354–367.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In Patricia Devine and Ashby Plant, editors, *Advances in Experimental Social Psychology*, volume 47, pages 55–130. Academic Press.

Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046.

Jesse Graham, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H. Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2):366–385.

Candida M. Greco, Lucio La Cava, and Andrea Tagarelli. 2026. Culturally grounded personas in large language models: Characterization and alignment with socio-psychological value frameworks. *Preprint*, arXiv:2601.22396.

Kobi Hackenburg, William J Brady, and Manos Tsakiris. 2023. Mapping moral language on us presidential primary campaigns reveals rhetorical networks of political division and unity. *PNAS Nexus*, 2(6):pgad189.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Heads or Tails. 2021. Trump twitter archive. Kaggle dataset. Accessed: 2026-05-13.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, Gabriela Moreno, Christina Park, Tingyee E. Chang, Jenna Chin, Christian Leong, Jun Yen Leung, Arineh Mirinjian, and Morteza Dehghani. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Ronald Inglehart and Wayne E. Baker. 2000. Modernization, cultural change, and the persistence of traditional values. *American Sociological Review*, 65(1):19–51.

Ronald Inglehart and Christian Welzel. 2005. *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. Cambridge University Press.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730, Melbourne, Australia. Association for Computational Linguistics.

Kristen Johnson and Dan Goldwasser. 2019. Modeling behavioral aspects of social media discourse for moral classification. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 100–109, Minneapolis, Minnesota. Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys*, 53(1):1–37.

Sanguk Lee, Tai-Quan Peng, Matthew H. Goldberg, Seth A. Rosenthal, John E. Kotcher, Edward W. Maibach, and Anthony Leiserowitz. 2024. Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8):1–14.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Mohsen Mosleh, Gordon Pennycook, and David G. Rand. 2020. Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *PLOS ONE*, 15(2):e0228882.

Diana C. Mutz and Byron Reeves. 2005. The new videomalaise: Effects of televised incivility on political trust. *American Political Science Review*, 99(1):1–15.

Brian L. Ott. 2017. The age of twitter: Donald j. trump and the politics of debasement. *Critical Studies in Media Communication*, 34(1):59–68.

Pew Research Center. 2026. The american trends panel. Accessed: 2026-05-13.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1095.

Shamik Roy, Maria Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9939–9958, Online and Punta Cana, Dominican Republic. 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*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Nicole Smith-Vaniz, Harper Lyon, Lorraine Steigner, Ben Armstrong, and Nicholas Mattei. 2025. Investigating political and demographic associations in large language models through moral foundations theory. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3):2419–2430.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21147–21170, Vienna, Austria. Association for Computational Linguistics.

The Trump Archive. 2026. The trump archive. Accessed: 2026-05-13.

Tom R. Tyler. 2006. Psychological perspectives on legitimacy and legitimation. *Annual Review of Psychology*, 57:375–400.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411.

Sze-Yuh Nina Wang and Yoel Inbar. 2021. Moral-language use by u.s. political elites. *Psychological Science*, 32(1):14–26.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, and 4 others. 2021. Ethical and social risks of harm from language models. *Preprint*, arXiv:2112.04359.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Lorenzo Zangari, Candida M. Greco, Davide Picca, and Andrea Tagarelli. 2025. ME2-BERT: Are events and emotions what you need for moral foundation prediction? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9516–9532, Abu Dhabi, UAE. Association for Computational Linguistics.

## A Research Question Overview

Table 5 summarizes how the three research questions differ in the data they use, the variables they

analyze, and the procedures applied to them. RQ1 and RQ2 both operate on synthetic responses alone, but at different persona scales: RQ1 compares the Top-5 persona groups selected for each topic-conditioned position, whereas RQ2 relies on the expanded IW-neigh and MFT samples required to make itemset support estimates stable. RQ3 is the only layer that introduces an external reference, pairing the RQ1 responses with the matched Pew ATP items. The table is intended as a compact reference for Sections 4–5, which describe each component in detail.

## B Prompt Template and Output Schema

Figure 3 reports the prompt template used to obtain persona-conditioned reactions to political-message bundles. The prompt is designed to separate two levels of evaluation: the persona's assessment of the underlying substantive position expressed by the message, and the persona's reaction to the rhetorical style of the message. This distinction is necessary because a persona may agree with the policy direction of a message while rejecting its tone, or may disagree with the substantive claim without finding the message especially offensive.

Each request contains four components: (i) the persona profile, including either IW cultural-positioning information or MFT moral-profile information; (ii) the institutional role of the author at the time the messages were written; (iii) a topic-specific bundle of anonymized political messages, each with an identifier, date, role, and text; and (iv) the required JSON output schema. The prompt instructs the model not to infer the hidden identity of the author and to base judgments only on the persona profile, message content, and author role.

The placeholder `{RULES FOR DISTINGUISHING...}` in Figure 3 summarizes the operational instructions used to separate substantive evaluation from rhetorical-style evaluation. In the full prompt, these rules require the model to first infer the most plausible topic-related claim or policy direction expressed by each message, and to base *stance* and *agreement* primarily on that content. By contrast, *offensiveness* is reserved for tone, hostility, insults, or disrespectful wording, while *endorsement* captures whether the persona would share the message in its exact written form. The rules also constrain label consistency: supportive stances should normally correspond to higher agreement, rejecting stances to lower agreement, and neutral stances to mixed, weak, or ambiguous substantive content.

The response schema contains both message-level and bundle-level judgments. The output format is summarized in Figure 2.

## C Corpus Construction and Topic Filtering

The post corpus is built from two Trump archive inputs: the 2009–2021 Kaggle Trump Twitter Archive (Heads or Tails, 2021) and a 2022–2024 manually prepared extension from The Trump Archive (The Trump Archive, 2026). The pipeline starts from posts marked as morally relevant by ME2-BERT (Zangari et al., 2025) and then applies deterministic preprocessing rules. These rules remove retweets, decode HTML entities, remove URLs, normalize whitespace, discard very short or low-information posts, and assign an author-role label from the posting date. Deleted or flagged posts are retained in a moderation-status archive but are excluded from the prompt-ready pool when the source explicitly provides that status. Posts from sources without deletion or flag metadata are marked as unknown rather than excluded solely for missing metadata.

Topic assignment uses a shared high-precision keyword registry for both post preprocessing and Pew selection. Each post is assigned to a topic only when exactly one issue-topic pattern matches. Posts with no topic hit are assigned to `other_campaign_generic`; posts with multiple topic hits are assigned to `review_needed` and excluded from prompt-ready prompting unless manually overridden. The canonical topics are `Immigration/border`, `Economy/jobs/trade`, `election_integrity_democracy`, `Foreign policy/security`, `crime_policing_criminal_justice`, `covid_public_health`, and `Judiciary/courts`. The keyword policy favors precision and auditability over maximal recall.

Anonymization is applied after topic labeling. The procedure replaces explicit references to the political actor with `[POLITICAL_ACTOR]`, campaign slogans with `[CAMPAIGN_TAG]`, and remaining user handles with `[USER]`, while preserving selected institutional names where they are needed for content interpretation. Prompt-ready posts are then checked for remaining identity leaks, anonymization degradation, topic ambiguity, context depen-

| RQ | Data / sample | Variables | Analysis / metrics |
|---|---|---|---|
| RQ1 | Top-5 personas from IW-neigh, IW-contour, and MFT for each topic-conditioned position, evaluated on topic-specific message bundles. | Primary outcome: stance. Additional collected ratings: agreement, legitimacy, offensiveness, and endorsement. | Mean stance at bundle and message level; profiles separation; Best vs. Top-5 comparison. |
| RQ2 | Expanded IW-neigh (100 personas) and MFT (80 personas) samples for each topic-position pair, queried with reduced persona profiles. | Observed bundle-level stance; demographic attributes; WVS-derived cultural variables; cross-theory MFT or IW attributes. | Closed frequent itemsets; support-based within-group filtering; lift against personas with a different stance. |
| RQ3 | RQ1 synthetic responses and 25 Pew ATP issue-specific items across the five matched topics. | Synthetic stance distributions and weighted Pew positive/negative response distributions. | Topic-level net scores; Pearson and Spearman correlations; mean absolute error; top-ranked topic recovery. |

Table 5: Mapping of each research question to the corresponding data, variables, and analytical procedures.

```
{
  "bundle_id": "<topic>_bundle_<id>",
  "responses": [
    {
      "tweet_id": "<message_id>",
      "stance": "support|neutral|reject",
      "agreement": 1,
      "legitimacy": 1,
      "offensiveness": 1,
      "endorsement": 1,
      "emotional_reaction": "short text",
      "brief_reason": "short text"
    }
  ],
  "bundle_overall_stance": "support|neutral|reject",
  "bundle_overall_agreement": 1,
  "bundle_overall_legitimacy": 1,
  "bundle_overall_offensiveness": 1,
  "bundle_overall_endorsement": 1,
  "bundle_overall_emotional_reaction": "short text",
  "bundle_overall_reason": "short text"
}
```

Figure 2: JSON output schema for persona-conditioned message-bundle reactions. Each message receives structured judgments over stance, agreement, legitimacy, offensiveness, endorsement, emotional reaction, and brief reason. The same labels are also collected at bundle level.

dence, and duplicate anonymized text.

The preprocessing stage in this work produces 23,230 cleaned posts after the initial moral-relevance and quality filters. Of these, 5,270 are prompt-ready across the seven canonical topics. The main experiments use the five topics that overlap with the Pew benchmark: `COVID/public health` (251 posts), `Economy/jobs/trade` (1,116), `Foreign policy/security` (1,112), `Immigration/border` (539), and `Judiciary/courts` (300).

The preprocessing rule list, topic registry, and anonymization rules are retained as audit artifacts for reproducibility.

## D Pew Matching Protocol

The Pew benchmark is constructed from Pew Research Center American Trends Panel wave folders (Pew Research Center, 2026). For each wave, we extract wave metadata, field dates, dataset filenames, candidate variable names, and question text from available documentation and `.sav` metadata. The merged inventory used here contains 981 candidate rows.

Each candidate row is then passed through a deterministic selector. A Pew item is eligible for RQ3 only if it targets Trump directly or through a constrained presidential context, uses a supported judgment family, and matches exactly one canonical issue topic. Supported judgment families are approval, confidence, and policy support. The selector excludes thermometer ratings, personal-

Persona Reaction Prompt Template

SYSTEM MESSAGE:

*You must role-play the person described in the provided persona profile. Adopt that person's values, dispositions, trust level, moral intuitions, and likely judgments as consistently as possible. You will evaluate a set of anonymized political messages. For each message, evaluate it on two distinct layers: (A) the underlying topic-related position, policy direction, or moral claim expressed by the message; (B) the rhetorical style, wording, hostility, and civility of the message. Do not collapse layer A into layer B.*
{RULES FOR DISTINGUISHING SUBSTANTIVE POLITICAL EVALUATION FROM RHETORICAL-STYLE EVALUATION AND FOR ENSURING COHERENT STANCE, AGREEMENT, OFFENSIVENESS, LEGITIMACY, AND ENDORSEMENT JUDGMENTS}.
*Base your evaluation only on: (1) the persona profile, (2) the message content, and (3) the institutional role of the message author at the time it was written. Do not use outside knowledge about any real politician, party, or historical event beyond what is explicitly stated. Return only valid JSON and no extra text.*

USER MESSAGE:

*Your task is to evaluate each message exactly as this persona profile would likely evaluate it. Use the persona profile consistently across all judgments.*

[*RESPONSE SCALE*]
*Use the following scales exactly:*
*- stance: one of ["support", "neutral", "reject"], referring to the underlying topic-related position or policy direction, not to tone alone*
*- agreement: 1..5 where 1 strongly disagree and 5 strongly agree with the underlying substantive claim or direction of the message*
*- legitimacy: 1..5 where 1 = completely illegitimate and 5 = completely legitimate from the persona's point of view; do not use this as a generic external fact-check score*
*- offensiveness: 1..5 where 1 = not offensive at all and 5 = extremely offensive in wording, hostility, or disrespect*
*- endorsement: 1..5 where 1 = would definitely not endorse/share and 5 = would definitely endorse/share the full message as written, considering both substance and tone*
*- emotional_reaction: short string, max 20 words*
*- brief_reason: short string, max 60 words; mention both substance and style when relevant*
*- After evaluating individual messages, also provide an overall evaluation of the whole bundle using the same logic.*

[*PERSONA PROFILE*]
{*Full persona profile, including demographic metadata, short biography, cultural-variable mapping, and cultural-positioning score in the IW map or MFQ-2 responses.* }

[*CONTEXT*]
*At the time these messages were written, the author was:* {*author institutional role* }

[*MESSAGES*]
{*Anonymized political messages with ID, date, role, and text* }

[*OUTPUT JSON SCHEMA*]
{*JSON output structure for message-level and bundle-level evaluations* }

Figure 3: Persona reaction prompt template used to obtain structured message-level and bundle-level judgments.

ity or trait batteries, affective reaction questions, knowledge or awareness items, broad favorability, generic presidential job approval, unsupported judgment forms, no-topic matches, and ambiguous multi-topic matches. Broad Trump job-approval questions are kept in the audit inventory but excluded from RQ3 because they measure global evaluation rather than issue-specific opinion.

This selection leaves 25 Pew rows across five topics and seven ATP waves. All retained rows are issue-specific confidence items about Trump's ability to handle a policy domain. The final Pew subset contains five COVID/public-health rows, seven `Economy/jobs/trade` rows, five foreign-policy/national-security rows, four `Immigration/border` rows, and four `Judiciary/courts` rows.

For each retained item, respondent-level answers are extracted from the corresponding `.sav` file and weighted using the wave weight vari-

able. Response labels are mapped into `positive`, `negative`, `neutral`, and `missing` buckets. For confidence scales, *very confident* and *somewhat confident* are positive, while *not too confident* and *not at all confident* are negative. Missing, refused, don't-know, blank, and non-substantive values are excluded from the valid denominator but retained in diagnostics. The selected Pew items do not contain a true neutral midpoint, so the Pew neutral bucket is zero in the main RQ3 benchmark.

Topic-level Pew distributions are computed by pooling weighted positive, negative, and neutral counts across retained questions within each topic and then calculating valid percentages. The main comparison statistic is the net score, defined as positive percentage minus negative percentage. Each included or excluded Pew row receives an exclusion code and a rule trace for auditability, and the full selection audit is retained as part of the reproducibility materials.

## E Persona Mapping Rationale and MFT Profiles

We define the persona-selection space by mapping each topic to a restricted set of theoretically relevant Inglehart–Welzel dimensions and Moral Foundations Theory dimensions. This mapping does not assign a single "correct" cultural or moral interpretation to a topic. Rather, it identifies the cultural-value dimensions and moral foundations that are theoretically most likely to organize contrasting audience interpretations of each topic. These dimensions are then used by the selection procedure in Appendix F to construct *high*, *mid*, and *low* persona groups for each topic. Table 6 summarizes the topic-level mapping, reporting which dimensions are treated as relevant for each topic.

**Inglehart–Welzel mapping.** Since the IW space is defined by two bipolar axes, a topic can be associated with more than one dimension when different audience interpretations plausibly emphasize different poles of the same axis. *The mapping is therefore used to define the relevant selection directions in the IW space, not to claim that all messages in a topic express the same cultural orientation.* The following list gives the rationale for each topic-to-IW association:

- `COVID/public health`: **Survival** is relevant because the topic concerns health risk, physical protection, and collective security. **Secular-rational** is relevant because the topic is often organized around science, expertise, institutions, and public policy.
- `Economy/jobs/trade`: **Survival** is relevant because the topic concerns material security, employment, income, prices, trade protection, and economic stability. **Self-expression** is relevant when economic claims are framed in terms of individual autonomy, opportunity, entrepreneurship, quality of work, or the capacity to pursue a preferred way of life beyond basic material security.
- `Foreign policy/security`: **Survival** is relevant because the topic involves threat perception, war, defense, military protection, and national security. **Traditional** is relevant because foreign-policy rhetoric often invokes duty, loyalty to the national community, respect for order, collective continuity, and the defense of established institutions or alliances.
- `Immigration/border`: **Survival** is relevant because immigration and border politics are frequently framed through security, control, resource pressure, and perceived social threat. **Traditional** is relevant when the issue is interpreted through national identity, borders, cultural cohesion, and social order. **Self-expression** is also relevant because contrasting interpretations may emphasize openness, inclusion, minority rights, pluralism, and tolerance toward outgroups.
- `Judiciary/courts`: **Secular-rational** is relevant because courts are associated with legal procedure, institutional rationality, impersonal rules, and rule-of-law reasoning. **Traditional** is relevant because judicial authority can also be interpreted through respect for order, hierarchy, legitimacy of established institutions, and obedience to socially recognized authority.

**Moral Foundations mapping.** For the MFT representation, we map each topic to the moral foundations that are most likely to structure audience judgments. As with the IW mapping, *this does not mean that every message in a topic expresses all of these foundations, nor that other foundations are irrelevant*. Rather, the selected foundations define the dimensions along which *high*, *mid*, and *low*

| Topic | IW dimensions | MFT foundations |
|---|---|---|
| COVID/public health | Survival; Secular-rational | Care; Authority; Purity |
| Economy/jobs/trade | Survival; Self-expression | Proportionality; Equality; Care |
| Foreign policy/security | Survival; Traditional | Loyalty; Authority; Care |
| Immigration/border | Survival; Traditional; Self-expression | Loyalty; Authority; Care |
| Judiciary/courts | Secular-rational; Traditional | Authority; Care |

Table 6: Topic-level mapping used for persona selection. IW dimensions specify the cultural-value axes used for topic-conditioned selection; MFT foundations specify the moral dimensions used to define topic-specific target profiles.

moral target profiles are constructed. The following list gives the rationale for each topic-to-MFT association:

- COVID/public health: **Care** is relevant because the topic centrally concerns suffering, vulnerability, harm prevention, and the protection of fragile individuals. **Authority** is relevant because public-health discourse often involves compliance with rules, expert guidance, institutional trust, and coordinated policy. **Purity** is relevant because the topic can involve contagion, contamination, hygiene, bodily integrity.

- Economy/jobs/trade: **Proportionality** is relevant because economic discourse often invokes merit, contribution, deserved reward, productivity, and fair exchange. **Equality** is relevant when the issue is framed around redistribution, wages, opportunity, inequality, or unequal access to economic security. **Care** is relevant when the frame emphasizes poverty, precarity, unemployment and social harm.

- Foreign policy/security: **Loyalty** is relevant because the topic often invokes national belonging, alliances, sacrifice for the group, patriotism, and defense of the national interest. **Authority** is relevant because foreign policy and security discourse involve state power, command, hierarchy, military institutions, and legitimate use of force. **Care** is relevant when the frame emphasizes victims, civilian suffering, humanitarian protection, or the prevention of harm.

- Immigration/border: **Loyalty** is relevant when immigration is framed through in-group protection, national community, belonging, and the boundary between "us" and "them." **Authority** is relevant when the issue is framed around legality, border control, enforcement, and respect for rules. **Care** is relevant when the discourse emphasizes refugees, family separation, vulnerability, humanitarian protection, or harm to migrants.

- Judiciary/courts: **Authority** is relevant because courts and rulings embody institutional legitimacy, legal order, procedural authority, and respect for binding decisions. **Care** is relevant when legal controversies are interpreted through harm, victim protection, or the consequences of judicial decisions for vulnerable groups.

# F Persona Selection Details

## F.1 Inglehart–Welzel selection

In the IW condition, each persona $p$ is represented by its coordinates $(x_p, y_p)$, where $x_p$ denotes the Survival–Self-expression axis and $y_p$ denotes the Traditional–Secular-rational axis. We standardize both axes over the full persona pool and refer to the standardized IW position of persona $p$ as $\mathbf{z}_p$.

For each topic-position pair $\langle t, pos \rangle$, we first select an anchor persona $a_{t,c}$. Let $\mathbf{u}_{t,c}$ be the unit vector corresponding to the topic-specific IW direction defined by Table 6.

**High-low positions.** For *high* and *low* positions, personas are ranked with a directional IW score:

$$s^{IW}_{t,c}(p) = \mathbf{z}_p^\top \mathbf{u}_{t,c} + \gamma \|\mathbf{z}_p\|_2 + \\ -\lambda \left\| \mathbf{z}_p - (\mathbf{z}_p^\top \mathbf{u}_{t,c}) \mathbf{u}_{t,c} \right\|_2 .$$

The dot product $\mathbf{z}_p^\top \mathbf{u}_{t,c}$ measures how strongly persona $p$ is aligned with the topic-specific IW direction. The Euclidean norm $\|\mathbf{z}_p\|_2$ measures how far the persona is from the center of the IW map. The final norm term measures lateral deviation, namely how far the persona lies from the selected topic direction rather than along it. We use fixed selection hyperparameters $\lambda = 0.22$ and $\gamma = 0.18$.

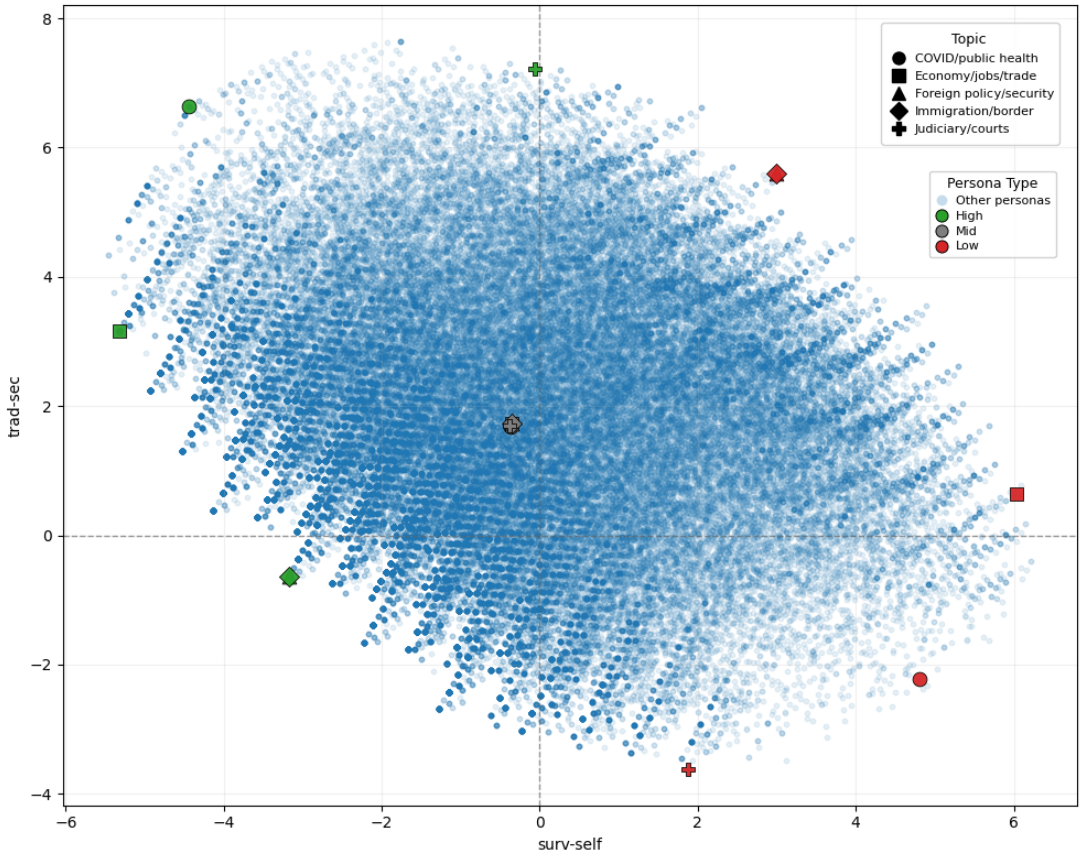


Figure 4: Inglehart–Welzel anchor personas selected for each topic-position pair. Background points show the full persona pool in the IW space; highlighted markers indicate the highest-scoring anchor persona for each topic and position. Colors denote topic-conditioned positions and marker shapes denote topics.

These values are chosen to keep directional alignment as the dominant criterion while softly discouraging lateral outliers and mildly favoring peripheral profiles.

The anchor for each *high* or *low* topic-position pair is then defined as the highest-scoring persona:

$$a_{t,c} = \arg\max_{p \in \mathcal{P}} s^{IW}_{t,c}(p), \qquad c \in \{high, low\},$$

where $\mathcal{P}$ denotes the full persona pool.

**Mid positions.** *Mid*-positioned anchor personas are intended to provide a non-polarized contrast, and are therefore ranked by

$$s^{IW}_{t,c}(p) = -\left(\|\mathbf{z}_p\|_2 + \alpha|\mathbf{z}_p^\top \mathbf{u}_{t,high}|\right), \; c \in \{mid\}$$

with $\alpha = 0.8$.

The Euclidean norm penalizes personas far from the center of the IW space, while the second term measures the magnitude of the projection on the *high* topic-specific direction $\mathbf{u}_{t,high}$. We use the absolute projection because a *mid* profile should be weakly aligned with both the *high* direction and its opposite. The minus sign makes higher scores correspond to more central and less topic-aligned profiles. The *mid* anchor is analogously defined as

$$a_{t,c} = \arg\max_{p \in \mathcal{P}} s^{IW}_{t,c}(p), \; c \in \{mid\}.$$

**Inglehart–Welzel personas.** Figure 4 shows the resulting IW anchor personas for all topic-position pairs before they are expanded into top-5 audiences.

The anchor identifies the single best-matching persona for a given topic-position pair. However, using only one profile would make the audience representation too sensitive to a single generated persona. We therefore expand each anchor into a top-5 audience using two selection variants (Figure 5) that differ in how the remaining four personas are chosen: the neighbor-based variant (IW-neigh) prioritizes local similarity to the anchor, while the contour-based variant (IW-contour) prioritizes broader coverage of the same topic-relevant IW region.

The IW-neigh condition expands each anchor

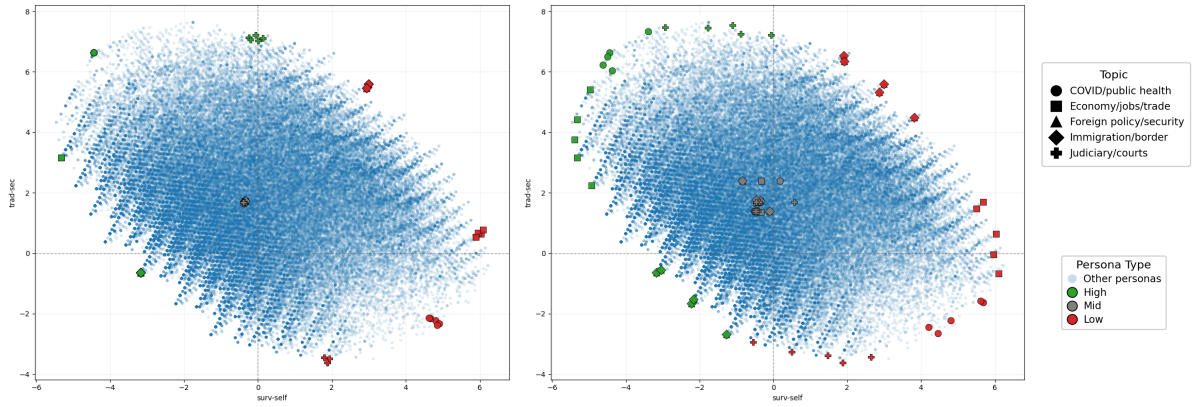


Figure 5: Inglehart–Welzel top-5 persona-selection variants: neighbor-based (Left) and contour-based (Right). Background points show the full persona pool in the IW space; highlighted markers show selected personas, with colors denoting topic-conditioned positions and marker shapes denoting topics.

with four local neighbors. We fit a $K$-means partition of the full standardized IW space with $K = 80$. For each anchor $a_{t,c}$, we identify its cluster and select the four personas in the same cluster with smallest Euclidean distance to the anchor.

In IW-contour, we use the same topic-specific ranking but divide the IW map into cells and select high-ranked personas from distinct cells along the topic-relevant contour. This variant therefore represents a less local audience than the neighbor-based selection. Together, the two IW variants let us compare reception patterns around the best-profile neighborhood with patterns obtained from a wider topic-relevant cultural contour.

### F.2 Moral Foundations selection

In the MFT condition, persona profiles are selected from the same underlying pool, but the selection is performed in the six-dimensional MFT space. Each persona $p$ is represented by a vector $\mathbf{m}_p \in \mathbb{R}^6$. We denote by $\mathbf{m}_{p,j}$ the score of persona $p$ on foundation $j \in \mathcal{M}$, where $\mathcal{M}$ contains the six moral foundations: Care, Equality, Proportionality, Loyalty, Authority, and Purity.

For each topic-position pair $\langle t, pos \rangle$, we define a target profile $\tau_{t,c}$ over the same six dimensions. The core dimensions of each target are given by the topic-to-foundation mapping in Table 6. For *high* and *low* positions, these core foundations are assigned high or low target values according to the intended topic-specific moral contrast. For instance, `COVID/public health` *high* profiles target high Care, Authority, and Purity, whereas the corresponding *low* profiles target low values on the same dimensions. Non-core foundations are targeted toward the neutral midpoint of the MFT scale with lower weight, so that selected personas possibly match the intended moral contrast without becoming unnecessarily extreme on foundations that are not central to the topic.

Personas are then ranked by weighted Euclidean distance from the target profile:

$$d_{t,c}^{MFT}(p) = \sqrt{\sum_{j \in \mathcal{M}} w_{t,c,j} \left(\mathbf{m}_{p,j} - \tau_{t,c,j}\right)^2},$$

where $w_{t,c,j}$ is the weight assigned to foundation $j$ for topic $t$ and position $pos$, which is defined as follows: core foundations receive weight $1.0$, while non-core foundations are assigned the neutral target value $3.0$ with lower weight $0.50$, so that they regularize the selection without dominating the topic-specific moral contrast. For *mid* profiles, all six foundations are targeted to $3.0$ and weighted equally.

For each $\langle t, pos \rangle$ pair, we retain the five personas with the smallest $d_{t,c}^{MFT}$. The closest persona is used as the corresponding MFT best profile. Figure 6 shows the corresponding best profiles for each topic. Table 7 summarizes the MFT scores of the selected top-5 personas, showing that the retained profiles instantiate the intended topic-conditioned contrasts.

## G Expanded Persona Selection for RQ2

RQ2 requires substantially more within-group variation than the RQ1 top-5 setup, because the goal is to associate response types with persona-level attributes. We therefore construct an expanded persona sample. The original top-5 personas remain fixed, and additional personas are selected to

| Position | Care | Equality | Proportionality | Loyalty | Authority | Purity |
|---|---|---|---|---|---|---|
| `COVID/public health` | | | | | | |
| *High* | $5.00 \pm 0.00$ | $2.93 \pm 0.09$ | $3.17 \pm 0.17$ | $2.77 \pm 0.15$ | $4.93 \pm 0.09$ | $4.93 \pm 0.09$ |
| *Mid* | $3.10 \pm 0.09$ | $2.73 \pm 0.25$ | $3.23 \pm 0.15$ | $2.83 \pm 0.39$ | $3.13 \pm 0.22$ | $2.77 \pm 0.19$ |
| *Low* | $1.63 \pm 0.43$ | $2.90 \pm 1.28$ | $3.23 \pm 0.49$ | $2.40 \pm 0.69$ | $1.87 \pm 0.18$ | $1.73 \pm 0.42$ |
| `Economy/jobs/trade` | | | | | | |
| *High* | $1.10 \pm 0.22$ | $1.17 \pm 0.12$ | $4.86 \pm 0.08$ | $2.80 \pm 0.22$ | $3.20 \pm 0.32$ | $2.30 \pm 0.25$ |
| *Mid* | $3.10 \pm 0.09$ | $2.73 \pm 0.25$ | $3.23 \pm 0.15$ | $2.83 \pm 0.39$ | $3.13 \pm 0.22$ | $2.77 \pm 0.19$ |
| *Low* | $5.00 \pm 0.00$ | $4.97 \pm 0.08$ | $2.03 \pm 0.18$ | $3.03 \pm 0.48$ | $2.94 \pm 0.25$ | $3.03 \pm 0.38$ |
| `Foreign policy/security` | | | | | | |
| *High* | $1.83 \pm 0.47$ | $2.50 \pm 0.86$ | $3.63 \pm 0.27$ | $4.77 \pm 0.19$ | $4.90 \pm 0.15$ | $3.53 \pm 0.43$ |
| *Mid* | $3.10 \pm 0.09$ | $2.73 \pm 0.25$ | $3.23 \pm 0.15$ | $2.83 \pm 0.39$ | $3.13 \pm 0.22$ | $2.77 \pm 0.19$ |
| *Low* | $5.00 \pm 0.00$ | $3.33 \pm 0.43$ | $3.40 \pm 0.51$ | $1.80 \pm 0.22$ | $1.73 \pm 0.25$ | $2.47 \pm 0.51$ |
| `Immigration/border` | | | | | | |
| *High* | $5.00 \pm 0.00$ | $3.33 \pm 0.43$ | $3.40 \pm 0.51$ | $1.80 \pm 0.22$ | $1.73 \pm 0.25$ | $2.47 \pm 0.51$ |
| *Mid* | $3.10 \pm 0.09$ | $2.73 \pm 0.25$ | $3.23 \pm 0.15$ | $2.83 \pm 0.39$ | $3.13 \pm 0.22$ | $2.77 \pm 0.19$ |
| *Low* | $1.83 \pm 0.47$ | $2.50 \pm 0.86$ | $3.63 \pm 0.27$ | $4.77 \pm 0.19$ | $4.90 \pm 0.15$ | $3.53 \pm 0.43$ |
| `Judiciary/courts` | | | | | | |
| *High* | $1.80 \pm 0.36$ | $2.20 \pm 0.36$ | $3.83 \pm 0.24$ | $2.77 \pm 0.34$ | $4.90 \pm 0.15$ | $2.97 \pm 0.22$ |
| *Mid* | $3.10 \pm 0.09$ | $2.73 \pm 0.25$ | $3.23 \pm 0.15$ | $2.83 \pm 0.39$ | $3.13 \pm 0.22$ | $2.77 \pm 0.19$ |
| *Low* | $5.00 \pm 0.00$ | $3.30 \pm 0.34$ | $3.50 \pm 0.26$ | $2.63 \pm 0.36$ | $1.60 \pm 0.19$ | $2.57 \pm 0.19$ |

Table 7: MFT scores of top-5 selected personas by topic-conditioned position. Values are reported as mean $\pm$ standard deviation. Cell shading is proportional to the mean score: lighter gray indicates lower values, darker gray indicates higher values.

stay close to the corresponding topic-position profile while increasing demographic diversity. More specifically, each candidate is assigned a representative demographic signature combining: demographic triple $(\text{continent}, \text{settlement}, \text{education})$, age band, gender, occupation group, and country group. The signature space is restricted to the most frequent values in the persona pool (Greco et al., 2026):

- five demographic triples, covering about 78% of the pool: ("Europe", "Urban", "Tertiary"), ("North America", "Urban", "Tertiary"), ("Asia", "Urban", "Tertiary"), ("South America", "Urban", "Tertiary"), ("North America", "Rural", "Tertiary").
- two age bands, covering about 98%:"30-39", "40-49". Together, these bands cover about 98% of the full persona pool.
- two gender categories, covering about 97%: "Man", "Woman".
- five occupation groups, covering about 62%: "Education & Academia", "Social Services & Nonprofit", "Sales & Marketing", "Tech / IT / Data", "Management / Operations / Admin"
- twenty country groups, covering about 75%: "United States", "Poland", "Mexico", "Germany", "Netherlands", "Philippines", "Brazil", "Sweden", "India", "Canada", "Spain", "Italy", "Turkey", "Czechia", "United Kingdom", "Saudi Arabia", "Pakistan", "Portugal", "Russia", "Indonesia".

**IW expansion.** For the IW condition, we expand only the main neighbor-based selection (IW-neigh), since this variant provides a local anchor neighborhood around the best-matching IW profile. For each topic-position pair $\langle t, pos \rangle$, we keep the original top-5 personas and then add nearby candidates from the full persona pool until reaching 100 personas.

Added IW personas are selected in order of Euclidean distance to the anchor, but only if they introduce a new representative demographic signature within the same topic-position cell. This yields a top-100 IW expanded sample that remains local to the original IW anchor while broadening demographic coverage. Figure 7 shows the resulting expanded samples in the Inglehart–Welzel space.

**MFT expansion.** For the MFT condition, the final RQ2 sample uses the top-80 selection. For each

topic-position group $\langle t, pos \rangle$, we keep the original top-5 personas and add 75 additional personas, yielding 80 personas per group. Candidates are first ranked by weighted Euclidean distance to the corresponding topic-specific MFT target profile in the six-dimensional MFT space. Thus, the primary selection criterion is proximity in moral-foundation scores.

We then apply two filters to the added candidates. First, we use a consistency filter to avoid adding personas whose moral profiles are close in overall distance but inconsistent with the intended direction of the target on the relevant foundations.

Second, we impose a demographic-diversity constraint. Figure 8 reports boxplots for these final top-80 MFT samples. We can see that the expanded samples preserve the main target-relevant contrasts most clearly for binding foundations such as Loyalty, Authority, and Purity, while exact target matching is more limited for Care, which is high across much of the underlying persona pool. Thus, the top-80 selection should be interpreted as a feasible proximity-based expansion rather than an exact replication of the target moral profiles.

## H Additional Results and Diagnostics

This appendix reports additional results and diagnostics supporting the main analyses. For RQ1, Table 8 expands the stance analysis by separating tweet- and bundle-level results and comparing the best persona with the corresponding top-5 persona group.

Table 9 reports the four scalar labels at bundle level. Agreement, legitimacy, and endorsement decrease across topics as offensiveness increases, and the high-minus-low contrasts move with stance within every condition, positively for agreement, legitimacy, and endorsement and negatively for offensiveness (Table 10). The label set therefore behaves as the design intends, with offensiveness oriented as a rejection signal rather than as a second support score. The message-level means preserve the same topic ordering with less polarized values (Table 11).

Level and separation nevertheless dissociate across topics. `Judiciary/courts` combines the highest mean offensiveness (3.49) with the lowest agreement (1.68), legitimacy (1.80), and endorsement (1.32), yet shows almost no persona separation on any label: under IW-neigh every high-minus-low contrast falls within $\pm 0.11$. Rejection is also topic-dependent in tone, not only in substance: mean offensiveness among rejecting responses rises from 2.70 for `COVID/public health` and 2.80 for `Economy/jobs/trade` to 3.69 for `Immigration/border` and 3.86 for `Judiciary/courts` (Table 12). This ordering holds within each selection strategy taken separately, so it is not an artifact of pooling. `Judiciary/courts` thus combines weak stance differentiation with strongly patterned rhetorical rejection, a message-driven rather than persona-driven response, which Section 6.3 draws on to explain the RQ3 mismatch.

Table 13 summarizes skipped malformed generations; the total parse-error rate is low.

Table 14 expands the RQ2 frequent closed itemset analysis reported in the main text. It excludes singleton itemsets and reports representative multi-item patterns across topics, together with support within the relevant response group and lift against personas from the same strategy and topic that produced a different response.

Table 15 gives the topic-level Pew and synthetic net scores behind the aggregate RQ3 metrics in Table 4. Finally, Table 16 provides qualitative examples from IW-neigh best-profile outputs. Each row reports one bundle-level response for a topic-position pair, including the observed stance, emotional reaction, and brief reason.

| Topic | Condition | Level | *High* Best | *High* Top-5 | *High* Δ | *Mid* Best | *Mid* Top-5 | *Mid* Δ | *Low* Best | *Low* Top-5 | *Low* Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| COVID/public health | IW-neigh | Tweet | 0.55 | 0.07 | -0.47 | -0.14 | 0.09 | +0.23 | -0.06 | -0.09 | -0.02 |
| | | Bundle | 0.65 | 0.08 | -0.58 | -0.27 | 0.08 | +0.35 | -0.12 | -0.15 | -0.03 |
| | IW-contour | Tweet | 0.51 | -0.02 | -0.53 | -0.18 | -0.11 | +0.07 | -0.07 | -0.09 | -0.02 |
| | | Bundle | 0.65 | -0.01 | -0.66 | -0.23 | -0.14 | +0.09 | -0.04 | -0.10 | -0.06 |
| | MFT | Tweet | 0.34 | 0.51 | +0.17 | -0.01 | 0.15 | +0.16 | -0.12 | 0.04 | +0.16 |
| | | Bundle | 0.42 | 0.61 | +0.18 | 0.04 | 0.23 | +0.19 | -0.15 | 0.11 | +0.26 |
| Economy/jobs/trade | IW-neigh | Tweet | 0.71 | 0.11 | -0.60 | 0.50 | 0.14 | -0.36 | -0.33 | -0.39 | -0.07 |
| | | Bundle | 0.96 | 0.20 | -0.75 | 0.75 | 0.29 | -0.46 | -0.45 | -0.56 | -0.12 |
| | IW-contour | Tweet | 0.77 | -0.08 | -0.85 | 0.46 | 0.35 | -0.12 | -0.38 | -0.34 | +0.04 |
| | | Bundle | 0.97 | -0.07 | -1.04 | 0.76 | 0.56 | -0.20 | -0.53 | -0.50 | +0.02 |
| | MFT | Tweet | 0.35 | 0.55 | +0.20 | -0.01 | 0.22 | +0.23 | -0.21 | -0.22 | -0.01 |
| | | Bundle | 0.62 | 0.82 | +0.19 | 0.04 | 0.39 | +0.35 | -0.35 | -0.34 | +0.01 |
| Foreign policy/security | IW-neigh | Tweet | -0.02 | -0.17 | -0.14 | -0.34 | -0.27 | +0.07 | -0.42 | -0.39 | +0.03 |
| | | Bundle | 0.05 | -0.10 | -0.16 | -0.36 | -0.26 | +0.10 | -0.54 | -0.45 | +0.08 |
| | IW-contour | Tweet | -0.06 | -0.14 | -0.08 | -0.39 | -0.31 | +0.08 | -0.41 | -0.46 | -0.05 |
| | | Bundle | 0.02 | -0.09 | -0.11 | -0.38 | -0.30 | +0.08 | -0.51 | -0.53 | -0.02 |
| | MFT | Tweet | 0.32 | 0.36 | +0.04 | -0.30 | -0.15 | +0.16 | -0.30 | -0.31 | -0.02 |
| | | Bundle | 0.49 | 0.52 | +0.03 | -0.30 | -0.09 | +0.22 | -0.35 | -0.33 | +0.01 |
| Immigration/border | IW-neigh | Tweet | 0.47 | -0.13 | -0.60 | -0.46 | -0.19 | +0.27 | -0.90 | -0.68 | +0.22 |
| | | Bundle | 0.67 | -0.08 | -0.74 | -0.48 | -0.13 | +0.35 | -1.00 | -0.74 | +0.26 |
| | IW-contour | Tweet | 0.48 | 0.05 | -0.44 | -0.38 | -0.27 | +0.11 | -0.88 | -0.85 | +0.03 |
| | | Bundle | 0.72 | 0.14 | -0.58 | -0.35 | -0.27 | +0.08 | -0.98 | -0.97 | +0.01 |
| | MFT | Tweet | -0.81 | -0.83 | -0.02 | -0.45 | -0.15 | +0.30 | 0.57 | 0.74 | +0.17 |
| | | Bundle | -0.94 | -0.95 | -0.01 | -0.52 | -0.10 | +0.42 | 0.80 | 0.93 | +0.14 |
| Judiciary/courts | IW-neigh | Tweet | -0.45 | -0.65 | -0.20 | -0.63 | -0.60 | +0.03 | -0.75 | -0.63 | +0.12 |
| | | Bundle | -0.47 | -0.73 | -0.27 | -0.63 | -0.65 | -0.01 | -0.77 | -0.69 | +0.08 |
| | IW-contour | Tweet | -0.44 | -0.54 | -0.11 | -0.66 | -0.53 | +0.13 | -0.79 | -0.61 | +0.19 |
| | | Bundle | -0.37 | -0.59 | -0.22 | -0.63 | -0.53 | +0.11 | -0.80 | -0.67 | +0.13 |
| | MFT | Tweet | -0.19 | -0.28 | -0.09 | -0.70 | -0.59 | +0.11 | -0.69 | -0.72 | -0.04 |
| | | Bundle | -0.17 | -0.29 | -0.13 | -0.87 | -0.65 | +0.22 | -0.80 | -0.81 | -0.01 |

Table 8: Extended RQ1 mean stance table. Results are reported separately at tweet level and bundle level. For each topic-conditioned persona position, Best is the highest-ranked persona, Top-5 is the average over the five selected personas, and Δ is Top-5 minus Best.

| Topic $t$ | $n$ | Agr. | Leg. | Off. | End. |
|---|---|---|---|---|---|
| COVID | 1,170 | 2.85 | 2.89 | 2.25 | 2.28 |
| Economy | 5,023 | 2.81 | 2.88 | 2.26 | 2.26 |
| Foreign policy | 5,031 | 2.38 | 2.42 | 2.77 | 1.79 |
| Immigration | 2,424 | 2.19 | 2.27 | 3.16 | 1.59 |
| Judiciary | 1,350 | 1.68 | 1.80 | 3.49 | 1.32 |

Table 9: Mean secondary reception ratings by topic at bundle level (1–5 scales), pooled over the three selection strategies and the three persona positions. Agr. = agreement, Leg. = legitimacy, Off. = offensiveness, End. = endorsement. Rows are ordered by increasing offensiveness; $n$ is the number of persona–bundle judgments.

| Strategy / Topic | Stance | Agr. | Leg. | Off. | End. |
|---|---|---|---|---|---|
| IW-neigh | | | | | |
| COVID | +0.22 | +0.14 | +0.09 | −0.29 | +0.07 |
| Economy | +0.77 | +0.92 | +0.73 | −0.49 | +0.79 |
| Foreign policy | +0.35 | +0.37 | +0.32 | −0.26 | +0.13 |
| Immigration | +0.66 | +0.96 | +0.85 | −0.57 | +0.28 |
| Judiciary | −0.05 | −0.07 | −0.11 | −0.10 | −0.09 |
| IW-contour | | | | | |
| COVID | +0.09 | −0.02 | −0.03 | −0.28 | −0.22 |
| Economy | +0.44 | +0.46 | +0.32 | −0.17 | +0.33 |
| Foreign policy | +0.43 | +0.49 | +0.38 | −0.36 | +0.32 |
| Immigration | +1.12 | +1.56 | +1.29 | −0.89 | +0.71 |
| Judiciary | +0.08 | +0.11 | +0.15 | −0.27 | +0.03 |
| MFT | | | | | |
| COVID | +0.50 | +0.80 | +0.65 | +0.02 | +1.08 |
| Economy | +1.16 | +1.47 | +1.06 | −0.46 | +1.46 |
| Foreign policy | +0.86 | +1.08 | +0.83 | −0.37 | +1.01 |
| Immigration | −1.88 | −2.81 | −2.19 | +0.99 | −1.94 |
| Judiciary | +0.51 | +0.73 | +0.63 | −0.45 | +0.55 |

Table 10: High-minus-low persona-position contrasts at bundle level, for stance and for the four scalar labels. Positive values mean the *high* position scores higher than the *low* position. Agreement, legitimacy, and endorsement track stance, whereas offensiveness moves in the opposite direction, as intended by the label design. The sign reversal for MFT Immigration/border follows from its topic mapping, where the *high* position is Care-oriented and the *low* position is Loyalty/Authority-oriented (Table 6).

| Topic $t$ | $n$ | Agreement | Legitimacy | Offensiveness | Endorsement |
|---|---|---|---|---|---|
| COVID | 11,295 | 2.72 | 3.03 | 1.85 | 2.41 |
| Economy | 50,050 | 2.62 | 2.94 | 1.81 | 2.28 |
| Foreign policy | 49,950 | 2.22 | 2.45 | 2.36 | 1.84 |
| Immigration | 24,195 | 2.11 | 2.33 | 2.80 | 1.63 |
| Judiciary | 13,500 | 1.66 | 1.92 | 3.07 | 1.43 |

Table 11: Mean secondary reception ratings by topic at message level, the counterpart of Table 9. The topic ordering is identical, but bundle-level values are more polarized, as expected when the judgment summarizes a coherent topic-matched bundle rather than isolated posts.

| | By stance | | | By strategy | | |
|---|---|---|---|---|---|---|
| Topic $t$ | support | neutral | reject | IW-neigh | IW-contour | MFT |
| COVID | 1.92 | 2.18 | 2.70 | 2.31 | 2.32 | 2.12 |
| Economy | 1.93 | 2.10 | 2.80 | 2.30 | 2.30 | 2.18 |
| Foreign policy | 2.24 | 2.39 | 3.26 | 2.82 | 2.84 | 2.65 |
| Immigration | 2.52 | 2.46 | 3.69 | 3.25 | 3.28 | 2.95 |
| Judiciary | 2.39 | 2.51 | 3.86 | 3.59 | 3.46 | 3.40 |

Table 12: Mean offensiveness at bundle level, broken down by response stance (support, neutral, reject) and by selection strategy (IW-neigh, IW-contour, MFT). The tone penalty attached to rejection is topic-dependent, and the topic ordering is preserved within each selection strategy taken separately.

| Mode | Variant | Topic | Records | Skipped | Rate |
|---|---|---|---|---|---|
| IW | contour | `Foreign policy/security` | 1680 | 1 | 0.06% |
| IW | neigh | `Foreign policy/security` | 1680 | 4 | 0.24% |
| IW | neigh | `Immigration/border` | 810 | 5 | 0.62% |
| MFT | top-5 | `Economy/jobs/trade` | 1680 | 17 | 1.01% |
| MFT | top-5 | `Foreign policy/security` | 1680 | 4 | 0.24% |
| MFT | top-5 | `Immigration/border` | 810 | 1 | 0.12% |
| Total | – | – | 15030 | 32 | 0.21% |

Table 13: RQ1 parse-error diagnostics. Rows report output files with at least one skipped parse error. Remaining output files had zero skipped parse errors.

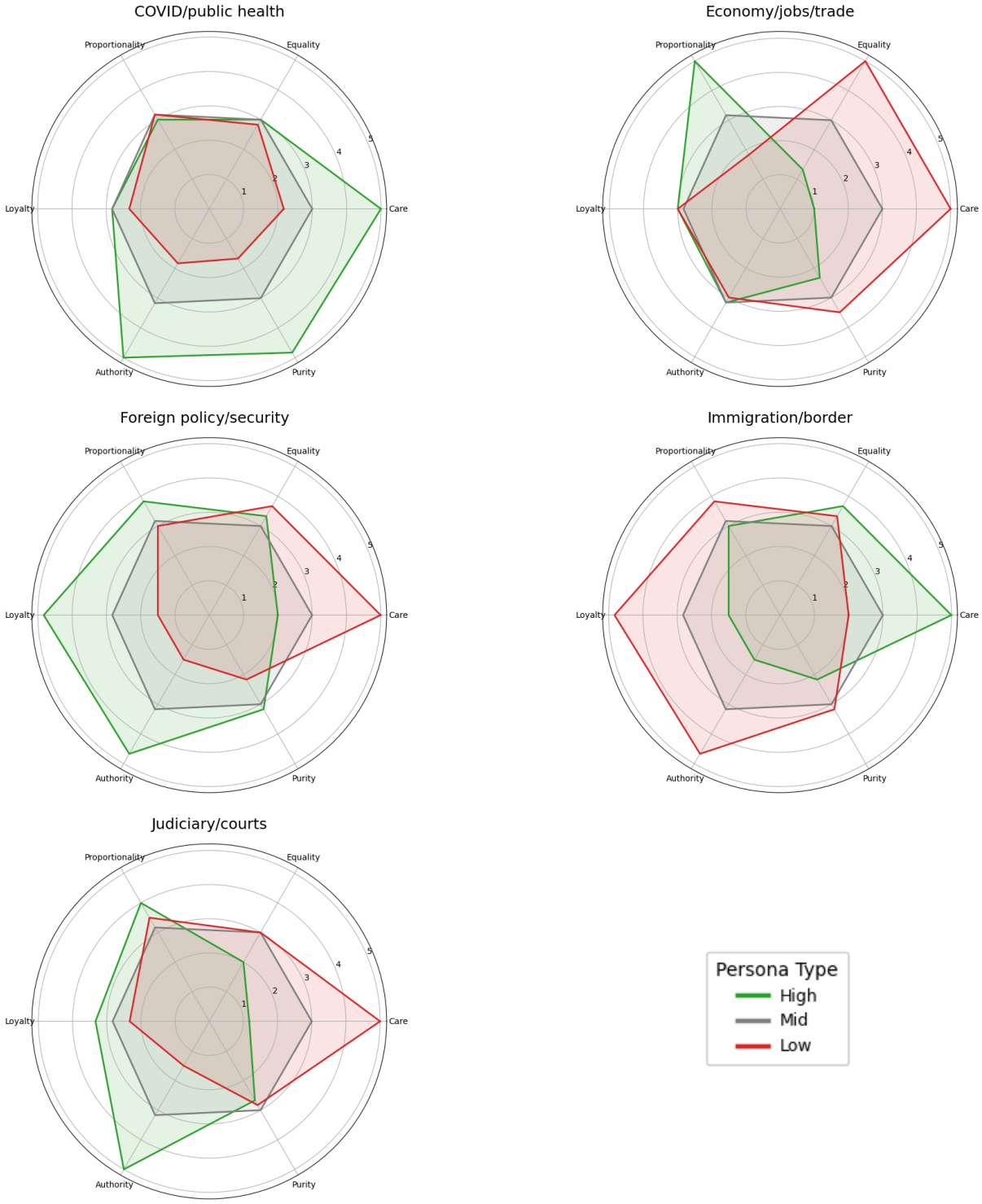


Figure 6: MFT best profiles selected for each topic-position pair. Each radar chart corresponds to one topic and shows the closest persona to the *high* (green), *mid* (grey), and *low* (red) moral target profiles in the six-dimensional MFT space.

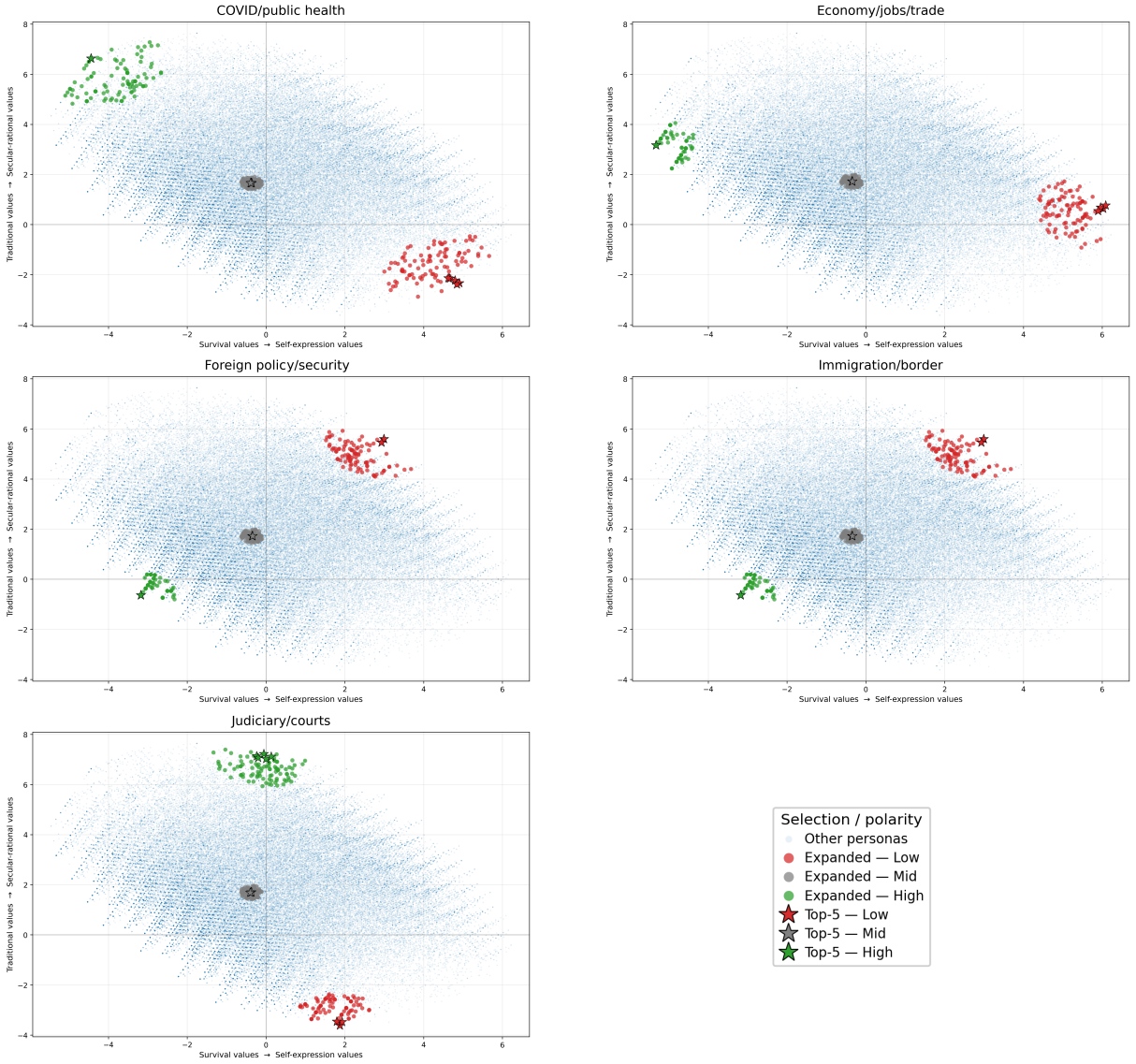


Figure 7: Expanded IW-neigh persona samples used for RQ2. Each panel corresponds to one topic and shows the expanded top-100 selection in the IW space. The selection preserves proximity to the original neighbor-based topic-position anchors while increasing demographic-signature diversity.

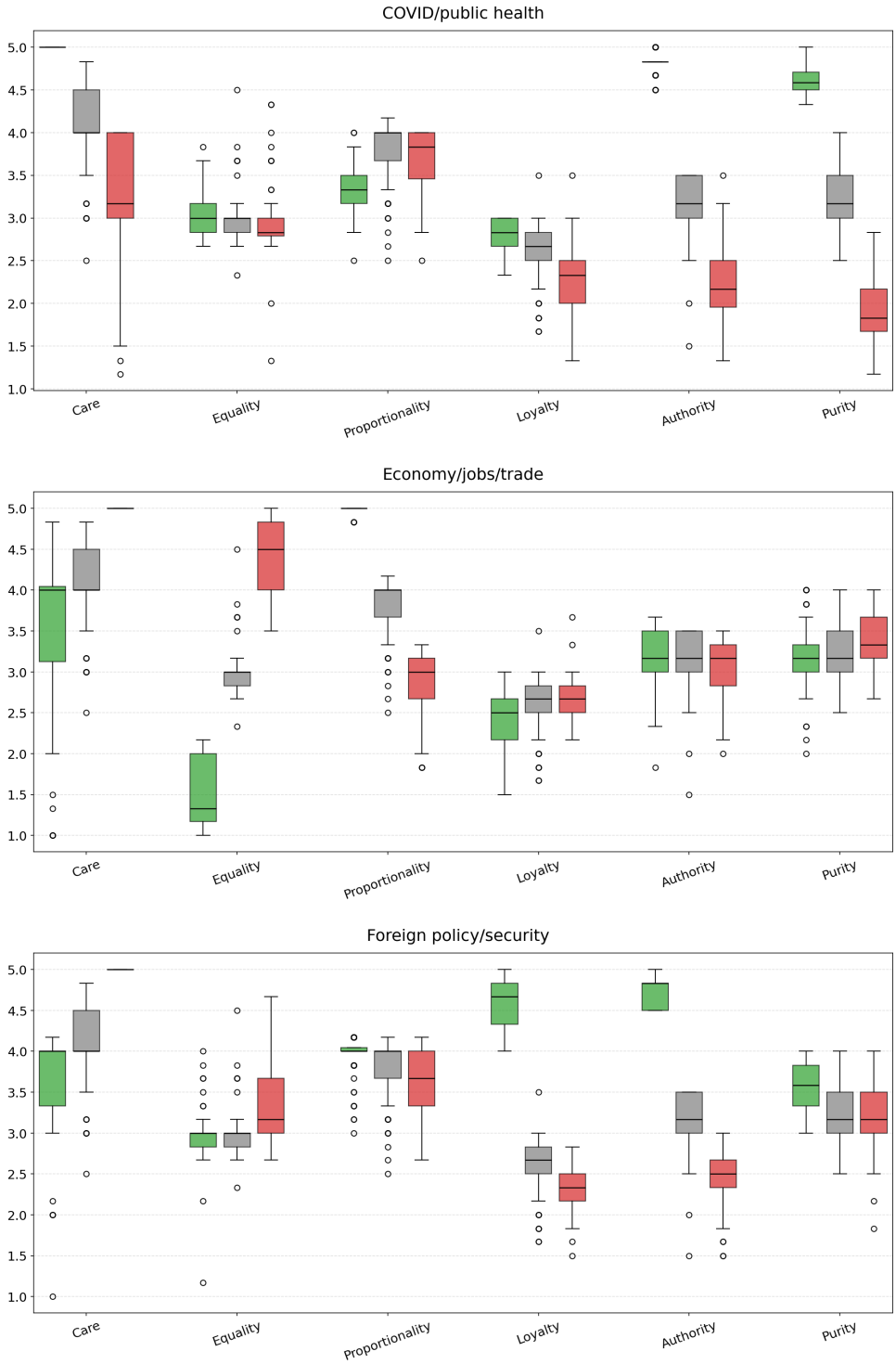


Figure 8: Boxplots for the expanded MFT top-80 persona samples used in RQ2. Each panel reports MFT score distributions by topic-conditioned position: *high* in green, *mid* in gray, and *low* in red.

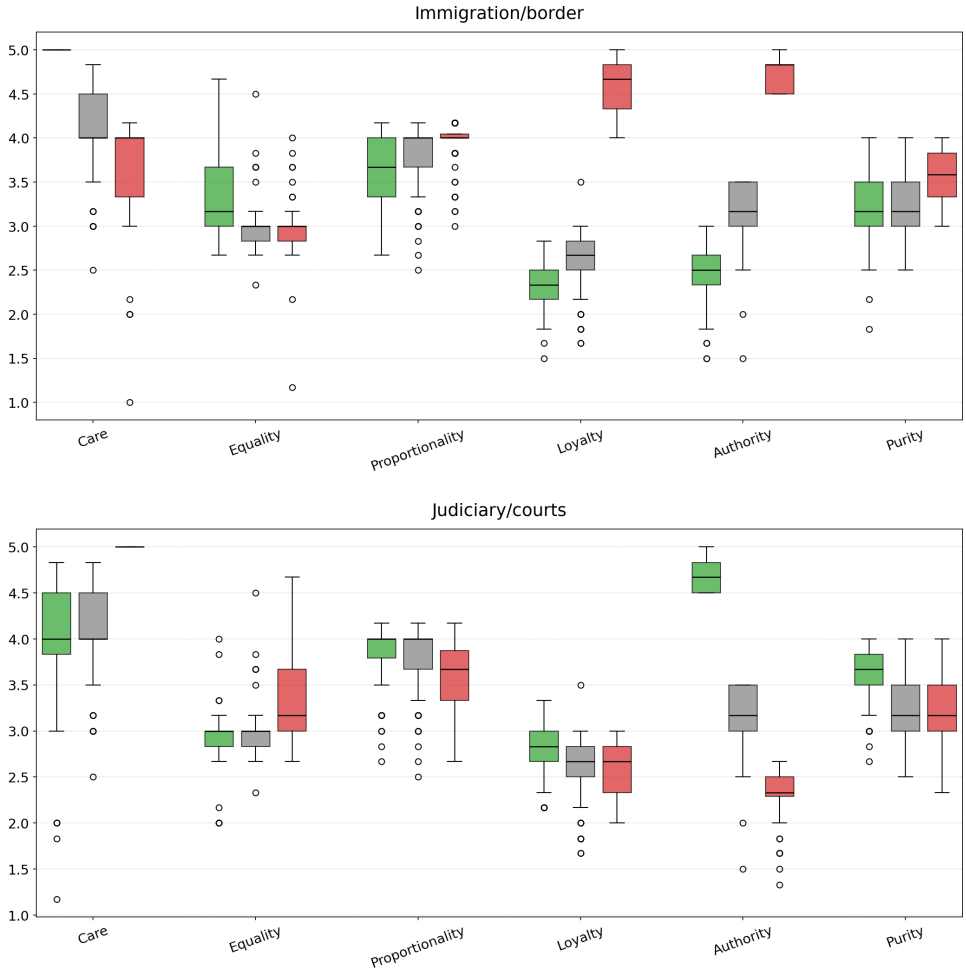


Figure 8: *(Continued)* Boxplots for the expanded MFT top-80 persona samples used in RQ2.

| Strategy $f$ | Topic $t$ | Stance $s$ | Closed frequent itemset $I$ | $\lvert G_{f,t,s}\rvert$ | supp($I$) | Lift |
|---|---|---|---|---|---|---|
| IW | Immigration | support | Authority = high; Proportionality = high; social trust = cannot trust people | 200 | 0.63 | 1.37 |
| IW | Immigration | support | education = tertiary; Authority = high; Proportionality = high; social trust = cannot trust people | 200 | 0.62 | 1.37 |
| IW | Immigration | support | Authority = high; social trust = cannot trust people | 200 | 0.64 | 1.35 |
| IW | Immigration | support | education = tertiary; Authority = high; social trust = cannot trust people | 200 | 0.64 | 1.35 |
| IW | Immigration | support | Loyalty = high; Proportionality = high | 200 | 0.60 | 1.34 |
| IW | Immigration | support | Authority = high; Proportionality = high | 200 | 0.78 | 1.33 |
| IW | Immigration | support | education = tertiary; Authority = high; Proportionality = high | 200 | 0.76 | 1.33 |
| IW | Immigration | support | Authority = high; Proportionality = high; settlement = urban | 200 | 0.70 | 1.32 |
| IW | Immigration | support | education = tertiary; Authority = high; Proportionality = high; settlement = urban | 200 | 0.70 | 1.32 |
| IW | Immigration | support | education = tertiary; Authority = high | 200 | 0.78 | 1.32 |
| IW | Immigration | support | Authority = high; Care = high; Proportionality = high | 200 | 0.68 | 1.32 |
| IW | Immigration | support | education = tertiary; Authority = high; Care = high; Proportionality = high | 200 | 0.68 | 1.32 |
| MFT | Immigration | reject | age = 30–39; settlement = urban | 225 | 0.54 | 1.17 |
| MFT | Immigration | reject | age = 30–39; education = tertiary | 225 | 0.56 | 1.16 |
| MFT | Immigration | reject | materialism orientation = postmaterialist; settlement = urban | 225 | 0.54 | 1.15 |
| MFT | Immigration | reject | education = tertiary; settlement = urban | 225 | 0.70 | 1.13 |
| IW | Judiciary | support | education = tertiary; Authority = high | 219 | 0.63 | 1.11 |
| IW | Judiciary | support | Authority = high; Proportionality = high | 219 | 0.61 | 1.11 |
| IW | Judiciary | support | education = tertiary; Authority = high; Proportionality = high | 219 | 0.60 | 1.10 |
| MFT | Immigration | reject | continent = Europe; education = tertiary | 225 | 0.53 | 1.09 |
| MFT | Judiciary | neutral | education = tertiary; IW Traditional–Secular = secular-rational | 221 | 0.48 | 1.09 |
| MFT | Immigration | reject | education = tertiary; materialism orientation = postmaterialist | 225 | 0.60 | 1.08 |
| MFT | Judiciary | reject | education = tertiary; IW Traditional–Secular = secular-rational | 225 | 0.48 | 1.07 |
| MFT | Judiciary | neutral | education = tertiary; settlement = urban | 221 | 0.58 | 1.06 |
| MFT | Judiciary | reject | education = tertiary; settlement = urban | 225 | 0.57 | 1.05 |
| MFT | Immigration | neutral | education = tertiary; settlement = urban | 194 | 0.68 | 1.05 |
| MFT | Judiciary | neutral | education = tertiary; materialism orientation = postmaterialist | 221 | 0.53 | 1.03 |
| MFT | Judiciary | neutral | education = tertiary; gender = woman | 221 | 0.52 | 1.03 |
| MFT | Judiciary | neutral | education = tertiary; national pride = not proud at all | 221 | 0.49 | 1.03 |
| MFT | Judiciary | reject | education = tertiary; gender = woman | 225 | 0.52 | 1.03 |
| MFT | Judiciary | neutral | age = 30–39; education = tertiary | 221 | 0.55 | 1.02 |
| MFT | Immigration | neutral | age = 30–39; education = tertiary | 194 | 0.53 | 1.02 |
| IW | COVID | neutral | education = tertiary; Care = high; settlement = urban | 297 | 0.76 | 1.01 |
| IW | COVID | neutral | education = tertiary; Care = high; Proportionality = high; settlement = urban | 297 | 0.69 | 1.01 |
| IW | COVID | neutral | education = tertiary; Care = high | 297 | 0.80 | 1.01 |
| IW | COVID | neutral | Care = high; Proportionality = high | 297 | 0.72 | 1.01 |
| MFT | COVID | neutral | age = 30–39; settlement = urban | 220 | 0.52 | 1.02 |
| MFT | COVID | neutral | IW Traditional–Secular = secular-rational; settlement = urban | 220 | 0.51 | 1.02 |
| MFT | COVID | neutral | education = tertiary; settlement = urban | 220 | 0.60 | 1.01 |
| MFT | COVID | reject | education = tertiary; materialism orientation = postmaterialist | 224 | 0.50 | 1.01 |
| IW | Economy | reject | age = 30–39; education = tertiary; settlement = urban | 295 | 0.65 | 1.01 |
| IW | Economy | reject | age = 30–39; education = tertiary; Proportionality = high; settlement = urban | 295 | 0.60 | 1.01 |
| IW | Economy | reject | age = 30–39; education = tertiary; Care = high; settlement = urban | 295 | 0.60 | 1.01 |
| IW | Economy | reject | education = tertiary; settlement = urban | 295 | 0.93 | 1.01 |
| MFT | Economy | reject | IW Traditional–Secular = secular-rational; settlement = urban | 223 | 0.59 | 1.01 |
| MFT | Economy | reject | education = tertiary; IW Traditional–Secular = secular-rational | 223 | 0.65 | 1.01 |
| MFT | Economy | reject | education = tertiary; settlement = urban | 223 | 0.78 | 1.00 |
| IW | Economy | support | education = tertiary; Authority = high | 300 | 0.60 | 1.01 |

Table 14: Broader selection of non-singleton closed frequent itemsets for RQ2. Support is computed within ⟨framework strategy, topic, observed response⟩ group. The first block reports the strongest multi-dimensional patterns; the second reports secondary non-immigration patterns; the third reports weaker patterns for `COVID/public health` and `Economy/jobs/trade`.

| Topic | Pew | IW-neigh B | IW-neigh T | IW-contour B | IW-contour T | MFT B | MFT T |
|---|---|---|---|---|---|---|---|
| `COVID/public health` | -17.02 | 0.26 | 2.68 | -8.21 | -7.36 | 31.54 | 23.59 |
| `Economy/jobs/trade` | 8.12 | -2.44 | -4.84 | -0.36 | -2.50 | 28.38 | 17.79 |
| `Foreign policy/security` | -3.29 | -27.03 | -27.61 | -30.61 | -30.36 | 3.34 | -3.34 |
| `Immigration/border` | -0.38 | -31.30 | -33.11 | -36.67 | -35.71 | -4.08 | -7.99 |
| `Judiciary/courts` | -5.71 | -68.89 | -62.38 | -59.33 | -55.96 | -58.22 | -53.16 |

Table 15: Topic-level net scores used for RQ3 diagnostics. Net is positive percentage minus negative percentage. Pew values are weighted valid-response distributions. Synthetic values are computed from stance labels, with support as positive and reject as negative. B = bundle level; T = tweet/message level.

| Topic $t$ | $pos$ | Stance $s$ | Emotional reaction | Brief reason |
|---|---|---|---|---|
| COVID | High | support | Mostly supportive | Most messages are supportive of aid and public health measures, with few neutral or negative remarks. |
| COVID | Mid | neutral | Mixed feelings, cautious optimism | Some claims are credible, many are conspiratorial and offensive. |
| COVID | Low | reject | Disgusted and concerned about political rhetoric | Most messages are hostile, conspiratorial, and lack constructive content. |
| Economy | High | support | Strongly supportive, some concerns about hyperbole | Messages align with Linda's values on low taxes, job protection, and anti-foreign influence, though some are hyperbolic or insulting. |
| Economy | Mid | support | Mixed but generally supportive of fiscal responsibility | Lily sees many messages as aligning with her preference for lower taxes, fiscal responsibility, and tough trade stance, though some claims lack evidence. |
| Economy | Low | reject | Discomfort with partisan, offensive rhetoric | All messages are partisan, many contain insults or hateful language, none align with my values. |
| Foreign policy | High | support | Strongly supportive of national security | Most messages align with anti-foreign influence, anti-leftist, and anti-terrorism, which Maria supports. |
| Foreign policy | Mid | neutral | Mixed feelings: supportive of peace, uneasy about hostility | Positive about peace deals, but some messages contain hostile rhetoric. |
| Foreign policy | Low | reject | Disgusted, uneasy | All messages blame Biden, promote militarism, lack evidence, and use hostile rhetoric, conflicting with values of peace and community. |
| Immigration | High | support | Concerned but uneasy | Most messages emphasize a border crisis, aligning with Maria's security concerns; style is often insulting, reducing endorsement. |
| Immigration | Mid | support | Concerned but wary of rhetoric | Supports border security but disapproves of hyperbole, insults, and conspiratorial claims. |
| Immigration | Low | reject | Deeply concerned and disgusted | All messages contain hostile, xenophobic rhetoric lacking evidence, contrary to inclusive values. |
| Judiciary | High | reject | Skeptical, annoyed | All messages are unsubstantiated, hostile, and lack credibility. |
| Judiciary | Mid | reject | Disgusted and skeptical | All messages contain unsubstantiated, hostile claims about judicial misconduct, perceived as conspiratorial and offensive. |
| Judiciary | Low | reject | Discomfort and disapproval | The messages promote a travel ban and hostile rhetoric, which Ingrid finds morally objectionable and lacking evidence. |

Table 16: Qualitative examples from IW-neigh best-profile outputs. Each row reports one bundle-level response for the best persona for each $\langle t, pos \rangle$ pair.