Values Are Not Single Labels: Distributional Value Profiling Across Groups and Contexts

Woojin Kim[1] **Jaeyoung Do**[1,2,†]
AIDAS Laboratory, [1]ECE & [2]APAI, Seoul National University
{wjk9904, jaeyoung.do}@snu.ac.kr

Abstract

Human values are not single labels, but distributions: they vary across individuals, contexts, and their relationships with other values. Yet most alignment approaches still compress group-level values into scalar scores or single representative targets, obscuring within-group variation and context-dependent expression. We introduce **Distributional Value Profiling**, a framework for representing group-level value expression as distributions over expression modes, intensity, and inter-value relationships. Using large scale text corpora associated with cultural, political, and religious groups, we construct profiles that capture how values are expressed, how strongly they are emphasized, and how they relate to one another. We then model context-dependent shifts in these profiles using optimal transport, revealing structured redistributions of value expression across situations. To investigate how these profiles can support alignment, we use them as targets for prediction and activation-based steering. A lightweight profiling model predicts held-out expression distributions, intensity summaries, and contextual shift magnitudes within the constructed value and context space, while profile-derived steering improves average agreement with survey-response distributions at inference time across the evaluated benchmarks. Together, these results establish distributional value profiling as a concrete step toward pluralistic alignment, treating within-group diversity not as noise to be averaged away, but as structure to be measured, predicted, and used to guide model responses.

1 Introduction

Large language models are increasingly deployed in settings where their responses shape how people interpret information and make decisions, from personal advice to public discourse [32, 58]. In these settings, alignment is not only about factual accuracy, but also about the values a model reflects in its responses [16, 26]. Yet human values are not simple or stable targets. They vary across individuals and situations, and are often expressed indirectly through framing, emphasis, qualification, and trade-offs, rather than stated explicitly [46, 47, 55]. For example, when asked about work-life balance, responses may express firm support or qualify that support in light of career demands. A single situation can involve multiple values whose relative importance and modes of expression shift with context [39, 47].

Recent work represents values explicitly and uses them to guide model behavior [64, 52, 12, 24]. However, approaches that summarize a group's values with scalar scores or fixed value weights can obscure distinct expression modes, within-group variation, and how expression shifts with context [38, 24]. For example, two groups can have similar average scores while differing in the proportions of firm support, qualification, and rejection. Retaining these proportions distinguishes a shared tendency toward qualified support from a mixture of contrasting modes of expression.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

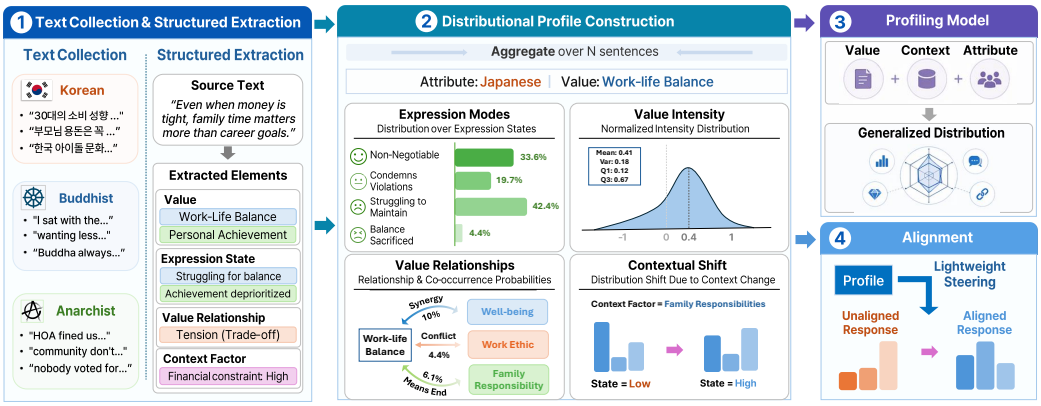


Figure 1: Overview of the framework. Structured value signals are extracted from group-associated text, aggregated into distributional profiles, and used for profile generalization and alignment.

To address this, we introduce **Distributional Value Profiling (DVP), which represents group-level value expression as distributions** over *how values are expressed*, *how strongly they are emphasized*, and *how they relate to one another*. This reflects distributional pluralism's view [53] that alignment should preserve within-group diversity rather than collapse it into a single representative response. We put this idea into practice by inferring distributions from sampled naturally occurring text, treating these signals not as direct measurements of individual belief but as **attribute-associated discourses** that capture patterns of engagement such as support, critique, and trade-off.

Following this discourse-level view, we construct **distributional value profiles** across 21 groups spanning cultural, political, and religious settings. Each profile represents value expression, intensity, and inter-value relationships as empirical distributions. We further use optimal transport to characterize how these distributions change across contexts. The transport cost accounts for both the amount of probability mass redistributed and the distances it moves between expression states. For example, a group may treat work-life balance as non-negotiable under low financial pressure, but express it as difficult to maintain as financial constraints increase. This reflects a redistribution of expression modes, not merely a decrease in an average score.

The resulting profiles reveal structured differences both across and within groups. For example, the language-associated corpora share *Equality Defended* as the dominant mode for gender equality, but differ in the proportions of the other expression modes (Figure 4). In Buddhism-associated discourse, compassion shifts from being defended under dismissive contexts to being treated as a duty under supportive ones (Figure 5). Beyond descriptive analysis, the constructed profiles also serve as actionable targets for prediction and alignment. A lightweight profiling model predicts expression distributions, intensity summaries, and contextual shift magnitudes for held-out profile records. Profile-based activation steering improves average agreement with survey-response distributions at inference time.

Below, we summarize our contributions:

- **A distributional formulation of value expression.** We represent group-level value expression through distributions over expression modes, intensity, and inter-value relationships, retaining variation that scalar scores or single representative labels omit.
- **Distributional value profiling with context modeling.** We construct group-level value profiles from large-scale text across 21 cultural, political, and religious groups, and model how these distributions shift across contexts using optimal transport.
- **Actionable distributional profiles for prediction and alignment.** We demonstrate prediction of held-out profile records and use expression profiles to steer model responses, improving average agreement with survey-response distributions across multiple models. We release the full set of constructed profiles and code to support further research.

2 Related Works

Human Values in LLMs Human values are commonly defined as desirable, trans-situational goals that guide behavior [47, 46, 40, 62]. Recent work incorporates human values into LLMs through representation, alignment, and evaluation. For representation, prior work maps text to predefined value taxonomies or learns value embeddings from data [35, 51, 61, 5, 24, 52]. For alignment, methods steer model behavior through prompting [41, 24, 12], fine-tuning [23, 61], or value-conditioned control [20, 67]. For evaluation, methods use structured questionnaires [38, 44, 66], as well as generative measurement and calibrated evaluators [63, 64, 60]. Closest to our work, Sorensen et al. [52] infer natural-language profiles from individual raters' example judgments to predict their ratings, whereas DVP estimates context-dependent distributions of group-level value expression from sampled discourse.

Distributional Alignment Recent work on distributional pluralism [28, 53, 30, 25, 10] and social simulation [1, 34, 2, 18] emphasizes representing variation in human preferences rather than collapsing it into a single response. Survey based approaches use large scale instruments such as the World Values Survey [13] to benchmark [43, 9, 66, 30] or improve distributional alignment [54, 6, 22, 10]. Human annotation approaches instead collect preference judgments or dialogue feedback [25, 21] and use them for reward-based alignment [36, 7, 19]. Russo et al. [42] model judgment distributions and the prevalence of values in open-ended moral rationales. DVP focuses on how individual values are expressed through expression modes, intensity, and inter-value relations across groups and contexts. Our focus on distributional alignment complements steerable and Overton approaches to pluralism [53].

3 Value Distributions and Profiling

3.1 Modeling Group-Level Value Distributions

Let $a \in \mathcal{A}$ denote an attribute (e.g., a demographic or cultural group), $v \in \mathcal{V}$ a value, and $z \in \mathcal{Z}$ a value expression state. We model how value v is expressed in discourse associated with attribute a as a conditional distribution $p(z \mid v, a)$. This formulation captures group-level value expression as a distribution over multiple expression modes, rather than a single summary. For example, a may denote Buddhism-associated discourse, v compassion, and z "Compassion as Duty" or "Compassion Under Strain."

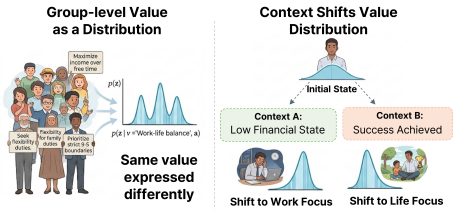


Figure 2: Schematic distributions of group-level value expression under different contexts.

Distributional Value Profile. The expression-state component of the value profile of value v under attribute a is defined as $P_a(v) := p(z \mid v, a)$, which we estimate from sampled text associated with a. We denote the collection of the three component distributions described below by $\mathcal{P}_a(v)$.

3.2 Decomposition of the Distributional Profile

Building on work on values and behavior [3, 47, 46, 40], we operationalize $\mathcal{P}_a(v)$ through three complementary components: (i) *how* a value is expressed, (ii) *how strongly* it is expressed, and (iii) *how it interacts* with other values. For example, in group-associated discourse, work–life balance may be actively upheld or described as difficult to maintain (expression mode), emphasized to varying degrees (intensity), and framed as either reinforcing well-being or conflicting with career ambition (relations).

(1) Value Expression Modes $p(z \mid v, a)$. We model the qualitative modes through which a value is expressed, capturing how it is manifested under different conditions (e.g., actively pursued, conditionally expressed, or suppressed), where $z \in \mathcal{Z}$ denotes the expression state.

(2) Value Intensity $p(s \mid v, a)$. We model the strength with which a value is expressed, where $s \in \mathbb{R}$ denotes the intensity. In practice, each instance yields a scalar score, estimated by projecting sentence representations onto value-specific axes.

(3) Value Relations. We model value relationships at two levels: which values co-occur with v, captured as a distribution $p(v' \mid v, a)$ over $v' \in \mathcal{V}$, and how they interact, captured as a distribution over relation types (e.g., reinforcement, tradeoff) for each pair (v, v').

3.3 Effect of Context

Value expression varies with the context in which it occurs [48, 57, 46]. We define **context** $c \in \mathcal{C}$ through situational factors such as framing, social setting, or competing constraints that are associated with how a value is expressed. As illustrated in Figure 2, the same group can exhibit different value distributions under different contextual conditions. To capture this, we condition the expression-state component on context: $P_a(v, c) := p(z \mid v, a, c)$, with intensity and relation distributions conditioned analogously to form the context-conditioned profile $\mathcal{P}_a(v, c)$.

4 Constructing Distributional Value Profiles

This section describes the practical pipeline for estimating distributional value profiles from large-scale text. We treat the collected corpora not as direct measurements of individual belief, but as attribute-associated discourse that reflects how value-related topics are framed across contexts. We first describe the data sources, curation, and corpus scale (Section 4.1), then present the structured extraction pipeline for value assignment, expression annotation, intensity parsing, and relationship extraction (Section 4.2). We then model context-dependent distributional shifts using optimal transport (Section 4.3) and analyze the resulting profiles across attributes (Section 4.4).

Table 1: Target attributes and data sources used for constructing distributional value profiles.

Category	Attributes	Source
Cultural	Arabic, German, French, Japanese, Korean, Russian, Chinese	CulturaX [33]
Political	Anarchism, Communism, Conservatism, Democrats, Libertarianism, Republicans, Socialism	Reddit [4]
Religious	Atheism, Buddhism, Catholicism, Christianity, Hinduism, Islam, Judaism	Reddit [4]

4.1 Dataset

We collect attribute-specific corpora [33, 4] across cultural, political, and religious groups, as summarized in Table 1. Because cultural profiles are built from language-specific web text whereas political and religious profiles are built from Reddit community discussions, cross-category differences may include source-domain effects. We examine source dependence in Section 6.3. After filtering low-quality content, we treat sentences as the annotation unit to capture local variation while retaining surrounding document or thread context. Each sentence is processed by the structured extraction pipeline described in Section 4.2.

Table 2: Sparsity estimates and implied corpus sizes.

Sparsity Estimates	
Quantity	**Value**
$P(r)$	0.108
$P(v \mid r)$	0.013
$P(c \mid r, v)$	0.5
$P(v, c)$	0.0007
Implied Corpus Size	
Target K	**# Sentences**
$K = 100$	$\sim$ 143K
$K = 200$	$\sim$ 286K
$K = 500$	$\sim$ 714K

Data Scale. Reliable estimation of value-context distributions requires large corpora, because value signals are sparse and are further divided across values and contexts. Let r denote the event that a sentence is value-relevant. Then $P(r)$ is the rate of value-relevant sentences, $P(v \mid r)$ is the probability that such a sentence is assigned to a specific value v, and $P(c \mid r, v)$ is the corresponding context-availability rate. Their product, $P(v, c)$, gives the rate of observing value v with contextual information. Obtaining K such observations on average requires approximately $K/P(v, c)$ sentences. The rates in Table 2 therefore motivate corpora on the order of 10^5–10^6 sentences per attribute. In our experiments, we use approximately 141K-1.25M sentences per attribute after text filtering. Per-attribute counts and subsequent filtering stages are reported in Appendix B.1.

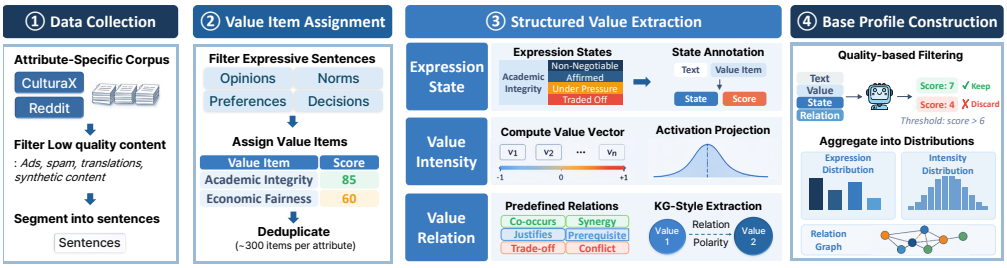


Figure 3: Pipeline for constructing base distributional value profiles.

4.2 Profile Construction

We construct distributional value profiles by extracting structured value signals from attribute-associated text, using each sentence and its surrounding context. As illustrated in Figure 3, the pipeline identifies value-relevant sentences, assigns value items, and annotates three profile components: expression states, intensity scores, and inter-value relations. Aggregating these annotations within each attribute-value pair yields the empirical distributions used as group-level value profiles. Appendix B describes the annotation procedures, model configurations, and filtering criteria.

Value Item Assignment. We identify sentences expressing opinions, preferences, or normative claims using an LLM-based relevance filter. For each value-relevant sentence, a value item extractor then assigns candidate value items with relevance scores, and we retain high confidence assignments, resulting in over 100K raw value item instances per attribute. To obtain a compact and semantically coherent value space, we cluster semantically similar items within each attribute using `Qwen3-Embedding-8B`, and map each cluster to a representative value. We then merge the attribute-level vocabularies through manual review, removing near-duplicates while retaining group-specific concepts. This produces approximately 2K unique representative value items with definitions. We map the original sentence-level assignments to this shared vocabulary by embedding similarity for downstream profile construction.

Expression State Annotation. For each value item, we define a set of expression states that characterize how the value is manifested in text, such as whether it is affirmed, constrained, or set aside. The schema comprises four invoked modes and a separate *Not Invoked* category for sentences that do not invoke the value. The state schema is generated and refined through LLM-based auditing and human inspection for textual observability, coverage, and separation between states. We then use a value expression parser to assign an expression state and confidence score to each sentence-value pair, discarding low confidence cases. We exclude *Not Invoked* and normalize the retained counts over the four invoked modes within each attribute-value pair, yielding the empirical expression distribution $p(z \mid v, a)$. Two corpora can therefore have similar expression-mode profiles even if the value is invoked much more frequently in one corpus than in the other.

Intensity Parsing. To quantify how strongly a value is expressed, we estimate a continuous intensity score for each sentence-value pair. Because scalar ratings can be unstable and ranking-based alternatives are limited for large value spaces [24], we adopt an activation-based approach inspired by representation engineering methods [68, 56, 8]. We construct value-specific axes from activation differences between positive and negative seed instances, project sentence representations onto these axes, and average the projections across selected layers within each model. We then normalize scores to $[-1, 1]$ within each model-value pair and average across four models. The resulting scores define the empirical intensity distribution, with higher scores indicating stronger expression toward the positive seed direction. We fit Gaussian summaries for downstream modeling.

Relationship Extraction. Finally, we model interactions between values by extracting pairwise relations among co-occurring value items within a sentence. For each pair, a relationship parser predicts the relation type, direction, and polarity using a knowledge graph style extraction process. We organize relation types into three broad groups: ASSOCIATION, where values co-occur or support each other; DIRECTIONAL, where one value enables, justifies, or leads to another; and TENSION, where values conflict or require trade-offs. For example, "Honesty is a prerequisite for accountability" links *honesty* to *accountability* through a PREREQUISITE relation. For each attribute-value pair,

Table 3: Group-level summary statistics of the constructed profiles. Value Supp. is the average number of samples per item; Eff. States and Eff. Part. are entropy-based effective counts of active expression modes and partners; Sig. is the fraction of context pairs with statistically significant shifts.

	Expression Mode				Intensity				Relations		
	Base		Context Shift		Base		Context Shift		Base		Context
Group	Value Supp.	Eff. States	OT	Sig.	Gauss. (%)	Skew.	OT	Sig.	Mean Deg.	Eff. Part.	Joint OT
Cultural	491	2.8	.33	58.8	97.7	-0.04	.12	67.5	16.5	9.5	.20
Political	684	2.9	.29	47.9	82.8	$+0.36$	.10	52.7	17.6	9.0	.20
Religious	714	3.0	.31	53.3	93.9	$+0.22$	.10	57.8	19.9	10.2	.21

we normalize co-occurrence counts to estimate the distribution over partner values, and relation counts to estimate the distribution of relation types for each partner. Their product gives the base joint value-relation representation. Contextual profiles use normalized joint edge counts within each context.

4.3 Dealing with Context

While the profiles above aggregate expression across contexts, real-world expression is context-dependent [46, 57]. To capture this variability, we model context as structured factors and estimate how value distributions shift under different conditions. We measure these shifts with optimal transport, which accounts for both the amount of probability mass redistributed and the distances it moves.

Context Info Extraction. We define a fixed set of context factor groups (e.g., resource constraints, risk and uncertainty; see Table 13 for the full list) and generate candidate context factors for each value item. Each factor is associated with a small set of discrete states, together with definitions and lexical cues. Given a sentence and a value item, a context parser then assigns the most likely state for each relevant factor, producing a structured context representation as factor-state assignments. For each group and value, these assignments define the subsets used to construct profiles for the states of a given factor. Expression profiles again normalize over invoked modes, intensity profiles pool the corresponding scores, and contextual relation profiles normalize joint edge counts within each state.

Contextual Shift via Optimal Transport. We use optimal transport (OT) to quantify context-dependent shifts because it accounts for the specified geometry of the underlying space. Given two context-conditioned distributions $\mu = p(\cdot \mid v, a, c_1)$ and $\nu = p(\cdot \mid v, a, c_2)$ for two states c_1 and c_2 of the same context factor, OT measures the minimal cost of transporting probability mass from μ to ν:

$$\mathcal{W}_d(\mu, \nu) = \min_{\gamma \in \Pi(\mu,\nu)} \sum_{i,j} \gamma_{ij}\, d(x_i, y_j),$$

where $\Pi(\mu, \nu)$ denotes the set of couplings with marginals μ and ν, x_i and y_j are support points of the two empirical distributions, and d is the transport cost. For expression modes, costs follow equally spaced positions of the invoked modes in the state-schema order; for intensity, we use absolute differences between empirical scores. For example, suppose the first two expression modes have probabilities 0.8 and 0.2. Moving 0.1 of the probability mass from the second mode to the third gives an OT distance of 0.1, while moving that same mass to the fourth gives a distance of 0.2. The first mode remains dominant in both cases. Relations use a structured greedy transport approximation over partner values and relation types. Section 6.1 examines sensitivity to alternative expression-state costs. Full cost definitions and support criteria are provided in Appendix B.

4.4 Profile Results

Table 3 summarizes the coverage, diversity, and contextual variation of the constructed profiles. The profiles contain several hundred annotations per value on average. Across categories, the median effective-state count is around three, while values connect to roughly ten effective relational partners on average. These summaries show that the profiles retain variation across expression modes and relationships with other values. Contextual variation also appears across all three components:

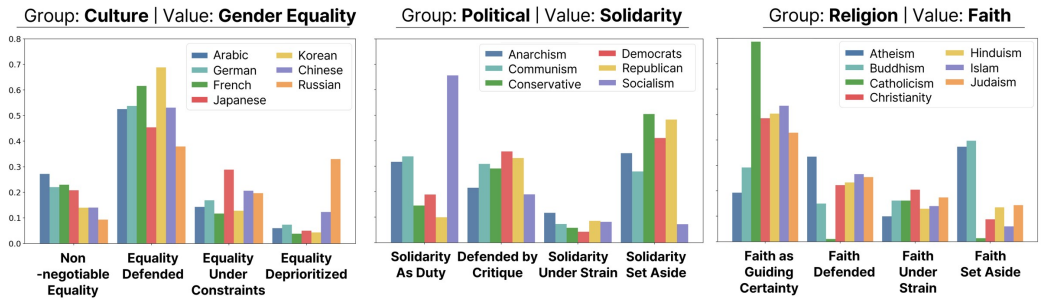


Figure 4: Base expression-mode distributions for selected values and groups.

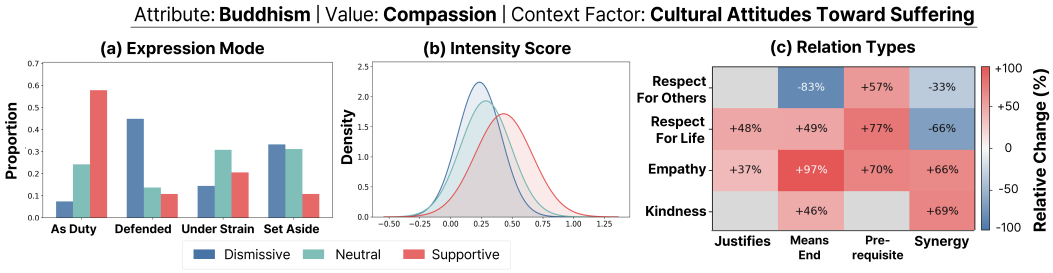


Figure 5: Contextual variation in *compassion* within Buddhism-associated discourse under different attitudes toward suffering: (a) expression-mode distributions, (b) intensity distributions, and (c) relative changes by partner value and relation type.

the mean OT distances reflect differences in expression-mode distributions, intensity scores, and value-relation structure across the compared context states.

The constructed profiles also reveal differences across groups that extend beyond the dominant expression mode. As shown in Figure 4, the language-associated profiles share *Equality Defended* as the dominant mode for *gender equality*, but differ in how the remaining probability mass is distributed across non-negotiable equality, constraints, and deprioritization. The profiles for *solidarity* in political discourse and *faith* in religious discourse also differ in their dominant modes across groups.

Pairwise EMD in Figure 6 further shows that the selected language-associated profiles are closer to one another than to the political or religious profiles. These comparisons involve differences in both group attributes and corpus sources. Section 6.3 examines group differences within Reddit and compares profiles across Reddit and CulturaX.

Turning to contextual variation, Figure 5 shows differences in both expression mode and intensity. For *compassion* within Buddhism-associated discourse, the dominant expression mode changes from *Defended* under dismissive contexts to *As Duty* under supportive ones, accompanied by a rightward intensity shift. Both modes affirm compassion: *Defended* does so by criticizing indifference, whereas *As Duty* presents compassion as an obligation.

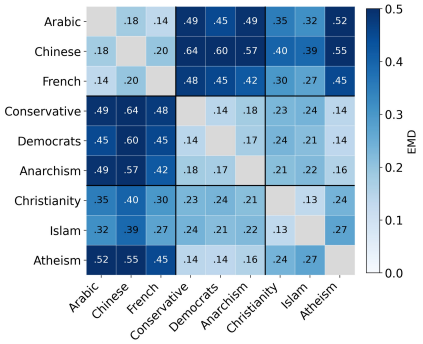


Figure 6: Cross-attribute divergence (EMD) in expression modes for selected groups.

5 Actionability of Distributional Profiles

The preceding section shows that distributional value profiles capture structured variation across groups and contexts. We next test whether a model can predict held-out profiles and contextual shifts, and whether the constructed expression profiles can guide model responses toward survey distributions.

5.1 Learning to Infer Value Profiles

Table 5: Evaluation on GlobalOpinionQA and OpinionQA. Lower is better for all metrics.

	GlobalOpinionQA						OpinionQA					
	Qwen3-8B		GLM-4-9B		Gemma-4-E4B		Qwen3-8B		GLM-4-9B		Gemma-4-E4B	
Method	KLD	JSD	KLD	JSD	KLD	JSD	KLD	JSD	KLD	JSD	KLD	JSD
Plain	5.58	0.25	1.70	0.22	3.02	0.28	4.32	0.21	1.61	0.18	2.16	0.20
Persona prompt	5.80	0.23	1.32	0.16	2.14	0.19	5.56	0.24	1.27	0.16	2.16	0.20
Distributional prompt	4.72	0.26	0.55	0.10	1.40	0.16	4.14	0.22	0.89	0.14	1.83	0.18
CultureLLM	2.47	0.21	1.76	0.18	9.37	0.26	2.94	0.21	1.97	0.20	11.01	0.29
CultureSPA	0.62	0.11	0.69	0.11	0.78	0.12	1.34	0.17	1.34	0.16	5.24	0.23
CultureBank	4.04	0.27	1.00	0.15	4.01	0.22	4.83	0.28	1.21	0.18	4.10	0.24
SimLLMCultureDist	3.01	0.31	1.28	0.17	5.75	0.33	2.97	0.31	1.91	0.18	6.08	0.34
Modular Pluralism	2.41	0.16	1.29	0.16	2.47	0.16	1.25	0.11	0.76	0.10	1.63	0.11
Ours	**0.56**	**0.11**	**0.25**	**0.07**	**0.34**	**0.08**	**0.58**	**0.11**	**0.35**	**0.08**	**0.46**	**0.10**

We train a lightweight profiling model that takes an attribute and a value item as input, together with descriptions of the context factor. It predicts base expression distributions, intensity summaries, and contextual OT distances. The model uses a frozen Qwen3-4B with task-specific prediction heads. Training uses the constructed profiles as supervision, and we evaluate on held-out profile records within each attribute. This tests prediction within the constructed value and context space. Implementation details are in Appendix C. Table 4 reports prediction errors for base profiles and contextual shift magnitudes on held-out records within each attribute.

Table 4: Profiling model performance on held-out base distributions and context shifts.

Component	Metric	Error
Base Distribution	Discrete KL ↓	0.07
	Intensity MAE ↓	0.09
Context Shift	Discrete OT MAE ↓	0.08
	Intensity OT MAE ↓	0.03

5.2 Using Profiles as Steering Targets

We next evaluate whether the constructed expression-state distributions, aggregated across contexts, can serve as targets for distributional alignment. Given a target profile, we construct auxiliary prompts with profile-derived expression-state options and target probabilities. We then optimize compact attribute-specific activation shifts on these prompts while keeping the model fixed. We train the steering vectors once per attribute using only the profile-derived prompts and reuse them across survey queries. The survey distributions serve only as evaluation targets.

We evaluate on GlobalOpinionQA [9] and OpinionQA [43] using attributes shared with our profiles, and compare predicted answer distributions against survey distributions using KLD and JSD. We compare against prompt-based methods (role prompting [29], distributional prompting), culture-aware methods (CultureLLM [27], CultureSPA [59], CultureBank [50]), and distribution-based methods (Modular Pluralism [10], SimLLMCultureDist [6]). All methods are evaluated on Qwen3-8B, GLM-4-9B, and Gemma-4-E4B. Full details are provided in Appendix D.

Table 5 shows that prompting-based methods improve fit inconsistently, while culture-aware tuning varies substantially across models. Across both benchmarks and all three models, profile-based steering yields the lowest reported KLD and the lowest or tied-lowest reported JSD. The steering vectors achieve this fit using targets derived from value expression in the sampled corpora, with survey responses used only for evaluation. Along with the held-out prediction results, this suggests that the profiles capture patterns of value expression that a model can both learn from descriptions and use to guide its responses on a different task.

6 Analyses

6.1 Profile Analyses

Contextual Shift. Figure 7(a) summarizes the OT distances between intensity distributions for pairs of context states. Smaller shifts are more frequent, with a tail of larger shifts. Across attribute categories, mean contextual OT distances are about 0.3 for expression modes and 0.1 for intensity (Table 3). Mean intensity OT is highest for resource constraints, followed by time pressure and psychological health (Table 16). For expression modes, using categorical or semantic costs in

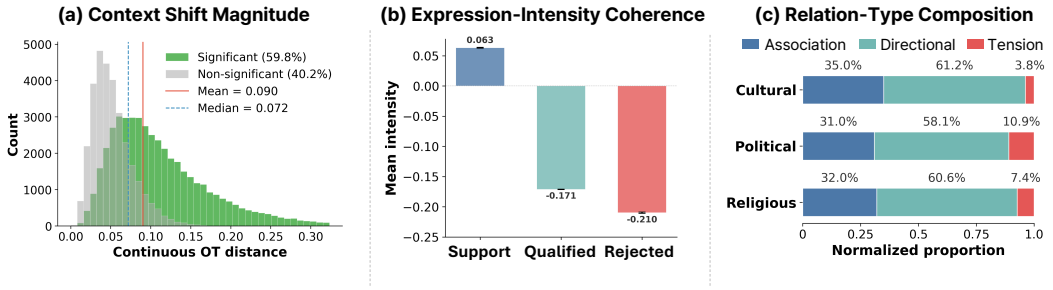


Figure 7: Profile analyses illustrating (a) contextual shifts in intensity, (b) mean intensity by expression category, and (c) relation-type frequencies by group category. In (a), colors indicate whether the 95% bootstrap interval for the change in mean intensity excludes zero.

place of the ordinal cost largely preserves the overall ranking of shift magnitudes, although the individual comparisons with the largest shifts vary with the cost. Examples of individual shifts and the comparison of transport costs appear in Appendix E.2.3.

Relation between Expression Mode and Intensity. We estimate expression mode and intensity separately. To examine how they relate, we group expression states into three categories (*affirmed*, *contested*, and *rejected*) and compare their mean intensity. As shown in Figure 7(b), mean intensity decreases from affirmed through contested to rejected expression. We reverse the expression-mode index so that higher scores indicate stronger endorsement on both measures. Across context pairs with nonzero changes in both means, the two measures agree in direction in 75.2% of cases. Intensity summarizes how strongly a passage endorses a value, while the finer modes distinguish how that stance is framed. For example, defending compassion and treating it as a duty both affirm the value (Figure 5).

Value Relation Structure. We then analyze inter-value relations across group categories. As shown in Figure 7(c), directional relations (e.g. *Justifies*) account for about 60% of annotated relations in each category. These relations describe cases in which one value serves as a condition, means, or reason for another. Tension relations are more frequent in the sampled political discourse than in the cultural discourse (10.9% vs. 3.8%). Examples include *capitalism* versus *economic justice* in political discourse and *sovereignty* versus *international cooperation* in cultural discourse.

6.2 Validation of Constructed Profiles

Human Evaluation. We conduct a human evaluation to assess the quality of the annotations, covering 1.8K samples across expression modes, context factors, and value relations. The evaluation involves approximately 280 participants recruited through Prolific and in-person annotation, with a target of at least 3 evaluators per sample, each fluent in the target language. Table 6 reports results by attribute. The reported overall score for expression labeling is 0.80, increasing to 0.89 under a one-level tolerance, while context and relation labeling have reported overall scores of approximately 0.75-0.85. These results indicate reliable annotation quality across domains, with most disagreements limited to adjacent labels, supporting the sentence-level annotations used to construct the profiles. Details are provided in Appendix E.1.

Table 6: Human evaluation results.

Attribute	Expression		Context	Relation
	Acc	PM1	Acc	Acc
Cultural	0.72	0.89	0.87	0.74
Political	0.85	0.90	0.83	0.75
Religious	0.81	0.88	0.81	0.73

Soft Profiles. Our primary profiles count one expression label per sentence-value pair, although human annotators can select different modes for the same pair. We examine whether representing this disagreement changes which mode dominates a profile and how far apart the profiles are across languages. Using the human judgments, we estimate a distribution of responses for each model-assigned state, shared across values and languages. We then average these probabilities within each language-value pair to obtain a soft profile. We compare hard and soft profiles on the same sentence-value pairs and normalize both over three invoked modes (Support, Qualified, and Rejected).

Across 50 values shared by six languages, *Support* is the most probable invoked mode in all 300 soft profiles. Of the 299 profiles with a unique dominant mode under both estimates, 169 shift from *Qualified* to *Support*. Differences across languages remain in the probability assigned to each mode, although they are typically smaller under the shared response model. For comparisons of the same value across languages, the median ratio of soft to hard total variation distance over invoked modes is 0.26 (Appendix E.3).

6.3 Source Dependence

The cultural profiles draw on CulturaX web text, whereas the political and religious profiles draw on Reddit. Differences across these categories may therefore partly reflect corpus source. We first examine whether category differences are also visible within a single platform by comparing expression profiles across 19 national, political, and religious Reddit communities. Mean Jensen-Shannon divergence is higher between communities in different categories than between communities in the same category, even after adjusting for the number of values shared by each pair.

We next compare how the same values are expressed in country-associated samples from CulturaX and Reddit. As shown in Table 7, profiles for the same country differ more across sources, on average, than profiles for different countries within a source. Across sources, the mean distance is nearly the same whether the countries match or differ.

Table 7: Mean total variation (TV) between profiles over the four invoked expression states for six country-associated groups in CulturaX and Reddit.

Country	Source	Mean TV
Same	Different	0.269
Different	Different	0.271
Different	Same	0.095

We also test whether accounting for source improves country matching on a separate panel of 38 values. For each profile, we compare a same-country and a different-country counterpart from the other source. Pooling information from other countries reduces the mean distance between same-country profiles across sources relative to uniform smoothing. A further source adjustment leaves the rate of selecting the same-country counterpart unchanged for one annotation procedure and slightly lower for the other two (Appendix E.4).

Further Analyses. We compare value vocabularies and expression labels across models, test sensitivity to confidence filtering and document-level aggregation, and compare group rankings with those derived from model judgments of stance toward five Moral Foundations (Appendix E.5). For contextual shifts, we examine how closely rankings by total variation distance agree across document splits and under alternative groupings of expression modes (Appendix E.2.3). For steering, we map free-text responses back to survey options and evaluate agreement with the survey distributions (Appendix D.4). We also compare steering trained on expression, intensity, and relation profiles, separately and in combination (Appendix D.5). Finally, we assess variation across training seeds and test how steering changes when the learned shift is applied at a subset of layers (Appendix D.6).

7 Conclusion

In this paper, we introduce **Distributional Value Profiling**, a framework for representing group-level value expression as structured distributions over multiple expression dimensions. Rather than reducing alignment targets to scalar scores or canonical labels, the framework preserves within-group variation in how values are expressed, emphasized, and related to one another. Across 21 corpora spanning cultural, political, and religious discourse, the profiles average roughly three effective expression modes per value and link values to multiple partners. The same value can take different expression modes, levels of emphasis, and relational roles across groups, while the relative frequencies of its expression modes vary across observed contexts within a group. Beyond descriptive analysis, a profiling model predicts base profiles and contextual shift magnitudes for held-out records within each attribute, and using expression profiles as steering targets improves average agreement with survey-response distributions at inference time. Overall, this work moves value alignment from selecting a representative group preference toward modeling the broader landscape of value expression, offering a foundation for alignment systems that can account for plurality, context, and relational structure rather than averaging them away.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) grant (RS-2025-00560762), the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants (RS-2025-25442338, RS-2026-25518317, RS-2026-25551943, RS-2021-II211343), and the National Supercomputing Center at KISTI. J. Do is with ASRI, Seoul National University.

References

[1] Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies, 2023.

[2] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 2023.

[3] Anat Bardi and Shalom Schwartz. Values and behavior: Strength and structure of relations. *Personality and social psychology bulletin*, 29, 2003.

[4] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *arXiv:2001.08435*, 2020.

[5] Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. High-Dimension Human Value Representation in Large Language Models. In *Proc. of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025.

[6] Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. In *Proc. of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025.

[7] Haoxian Chen, Hanyang Zhao, Henry Lam, David Yao, and Wenpin Tang. MallowsPO: Fine-tune your LLM with preference dispersions. In *Proc. of Int'l Conf. on Learning Representations (ICLR)* , 2025.

[8] Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. In *arXiv:2507.21509*, 2025.

[9] Esin DURMUS, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*, 2024.

[10] Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[11] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. *Chapter Two - Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism*, volume 47 of *Advances in Experimental Social Psychology*. 2013.

[12] Hanze Guo, Jing Yao, Xiao Zhou, Xiaoyuan Yi, and Xing Xie. Counterfactual reasoning for steerable pluralistic value alignment of large language models. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2025.

[13] Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Juan Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. World values survey: Round seven – country-pooled datafile version 6.0, 2022.

[14] Jonathan Haidt. *The righteous mind: Why good people are divided by politics and religion*. 2012.

[15] Jongwook Han, Dongmin Choi, Woojung Song, Eun-Ju Lee, and Yohan Jo. Value portrait: Assessing language models' values through psychometrically and ecologically valid items. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[16] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning AI with shared human values. In *Proc. of Int'l Conf. on Learning Representations (ICLR)* , 2021.

[17] Geert Hofstede and Michael H Bond. Hofstede's culture dimensions: An independent validation using rokeach's value survey. *Journal of cross-cultural psychology*, 15, 1984.

[18] EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. Aligning language models to user opinions. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

[19] Saki Imai, Pedram Heydari, Anthony Sicilia, Asteria Kaeberlein, Katherine Atwell, and Malihe Alikhani. Mixdpo: Modeling preference strength for pluralistic alignment. In *arXiv:2601.06180*, 2026.

[20] Haoran Jin, Meng Li, Xiting Wang, Zhihao Xu, Minlie Huang, Yantao Jia, and Defu Lian. Internal value alignment in large language models through controlled value vector activation. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[21] Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez Adauto, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, and Bernhard Schölkopf. Language model alignment in multilingual trolley problems. In *Proc. of Int'l Conf. on Learning Representations (ICLR)* , 2025.

[22] Gauri Kambhatla, Sanjana Gautam, Angela Zhang, Alex Liu, Ravi Srinivasan, Junyi Jessy Li, and Matthew Lease. Improving the distributional alignment of llms using supervision. In *arXiv:2507.00439*, 2026.

[23] Dongjun Kang, Joonsuk Park, Yohan Jo, and JinYeong Bak. From values to opinions: Predicting human behaviors and stances using value-injected large language models. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[24] Woojin Kim, Sieun Hyeon, Jusang Oh, and Jaeyoung Do. Valueflow: Toward pluralistic and steerable value-based alignment in large language models. In *arXiv:2602.03160*, 2026.

[25] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Michael Bean, Katerina Margatina, Rafael Mosquera, Juan Manuel Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2024.

[26] Oliver Klingefjord, Ryan Lowe, and Joe Edelman. What are human values, and how do we align ai to them? In *arXiv:2404.10636*, 2024.

[27] CHENG LI, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. CultureLLM: Incorporating cultural differences into large language models. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2024.

[28] Elinor Mason. Value Pluralism. In *The Stanford Encyclopedia of Philosophy*. 2023.

[29] Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the International Conference on Computational Linguistics*, 2025.

[30] Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. Benchmarking distributional alignment of large language models. In *Proc. of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025.

[31] Rustin Meyer, Reeshad Dalal, and Richard Hermida. A review and synthesis of situational strength in the organizational sciences. *Journal of Management*, 36, 2010.

[32] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. In *arXiv:2402.06196*, 2025.

[33] Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024.

[34] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023.

[35] Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. Valuenet: A new dataset for human value driven dialogue system. In *Proc. of Int'l Conf. on Artificial Intelligence (AAAI)*, 2022.

[36] Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free RLHF. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2024.

[37] John Rauthmann, David Gallardo-Pujol, Esther Hanes, Elysia Todd, Christopher Nave, Ryne Sherman, Matthias Ziegler, A. Bell Cooper, and David Funder. The situational eight diamonds: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, 107, 2014.

[38] Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. ValueBench: Towards Comprehensively Evaluating Value Orientations and Understanding of Large Language Models. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[39] Sonia Roccas and Lilach Sagiv, editors. *Values and Behavior: Taking a Cross Cultural Perspective*. Springer International Publishing, 2017.

[40] Milton Rokeach. *The nature of human values.* 1973.

[41] Naama Rozen, Liat Bezalel, Gal Elidan, Amir Globerson, and Ella Daniel. Do LLMs have consistent values? In *Proc. of Int'l Conf. on Learning Representations (ICLR)* , 2025.

[42] Giuseppe Russo, Debora Nozza, Paul Röttger, and Dirk Hovy. The pluralistic moral gap: Understanding moral judgment and value differences between humans and large language models. In *Proc. of the European Chapter of the Association for Computational Linguistics (EACL)*, 2026.

[43] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2023.

[44] Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. Evaluating the Moral Beliefs Encoded in LLMs. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2023.

[45] Shalom Schwartz. *The Refined Theory of Basic Values*. 2017.

[46] Shalom Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, Ozlem dirilen gumus, and Mark Konty. Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103, 2012.

[47] Shalom H. Schwartz. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. volume 25 of *Advances in Experimental Social Psychology*. 1992.

[48] Shalom H. Schwartz and Anat Bardi. Value hierarchies across cultures: Taking a similarities perspective. *Journal of Cross-Cultural Psychology*, 32, 2001.

[49] Shalom H Schwartz and Klaus Boehnke. Evaluating the structure of human values with confirmatory factor analysis. *Journal of research in personality*, 38, 2004.

[50] Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024.

[51] Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. Value kaleidoscope: engaging AI with pluralistic human values, rights, and duties. In *Proc. of Int'l Conf. on Artificial Intelligence (AAAI)*, 2024.

[52] Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel A. Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. Value profiles for encoding human variation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

[53] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A Roadmap to Pluralistic Alignment. In *Proc. of Int'l Conf. on Machine Learning (ICML)*, 2024.

[54] Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[55] Philip E. Tetlock. Thinking the unthinkable: sacred values and taboo cognitions. *Trends in Cognitive Sciences*, 7, 2003.

[56] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. In *arXiv:2308.10248*, 2024.

[57] Bas Verplanken and Rob Holland. Motivated decision making: Effects of activation and self-centrality of values on choices and behavior. *Journal of Personality and Social Psychology*, 82, 2002.

[58] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18, 2024.

[59] Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. Self-pluralising culture alignment for large language models. In *Proc. of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2025.

[60] Jing Yao, Shitong Duan, Xiaoyuan Yi, Dongkuan Xu, Peng Zhang, Tun Lu, Ning Gu, Zhicheng Dou, and Xing Xie. AdAEM: An adaptively and automated extensible measurement of LLMs' value difference. In *Proc. of Int'l Conf. on Learning Representations (ICLR)* , 2026.

[61] Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. Value FULCRA: Mapping Large Language Models to the Multidimensional Spectrum of Basic Human Value. In *Proc. of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2024.

[62] Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. From instructions to intrinsic human values – a survey of alignment goals for big models. In *arXiv:2308.12014*, 2023.

[63] Jing Yao, Xiaoyuan Yi, and Xing Xie. CLAVE: An Adaptive Framework for Evaluating Values of LLM Generated Responses. In *Proc. of Neural Information Processing Systems (NeurIPS)*, 2024.

[64] Haoran Ye, Yuhang Xie, Yuanyi Ren, Hanjun Fang, Xin Zhang, and Guojie Song. Measuring human and AI values based on generative psychometrics with large language models. In *Proc. of Int'l Conf. on Artificial Intelligence (AAAI)*, 2025.

[65] Haoran Ye, TianZe Zhang, Yuhang Xie, Liyuan Zhang, Yuanyi Ren, Xin Zhang, and Guojie Song. Generative Psycho-Lexical Approach for Constructing Value Systems in Large Language Models. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[66] Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024.

[67] Shenyan Zheng, Jiayou Zhong, Anudeex Shetty, Heng Ji, Preslav Nakov, and Usman Naseem. Vispa: Pluralistic alignment via automatic value selection and activation. In *arXiv:2601.12758*, 2026.

[68] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. In *arXiv:2310.01405*, 2025.

Supplementary Material

Contents

A Related Works **18**

A.1 Human Values 18

A.2 Human Values in LLMs 18

A.3 Distributional Alignment 18

B Distributional Value Profiles **20**

B.1 Data 20

B.2 Profile Construction 21

B.3 Contextual Shift 24

C Profiling Model **28**

C.1 Dataset 28

C.2 Modeling 28

C.3 Staged Training 28

C.4 Examples 29

D Distributional Steering **31**

D.1 Method 31

D.2 Baselines 31

D.3 Evaluations 32

D.4 Free-Text Survey Responses 32

D.5 Steering with Other Profile Components 32

D.6 Training Seeds and Layer Contributions 33

E Additional Analyses **34**

E.1 Human Evaluation 34

E.2 Profile Analyses 34

E.3 Soft-Profile Sensitivity 44

E.4 Source Dependence 44

E.5 Sensitivity to Profile Construction 44

F Limitation and Impact Statement **46**

F.1 Limitation 46

F.2 Broader Impacts 46

F.3 Ethical Concerns & Safeguards 46

G License **47**

H LLM Usage **47**

A Related Works

A.1 Human Values

Human values are commonly understood as desirable, trans-situational goals that guide how people evaluate actions and events [47, 40]. Rokeach [40] distinguishes terminal values, which refer to desired end states, from instrumental values, which refer to preferred modes of conduct. Schwartz organizes values in a circular structure according to the motivations they express. Neighboring values tend to be compatible, while pursuing values on opposite sides of the circle can bring their goals into conflict [47, 49]. The refined theory preserves this continuum while distinguishing more specific motivations [46, 45]. Other traditions examine cultural dimensions such as individualism and power distance [17], moral domains such as care, fairness, and loyalty [11, 14], or how the relative importance of values varies across cultures [48, 39].

Although values are relatively stable, their expression depends on the situation and the competing goals at stake. A value is more likely to guide a choice when it is activated and central to the self [57]. The relation between values and behavior also reflects which other values support or conflict with a particular course of action [3]. Situations differ in the cues they provide and in the constraints they place on behavior [37, 31]. Philosophical accounts of value pluralism further allow that competing values can be incommensurable [28]. In discourse, a speaker may defend a value in one passage and qualify it in another when a competing commitment takes precedence. Both passages invoke the value, but they express it in different ways.

A.2 Human Values in LLMs

Research on human values in LLMs uses several kinds of representation. Some approaches map text to predefined value taxonomies, including Schwartz's basic values, or annotate the values, rights, and duties at stake in a situation [35, 51, 61]. Others derive value vocabularies from model outputs [65] or learn high-dimensional representations to compare the values models express across languages [5]. A related approach constructs profiles for individual raters. Sorensen et al. [52] infer a natural-language profile from an individual's judgments and use it to predict how that person will rate new responses. Our profiles describe how a particular value is expressed across passages in a group-associated corpus.

A second line of work evaluates the values that models express. Structured questionnaires and moral scenarios give models a common set of questions and make their responses comparable across value dimensions [44, 38, 15]. Studies of consistency examine whether an inferred value profile holds when the elicitation setting changes [41]. Open-ended evaluation poses a related measurement problem because a model can express a value without naming it. Generative psychometrics and calibrated evaluators infer values from the content of a response [64, 63]. Adaptive assessments also vary the questions themselves to elicit differences between models [60].

A third line uses values as alignment targets. Prompting methods specify a perspective or value priority in natural language [24]. Training-based approaches incorporate value information to predict opinions or generate responses consistent with specified values [23, 61]. Guo et al. [12] model the dependencies among value priorities and use counterfactual reasoning to guide generation toward a specified combination. Other methods intervene on internal representations, using value vectors to steer generation at inference time [20, 67]. Our steering study builds on this last approach and uses the distribution of expression modes in a corpus to specify the target.

A.3 Distributional Alignment

Pluralistic alignment distinguishes several goals for model behavior [53]. *Overton* pluralism concerns the range of reasonable perspectives a model can present. *Steerable* pluralism concerns its ability to respond from a specified perspective, while *distributional* pluralism concerns how often different responses occur relative to a reference group. A model can produce all of a group's common answers yet assign them quite different probabilities from those observed in the group. Distributional benchmarks evaluate how closely those probabilities agree [30].

A major line of work evaluates distributional alignment using survey data. OpinionQA and GlobalOpinionQA compare model answer distributions with those of respondents across demographic groups or countries [43, 9]. These studies find that models reflect some groups' responses more closely than

others and that demographic prompting alone does not reliably recover the observed distributions. Follow-up work uses survey supervision and population-specific training to predict how answers are distributed within a group [54, 6, 22]. Related work studies how well language models simulate individual opinions or reproduce patterns in survey and experimental responses [1, 2, 18].

Another line starts from variation in human feedback. Dialogue studies collect individual judgments [25], while preference-based methods account for differences across groups, dispersion in judgments, or the strength of a preference [36, 7, 19]. Modular approaches combine models representing different communities or perspectives [10]. Russo et al. [42] compare distributions of human and model judgments on moral dilemmas and trace the values invoked in the accompanying rationales. DVP starts with a given value and examines how its expression modes, intensity, and relations with other values vary across contexts in the sampled discourse.

B Distributional Value Profiles

We describe how we collect and filter the corpora, annotate expression modes, intensity, and inter-value relations, and aggregate these annotations into base and context-conditioned profiles. We then describe how we measure contextual shifts using optimal transport. Our computing setup included eight NVIDIA A100 GPUs.

B.1 Data

B.1.1 Data Collection and Filtering Pipeline

For cultural attributes, we use seven language-specific web corpora from CulturaX [33]. We first isolate the main text by removing navigation menus, footers, login prompts, and other boilerplate. Language-specific filters remove spam and promotional material, gambling and adult content, excessive URLs, contact information, and likely machine translations.

For political and religious attributes, we use submissions and comments from the Pushshift Reddit archive [4], spanning June 2005 through December 2023. Within each subreddit, we match comments to submissions by submission ID and reconstruct each thread from its title, self-text, and comments in chronological order. We decode HTML entities, skip malformed JSON lines, and discard comments marked as deleted or removed. We also remove AutoModerator comments, replace usernames with thread-local identifiers such as `user_1`, and retain threads with at least five remaining comments.

For both sources, we segment the retained documents into sentences and preserve the surrounding text when available. We first identify sentences that express an opinion, norm, preference, or decision. We then extract candidate value items from these sentences for expression-state annotation, intensity parsing, and relationship extraction.

Table 8: Corpus statistics for the 21 target attributes. *Docs* counts CulturaX articles or Reddit threads, and *Sentences* counts sentences after text filtering. *Expressive* counts sentences retained by the expressive-language filter, and *Expr. %* gives their share of the *Sentences* column. *Value Sents.* counts sentences with at least one retained value assignment. *Value Items* counts all such assignments, including multiple assignments to the same sentence.

Group	Attribute	Docs	Sentences	Expressive	Expr. %	Value Sents.	Value Items
Cultural	Arabic	66,095	731,501	226,508	31.0	68,671	476,376
	Chinese	59,674	569,830	231,825	40.7	96,759	639,904
	French	70,737	885,547	239,807	27.1	85,080	513,959
	German	68,900	1,232,841	260,233	21.1	113,968	676,869
	Japanese	64,951	489,183	246,661	50.4	94,935	546,839
	Korean	65,272	846,175	299,462	35.4	91,226	588,926
	Russian	70,684	1,249,808	240,047	19.2	85,432	510,347
Political	Anarchism	39,002	218,081	183,528	84.2	107,213	729,124
	Communism	30,734	141,353	111,147	78.6	90,483	608,544
	Conservatism	100,000	379,459	299,404	78.9	79,304	500,637
	Democrats	100,000	355,342	234,481	66.0	84,897	496,319
	Libertarianism	49,801	226,003	189,782	84.0	83,601	548,229
	Republicans	60,973	269,321	190,845	70.9	77,357	479,030
	Socialism	41,757	211,936	179,599	84.7	95,641	655,564
Religious	Atheism	51,162	279,824	239,570	85.6	110,989	714,476
	Buddhism	62,057	504,896	438,541	86.9	162,014	1,144,983
	Catholicism	50,000	358,991	303,594	84.6	139,976	952,583
	Christianity	50,000	334,725	295,902	88.4	143,069	985,157
	Hinduism	50,000	251,452	156,797	62.4	90,869	605,700
	Islam	50,000	312,172	260,819	83.5	132,908	905,413
	Judaism	50,000	256,561	200,948	78.3	89,700	563,328

B.1.2 Corpus Statistics

Table 8 reports per-attribute statistics at each stage of the processing pipeline. The corpus sizes summarized in Section 4.1 range from 141,353 sentences for Communism to 1,249,808 for Russian

after text filtering. The expressive-language filter retains 19.2–50.4% of sentences in the CulturaX corpora and 62.4–88.4% in the Reddit corpora. Across all 21 corpora, 42.2% of expressive sentences have at least one recorded value assignment. Those sentences receive an average of 6.5 assignments each.

B.2 Profile Construction

B.2.1 Value Item Assignment

Expressive sentence filtering. We use `Qwen3-32B` to identify sentences that express an opinion or belief, a value or norm, a preference or trade-off, or a decision or intent. The filter returns an expressive type and a confidence score on a 0–100 scale. We retain sentences with an explicit stance cue and exclude purely factual reports, narrative descriptions, and headline fragments.

Value item extraction. For each retained sentence, we provide `Qwen3-32B` with the target sentence and its source text. The model returns up to eight value items ranked by relevance, each with a score from 0 to 100. Items may name values, norms, duties, rights, customs, or professional codes. We retain assignments with relevance scores above 60. This yields over 100K raw value-item assignments per attribute before deduplication.

Deduplication and vocabulary construction. The extractor can return several names for the same value, such as *freedom of speech* and *free expression*. Within each attribute, we cluster value names by semantic similarity using `Qwen3-Embedding-8B` and select one phrase per cluster. This yields approximately 300 representative phrases per attribute. Human reviewers then merge near-duplicates across attributes while retaining group-specific concepts. The resulting vocabulary contains 1,946 value items, each with a short definition generated by GPT-5 (Table 9). We map each original assignment to its nearest neighbor in the shared vocabulary by embedding similarity.

Table 9: Sample of representative value items and their definitions.

Value Item	Definition
academic freedom	The principle that scholars and students should be free to teach, learn, and communicate ideas without undue interference or censorship.
compassion	Concern for others' suffering accompanied by a desire to alleviate it; motivates caring and humane responses.
economic fairness	The principle that economic benefits, burdens, and opportunities should be distributed justly, with fair treatment in wages, taxation, and access.
personal autonomy	The principle that individuals should be free to make their own choices and govern their lives without undue coercion or interference.
religious devotion	A commitment to religious faith expressed through worship, practice, and adherence to spiritual teachings.
respect for elders	The principle of showing deference and care toward older people, acknowledging their experience, wisdom, and dignity.

B.2.2 Expression State Annotation

State schema generation. For each value item, we prompt `Qwen3-30B-A3B` with its name, definition, and a few examples to generate a set of expression states. The four invoked modes distinguish passages that treat the value as non-negotiable, defend it by criticizing a violation, uphold it under difficulty, or explicitly set it aside. A separate label, *Not Invoked*, applies when the passage makes no appeal to the value. The state names and cues are specific to each value. For compassion, for example, a sentence can report suffering without invoking compassion, or it can condemn indifference to that suffering (Table 10).

Auditing and revision. GPT-5, GPT-OSS-120B, and Qwen3-Next-80B independently review each proposed schema. On a 1–5 scale, they score how readily the states can be recognized from the text, how clearly they are separated, how well they cover plausible modes of expression, how much their

interpretation depends on context, and the quality of their textual cues. Human reviewers revise schemas that fail a criterion or receive a flag of medium or higher severity, paying particular attention to overlapping state descriptions.

Table 10: Complete expression schemas for three value items: four invoked modes and a separate *Not Invoked* state. The value-specific descriptions define the annotation categories.

Value Item	State	Description
Compassion	Compassion as Duty	Compassion treated as a non-negotiable guiding principle; active care for others is urged or modeled.
	Condemned Lack	Compassion affirmed by criticizing cruelty or indifference as moral failures.
	Compassion Under Strain	Compassion endorsed but presented as difficult to sustain due to fear, burnout, or competing demands.
	Compassion Set Aside	Compassion minimized, rejected, or overridden in this situation, prioritizing punishment or self-interest.
	Not Invoked	Text does not position or appeal to compassion even if hardship is mentioned.
Fairness	Fairness as Nonnegotiable	Fairness treated as a primary standard; outcomes or procedures must be equitable without exception.
	Affirmed by Critique	Fairness endorsed by condemning people or systems as unfair, biased, or rigged.
	Fairness Under Constraints	Fairness affirmed but framed as difficult due to limited information, competing needs, or practical trade-offs.
	Fairness Traded Off	Fairness explicitly set aside in this situation; another goal is prioritized.
	Not Invoked	Text does not position or appeal to fairness even if decisions or outcomes are discussed.
Honesty	Non-Negotiable Honesty	Honesty framed as a core duty; truth is told even when costly.
	Defended by Condemnation	Honesty affirmed by criticizing deception, hypocrisy, or cover-ups.
	Honesty Under Pressure	Honesty endorsed but presented as difficult to maintain due to fear, incentives, or social consequences.
	Honesty Set Aside	Deception or withholding justified or normalized in this situation.
	Not Invoked	Text does not position or discuss honesty or deception even if communication is mentioned.

Expression state tagging. For each sentence–value pair, `Qwen3-30B-A3B` receives the target sentence, the value name and definition, and the finalized schema with its textual cues. When available, the surrounding text helps resolve references in the target sentence. The model selects one state and returns an integer confidence score on a 0–100 scale. After confidence filtering, we exclude *Not Invoked* and normalize the counts over the four invoked modes within each attribute–value pair. The resulting profile describes how the value is expressed when it is invoked.

B.2.3 Intensity Parsing

Value vector construction. We estimate signed intensity using a value-specific direction in a model's hidden states [68, 56, 8]. For each of the 1,946 values, `Qwen3-32B` generates positive seed sentences that endorse the value and negative seed sentences that oppose or dismiss it. Each seed contains one clause of at most 30 words, and the prompts vary the speaker, role, and setting. Let $\mathbf{h}_x^{(m,\ell)}$ denote the mean hidden state of the last three tokens of sentence x at layer ℓ of model m. For positive and negative seed sets $\mathcal{P}_v$ and $\mathcal{N}_v$, the value direction is the difference between their mean representations,

$$\mathbf{d}_v^{(m,\ell)} = \frac{1}{|\mathcal{P}_v|} \sum_{x \in \mathcal{P}_v} \mathbf{h}_x^{(m,\ell)} - \frac{1}{|\mathcal{N}_v|} \sum_{x \in \mathcal{N}_v} \mathbf{h}_x^{(m,\ell)}. \tag{1}$$

Intensity score computation. For a target sentence x, we compute its cosine similarity to the direction for value v at each selected layer. We first average these similarities over the layers L_m listed for each model in Table 11,

$$u_{xv}^{(m)} = \frac{1}{|L_m|} \sum_{\ell \in L_m} \cos\left(\mathbf{h}_x^{(m,\ell)}, \mathbf{d}_v^{(m,\ell)}\right). \tag{2}$$

Within each model–value pair, we clip the layer-averaged scores to their 1st and 99th percentiles and scale them to $[-1, 1]$ by dividing by the larger absolute percentile when that scale is nonzero. The final intensity score is the equally weighted average of the four normalized model scores. Higher scores indicate expression closer to the endorsement seeds.

Table 11: Frozen models and layers used for intensity scoring. Index 0 denotes the embedding output, and index $\ell \geq 1$ denotes the output of transformer block ℓ.

Model	Total Layers	Layers Used
`Qwen3-32B`	64	16, 24, 32, 40, 48
`Qwen3-30B-A3B`	48	16, 24, 32, 40, 48
`GPT-OSS-20B`	24	8, 12, 16, 20
`GLM-4.7-Flash`	47	8, 16, 24, 32, 40

Scores for a given attribute and value form its empirical intensity distribution. For downstream modeling, we fit a Gaussian summary using the sample mean and standard deviation. The *Gauss.* column in Table 3 counts a profile as approximately Gaussian when it has at least 20 scores, absolute skewness at most 0.5, absolute excess kurtosis at most 1.0, and a Kolmogorov–Smirnov distance of at most 0.12 from the fitted normal.

B.2.4 Relationship Extraction

For each sentence with at least two assigned values, `Qwen3-30B-A3B` receives the sentence, its candidate values, and the surrounding text when available. For each extracted pair, it records one of the 11 relation labels in Table 12, a direction (`source_to_target`, `target_to_source`, or `none`), and a polarity (`supports`, `opposes`, `mixed`, or `neutral`).

We retain a relation when the sentence links the values explicitly, for example through a causal or conditional phrase, or when the link follows from the argument in the passage itself. A general association between the values is insufficient. An optional verification pass can remove an edge if the cited text does not support it. We then deduplicate the edges, check their labels, directions, and polarities against the schema, and discard edges whose endpoints are absent from the candidate list.

For the base profile of value v under attribute a, we estimate partner frequencies $p(v' \mid v, a)$ from value co-occurrences. Within each pair (v, v'), we normalize the relation counts to estimate $p(r \mid v, v', a)$. Their product gives the joint distribution over partner values and relation types. For a context-conditioned profile, we count edges involving v within the context state and normalize the joint counts of (v', r) directly.

B.2.5 Quality Control & Aggregation

Annotation quality filtering. GPT-OSS-120B evaluates each integrated annotation record. It scores the sufficiency of the target sentence, the relevance of the assigned values, the plausibility of the expression labels, context states, and relation edges (one score each), and the consistency of the annotations across fields. Each of the six scores is an integer from 1 to 10. A record enters the filtered dataset only if all six scores are at least 5. For profiling-model supervision (Section 5.1), we require at least 7 on every dimension.

Aggregation of expression-mode distributions. For each attribute–value pair, we exclude *Not Invoked* and count the assignments to each of the four invoked modes. We retain pairs with more than 200 invoked assignments. For each retained pair, $p(z \mid v, a)$ is the count for mode z divided by the total count across the four modes.

Table 12: Relation labels and their definitions.

Group	Relation Label	Definition
ASSOCIATION	CO-OCCURS	Both values are salient in the text, with no more specific interaction expressed.
	SYNERGY	One value reinforces or supports the other.
DIRECTIONAL	MEANS-END	One value serves as a means to achieve the other.
	PREREQUISITE	One value is necessary for the other.
	JUSTIFIES	One value provides a normative reason to permit or legitimize the other.
	PRIORITIZES	The text ranks one value above the other.
	RESOLVES-BY	One value helps resolve a tension involving the other.
	SUBSTITUTES	One value replaces the other in an either-or choice.
TENSION	CONFLICT	One value undermines or is incompatible with the other.
	TRADEOFF	The text weighs one value against the other or describes a sacrifice.
	CONSTRAINT	One value limits the scope of the other or the conditions under which it applies.

Aggregation of intensity distributions. We pool the sentence-level scores from Section B.2.3 within each attribute–value pair to obtain the empirical intensity distribution $p(s \mid v, a)$. We use the empirical scores for contextual comparisons and the fitted Gaussian summaries for downstream prediction.

Aggregation of value-relation distributions. For each anchor value v and retained partner v', co-occurrence counts determine the partner weight $p(v' \mid v, a)$. For pairs with annotated edges, the relative frequency of each relation label gives $p(r \mid v, v', a)$. The product of these quantities gives the base profile entry for (v', r).

For contextual comparisons, we count edges with source v, partner v', and relation label r within each context state. For a given anchor value and context factor, we select up to 20 partners by edge frequency pooled across states, requiring at least three pooled edges per partner. Within each state, we normalize the joint counts of (v', r) over this shared support.

B.3 Contextual Shift

The base profiles aggregate value expression across observed contexts. To measure contextual variation, we hold the attribute and value fixed and compare profiles across two states of the same factor. We first define the factors relevant to each value, then assign factor states to individual sentences and measure the resulting shifts using optimal transport.

B.3.1 Contextual Factor Construction

Context factor groups. Drawing on work in social psychology and value theory [37, 31, 46], we define 13 groups of situational factors (Table 13). These include material conditions such as resource availability, social and institutional pressures, the information available to a speaker, and the time horizon of a decision. The groups provide a common vocabulary across values, while the factors within each group are specific to the value under study.

Per-value-item factor generation. For each of the 1,946 value items, we prompt `Qwen3-30B-A3B` to propose factors and assign each to one of the 13 groups. A factor record contains a name, a one-sentence definition, 2–5 discrete states, and 4–10 lexical cues. The proposed polarity indicates whether stronger or weaker value expression is expected across the factor's states, or whether the pattern is mixed. The model proposed 19,398 factors, about ten per value item.

Multi-model auditing. GPT-5, GPT-OSS-120B, and Qwen3-Next-80B independently review each proposal. They assess whether the group fits the factor, whether the states describe distinguishable conditions along a coherent dimension, whether the proposed polarity fits those states, and whether

Table 13: The 13 context-factor groups used to generate value-specific factors.

Group	Definition
stakes	Magnitude of potential gains or losses in the situation.
time_pressure	Urgency or temporal constraints on decisions.
role_and_identity	Social roles, professional identities, or self-conceptions activated in context.
social_context	Social environment, including norms, peer expectations, and group dynamics.
institutional_constraints	Formal rules, laws, regulations, or organizational policies constraining behavior.
resource_constraints	Availability or scarcity of material, financial, or human resources.
risk_and_uncertainty	Uncertainty, ambiguity, or risk in the situation.
moral_salience_and_harm	Visibility and perceived severity of moral consequences or potential harm.
cultural_norms_and_sensitivity	Cultural expectations, taboos, or sensitivities relevant to the situation.
information_environment	Quality, availability, and framing of accessible information.
conflict_tradeoff_structure	Competing goals, values, or interests requiring trade-offs.
temporal_scope	Time horizon over which consequences are considered.
psychological_health	Mental and emotional state, including stress, burnout, or resilience.

the definition and lexical cues are clear. Human reviewers revise flagged records or discard factors that cannot be resolved.

Examples. Table 14 shows how a broad group translates into a value-specific factor. For work–life balance, the length of the working day is a time-pressure factor. For freedom of expression, the severity of censorship laws is an institutional constraint.

Table 14: Examples of value-specific factors and their allowed states. In the last row, *none* denotes the absence of censorship laws.

Value item	Group	Factor	States
work–life balance	time_pressure	work_hours_length	short, moderate, long
honesty	moral_salience_and_harm	personal_consequences	low, moderate, high
compassion	social_context	social_norms_of_empathy	low, moderate, high
freedom of expression	institutional_constraints	censorship_laws	none, mild, moderate, severe

B.3.2 Contextual Info Parsing

For each sentence–value–factor triple, Qwen3-30B-A3B receives the target sentence, the value name, the factor definition, and its allowed states. The parser selects the state best supported by the text and returns a confidence score from 0 to 100. If the factor is unmentioned or the evidence is ambiguous, the parser returns NONE. We retain NONE as a separate context category. This label describes the evidence for the factor, whereas *Not Invoked* describes whether the sentence appeals to the value itself.

B.3.3 Optimal Transport Formulation

Optimal transport framework. Let $\mu = p(\cdot \mid v, a, c_1)$ and $\nu = p(\cdot \mid v, a, c_2)$ denote the distributions for two states of a contextual factor. Given a ground cost d, the transport distance on finite support is

$$\mathcal{W}_d(\mu, \nu) = \min_{\gamma \in \Pi(\mu,\nu)} \sum_{i,j} \gamma_{ij}\, d(x_i, y_j),$$

where $\Pi(\mu, \nu)$ is the set of transport plans with marginals μ and ν. The ground cost depends on whether we are comparing expression modes, intensity scores, or value relations.

(1) Value expression modes (discrete OT). Let p_i and q_i be the probabilities of the four invoked modes under c_1 and c_2, respectively. We assign unit distance to adjacent modes in schema order, so $d(z_i, z_j) = |i - j|$. The Wasserstein-1 distance is then the sum of the absolute differences in cumulative probability across the three bin boundaries,

$$\mathcal{W}_{\text{disc}}(p, q) = \sum_{k=1}^{3} \left| \sum_{i=1}^{k} (p_i - q_i) \right|.$$

For uncertainty estimates, we use an empirical-Bayes Dirichlet prior with concentration vector $20\hat{\boldsymbol{\pi}}$, where $\hat{\boldsymbol{\pi}}$ is the vector of mode proportions pooled across the two states. We add this vector to the observed counts in each state and draw 2,000 independent pairs of posterior profiles. We report the mean distance and its 95% credible interval. To flag directional shifts, we use the signed difference in mean bin position, $\Delta_z = \sum_{i=1}^{4} i(p_i - q_i)$, and flag a pair when the 95% posterior interval for Δ_z excludes zero.

(2) Value intensity (continuous OT). For intensity scores on $[-1, 1]$, we use $d(s, s') = |s - s'|$. The one-dimensional Wasserstein distance is the integral of the absolute difference between the two quantile functions,

$$\mathcal{W}_1(\mu, \nu) = \int_0^1 \left| F_\mu^{-1}(u) - F_\nu^{-1}(u) \right| du.$$

For equally sized samples, we average the absolute differences between the sorted scores. For unequal sample sizes n_1, n_2, we use linearly interpolated sample quantiles on a midpoint grid of size $M = \max(n_1, n_2)$,

$$\widehat{\mathcal{W}}_{\text{int}} = \frac{1}{M} \sum_{k=1}^{M} \left| \widehat{F}_1^{-1}(u_k) - \widehat{F}_2^{-1}(u_k) \right|, \qquad u_k = \frac{k - \frac{1}{2}}{M}.$$

We bootstrap the scores independently within each state, sampling with replacement at the original sample size. From 2,000 draws, we report the mean and 95% percentile interval for the distance and for the signed mean difference $\Delta_s = \bar{s}_1 - \bar{s}_2$. The intensity-shift flag records whether the interval for Δ_s excludes zero.

(3) Value relationships (joint OT). For a fixed attribute a and value v, the joint profile $p(v', r \mid v, a, c)$ records how often v is linked to partner v' by relation r in context state c. Two such profiles can differ because the same partner is linked to v in a different way, or because a different partner appears more often. We assign a unit cost to moving one unit of probability mass to a different partner. For the same partner, changing the relation from r_i to r_j costs $0.5\, d_r(r_i, r_j)$.

We specify d_r by hand so that changes between similar relations cost less. Equation (3) gives the distances within two sets of labels. The rows and columns of D_{support} follow the order (PREREQUISITE, MEANS-END, JUSTIFIES, SYNERGY). For D_{tension}, the order is (CONFLICT, TRADEOFF, CONSTRAINT).

$$D_{\text{support}} = \begin{pmatrix} 0 & .35 & .50 & .55 \\ .35 & 0 & .45 & .50 \\ .50 & .45 & 0 & .60 \\ .55 & .50 & .60 & 0 \end{pmatrix}, \qquad D_{\text{tension}} = \begin{pmatrix} 0 & .35 & .45 \\ .35 & 0 & .45 \\ .45 & .45 & 0 \end{pmatrix}. \tag{3}$$

For example, the first row and second column of D_{support} give $d_r(\text{PREREQUISITE}, \text{MEANS-END}) = 0.35$. Changing between these labels for the same partner therefore costs $0.5 \times 0.35 = 0.175$ per unit of mass. All observed labels are included. Identical labels have distance zero, and distinct labels whose pairing is absent from both matrices have distance one.

We draw 200 pairs of joint profiles using a Dirichlet prior with concentration 20 and proportions pooled across the two context states. We average the draws separately for each state and compute the greedy transport approximation once, using the two posterior-mean profiles. Within each partner v', we match as much mass as possible, starting with the least costly label pairs. Let $G_{v'}$ be the cost of these matches and let $p_{v'}$ and $q_{v'}$ be the total probability assigned to that partner in the two profiles. The resulting distance is

$$\widetilde{\mathcal{W}}_{\text{joint}} = \underbrace{\sum_{v'} G_{v'}}_{\text{cost of changes in relation}} + \underbrace{\frac{1}{2} \sum_{v'} |p_{v'} - q_{v'}|}_{\text{cost of changes in partner}} .$$

The second term is the mass that must move between partners, multiplied by its unit cost. The factor of one half accounts for the fact that each transfer is counted once at the partner losing mass and once at the partner gaining it.

We also use the posterior draws to estimate intervals for changes in individual (v', r) probabilities and for the Jensen–Shannon divergence between the joint profiles. To summarize changes in partners alone or in relations alone, we compute the divergence between the corresponding marginals of the posterior-mean profiles.

Support pruning and filtering. For each component, we require at least 200 distinct supporting sentence keys for a value. Among pairs with observations in both states, we flag those with fewer than five observations in either state for expression or relations, or fewer than ten for intensity. For joint OT, we retain up to 20 partners with at least three edge events each, ranked by frequency pooled across the states of the anchor value and factor. Before normalizing a state pair, we remove partner–relation cells with zero counts in both states.

C Profiling Model

C.1 Dataset

We train the profiling model on sentence-level annotations and on the profiles constructed from them. For sentence-level training, each record contains the source attribute, the sentence, and its annotated values, expression states, context factors and states, and value relations. At this stage, we retain the 200 most frequent value items per attribute.

For profile prediction, a base record pairs an attribute and a value (a, v) with the distribution over the four invoked expression modes and the mean and standard deviation of the intensity scores. A context record adds a factor and an ordered pair of its states. Its targets include contextual expression probabilities and intensity summaries, the signed changes in expected mode index and mean intensity, and the OT distance for each component. We use context records with at least ten observations in each state.

C.2 Modeling

For profile prediction, a frozen `Qwen3-4B` encoder reads separate text segments for the attribute, value item, value metadata, context factor, and state pair. The last two segments are empty for a base record. We use masked mean pooling over the hidden states within each segment and pass the result through a segment-specific adapter. The model combines the attribute and value representations with the value metadata to form the base representation. For a context record, it also encodes the factor definition, its states, polarity and cues, and the two states being compared. To combine two vectors x and y, each composition module uses both vectors, their elementwise product, and their absolute difference.

The base heads predict the four expression-mode probabilities and the parameters of the Gaussian intensity summary. The context heads predict the signed mean change and the OT distance between the two states for each component. For example, probability can move among expression modes even when the expected mode index changes little. Auxiliary heads can also reconstruct contextual profiles by adjusting the base predictions.

C.3 Staged Training

Stage 1. Ontology grounding. We train a sentence-level extraction model on ten instruction formats derived from the structured annotations. The tasks include identifying values, labeling their expression modes, and identifying context factors and states, either from the sentence alone or given a candidate value or value list. We use a language-modeling loss on the target outputs.

Stage 2. Base profile prediction. We train the adapters, composition modules, and base heads on (a, v) records. For expression, the objective combines KL divergence from the target distribution to the prediction with a Huber loss on the expected mode index. For intensity, we use Huber losses on the mean and log standard deviation and a squared Wasserstein-2 loss between the predicted and target Gaussian summaries.

Stage 3. Joint training on base and context records. We continue training on both record types. On context records, Huber losses supervise the signed changes and OT distances. Auxiliary losses supervise the reconstructed profiles and encourage agreement between the directly predicted shifts and those implied by the reconstructions. We randomly omit value metadata on a subset of training records.

Evaluation. We evaluate on held-out profile records within each attribute. Values and context factors can appear in both training and evaluation. For base profiles, Table 4 reports KL divergence for expression-mode probabilities and mean absolute error (MAE) for mean intensity. For context records, it reports MAE between predicted and target OT distances, separately for expression and intensity.

C.4 Examples

The two demo views illustrate the difference between extracting an annotation from a sentence and predicting a distribution from an attribute and value (Figures 8 and 9).

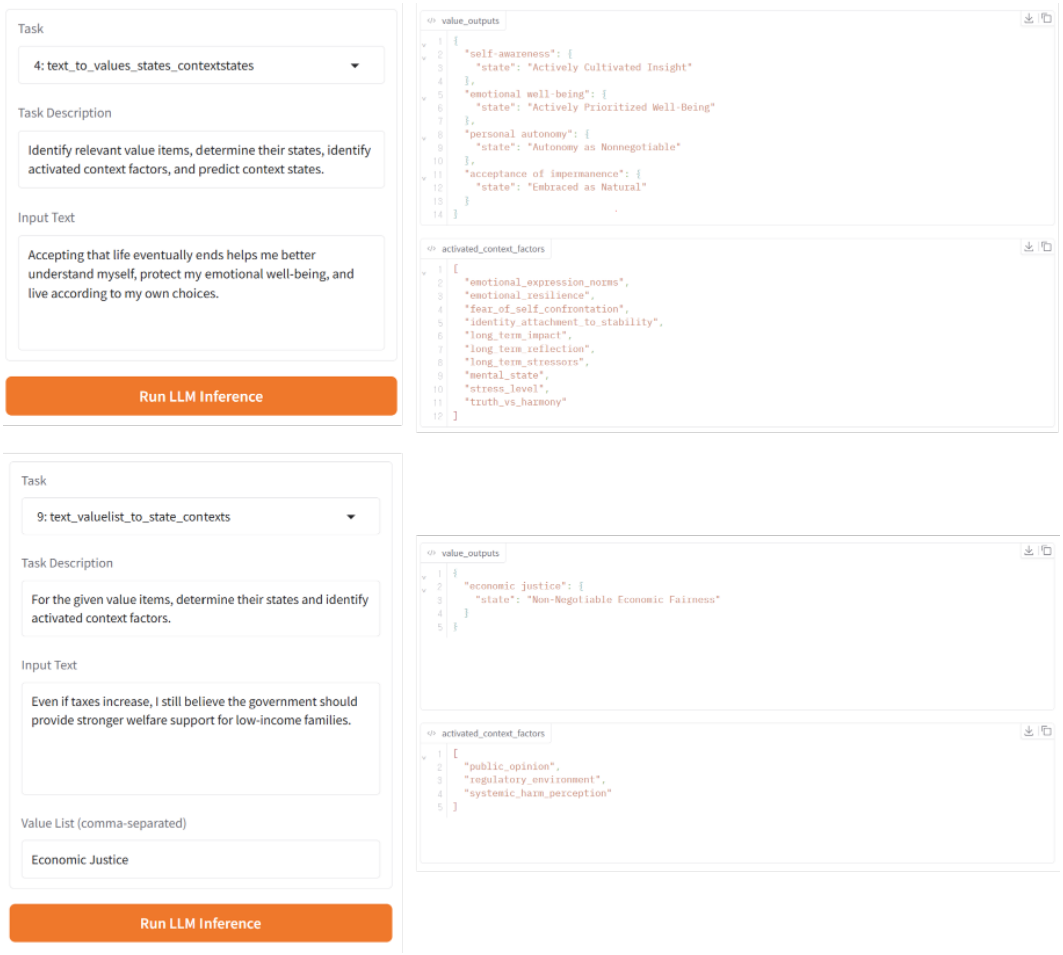


Figure 8: Sentence-level extraction. In the lower example, the speaker supports stronger welfare provision even if taxes rise. Given *economic justice* as the candidate value, the model assigns the expression state *Non-Negotiable Economic Fairness*.

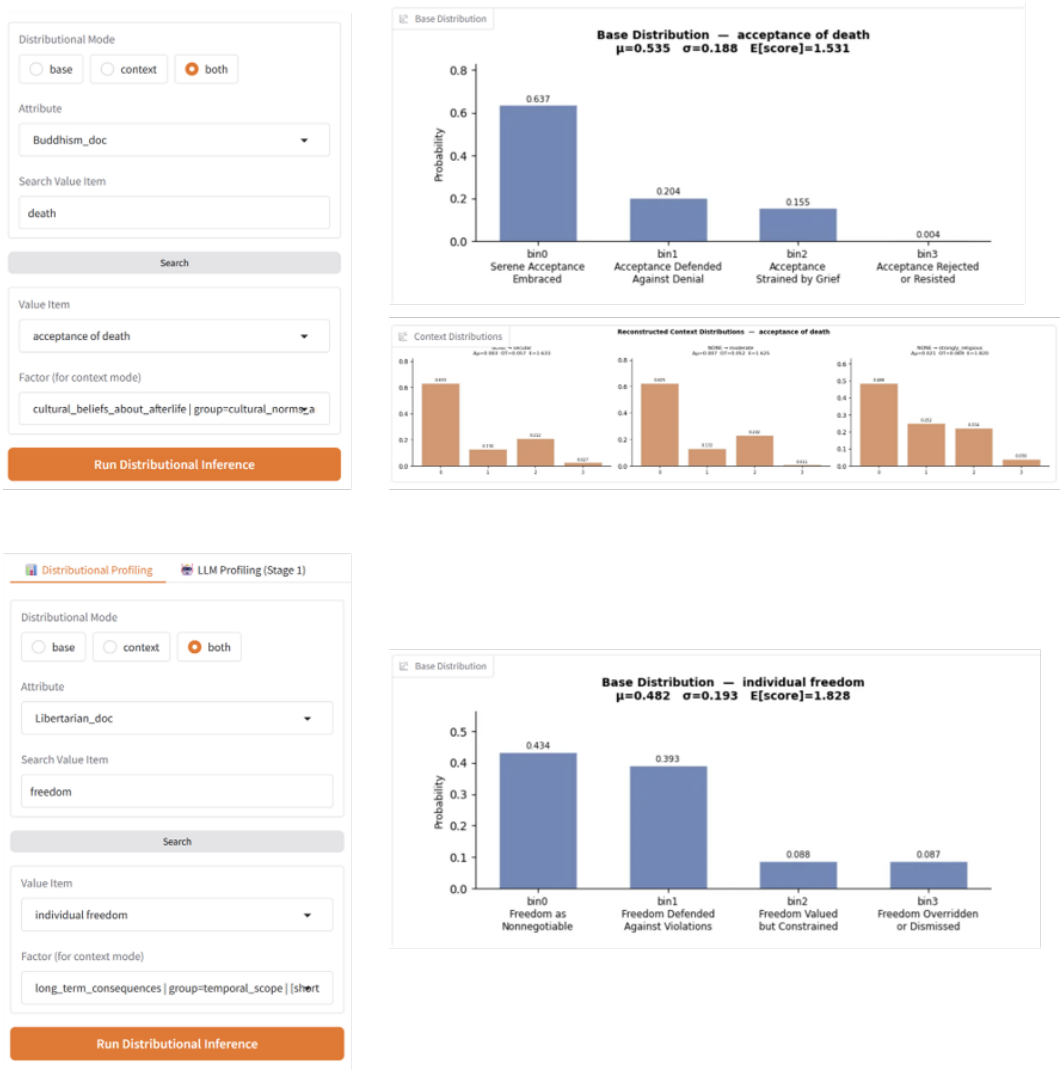


Figure 9: Profile prediction. For *acceptance of death* in Buddhism-associated discourse, the model assigns the highest base probability to *Serene Acceptance Embraced*. The panels below the base profile show how the prediction varies across states of the selected context factor.

D Distributional Steering

D.1 Method

We use the constructed distributional value profiles as supervision for lightweight activation-based steering. We train the shifts to match the expression-mode distributions in each attribute's profile while keeping the base model parameters fixed. For each attribute a, we learn a compact set of activation shifts $\{\Delta_\ell^{(a)}\}$ at selected transformer layers. These shifts are optimized once per attribute and then reused across all downstream queries for that attribute.

Distributional supervision from profiles. For each attribute, we select value items whose expression profiles place mass on more than one invoked mode. Each value item v is associated with a set of expression-state categories and their empirical proportions. We convert these profile entries into multiple-choice training prompts, where each option corresponds to one expression-state category. The target distribution over answer options is given by the empirical profile distribution.

To reduce positional artifacts, we randomly permute the order of expression-state categories across prompts. After permutation, the target probabilities are remapped to the corresponding option letters and renormalized. The prompt variants use category names alone, names with descriptions, or names with descriptions and textual cues.

Activation tuning. Given the generated training prompts, we optimize activation shifts while freezing the base model. Let $p_\theta(\cdot \mid x; \Delta^{(a)})$ denote the model's probability distribution over option letters when the attribute-specific shifts are applied. The optimization objective is

$$\mathcal{L}_{\text{steer}} = \mathrm{KL}\Big(q^\star(\cdot \mid a, v) \;\Big\|\; p_\theta(\cdot \mid x; \Delta^{(a)})\Big) + \lambda \sum_\ell \|\Delta_\ell^{(a)}\|_2^2.$$

The shifts are inserted through forward hooks at selected transformer blocks. A shift can be added to the last-token hidden state alone or to all token positions. In the experiments reported in the main paper, we use the last-token mode, since the prediction is obtained from the next-token distribution over answer letters.

Prediction with steering. At evaluation time, the learned shifts are loaded for the corresponding attribute and applied during the model forward pass. For each multiple-choice question, we compute the next-token logits for the option letters, restrict the distribution to the valid answer options, and apply a softmax over these selected logits. The resulting normalized option distribution is used as the model prediction.

D.2 Baselines

We compare against three groups of baselines.

Prompt-based baselines. The first group evaluates whether simple prompting can induce group-level response distributions without any parameter or activation modification. **Plain** uses the original question and asks the model to answer with a single option letter. **Persona prompt [29]** prepends an attribute description, such as a cultural, political, or religious identity description, before the question. **Distributional prompt** explicitly instructs the model to reflect group-level response distributions. For all prompt-based baselines, we compute answer probabilities directly from the next-token distribution over option letters.

Culture-aware tuning baselines. The second group includes prior tuning-based approaches for cultural or group-aware alignment. We implement and evaluate CultureLLM [27], CultureSPA [59], and CultureBank [50] under their corresponding training setups. These baselines use fine-tuning or culture-specific adaptation.

Distribution-based baselines. The third group includes methods that explicitly target group-level response distributions. Modular Pluralism [10] aggregates or composes outputs from separately adapted models, while SimLLMCultureDist [6] trains models to better match cultural response

distributions. Our steering vectors are trained on profile-derived prompts. We use survey response frequencies only for evaluation.

D.3 Evaluations

We evaluate distributional alignment on GlobalOpinionQA [9] and OpinionQA [43]. Each benchmark example consists of a question, a set of answer options, an attribute label, and a gold survey response distribution. We restrict evaluation to attributes that overlap with our constructed profiles. For GlobalOpinionQA, we apply shifts learned from language-associated corpora to the corresponding country-associated survey comparisons. For OpinionQA, we evaluate political ideology, political party, and religious groups that are shared with our profile attributes.

For each model and method, we compute a predicted distribution over answer options. Table 5 compares the predicted distribution q with the survey distribution p using KL divergence and Jensen–Shannon divergence (JSD) on Qwen3-8B, GLM-4-9B, and Gemma-4-E4B. For each attribute, our method trains one activation-shift file per backbone and reuses it across all questions for that attribute.

D.4 Free-Text Survey Responses

The main evaluation scores the probabilities of answer letters. To test whether steering also changes what the model says in its own words, we ask Qwen3-8B to answer 78 OpinionQA questions without showing the answer options. We use the Democrat, Republican, and Conservative groups and exclude questions with corrupted options or questions about personal behavior or experience. For each question, group, and condition, we sample 24 responses at temperature 1 and top-$p = 1$. A separate model, `Mistral-Small-24B-Instruct-2501`, maps each response to a survey option without seeing the generation condition. We compare the resulting option frequencies with the group's survey responses using natural-log JSD.

Table 15: Free-text OpinionQA evaluation with Qwen3-8B. We average JSD across the three groups within each question, then across questions. The final column gives the paired 95% question-bootstrap interval for the change from the unconditioned model.

Condition	Mean JSD	95% interval for change
Unconditioned	0.280	–
Distributional prompt	0.290	$[-0.015, \quad 0.036]$
Verbalized profile	0.271	$[-0.029, \quad 0.010]$
Persona prompt	0.250	$[-0.053, -0.007]$
Profile-based steering	0.241	$[-0.056, -0.023]$

The verbalized-profile condition gives the model the expression-mode probabilities for values relevant to the question. Steering uses activation shifts trained on 100 value items per attribute. With the questions, decoding settings, and judge held fixed, steering reduces mean JSD by 0.039 relative to the unconditioned model (Table 15, Holm-adjusted $p < 0.001$). Steering has the lowest mean JSD overall, while persona prompting gives a closer fit for the Conservative group (0.256 versus 0.274 for steering).

D.5 Steering with Other Profile Components

The shifts in Table 5 use expression-mode targets. In a separate Qwen3-8B experiment on nine OpinionQA attributes, we train shifts using expression probabilities, intensity probabilities over 11 bins, or probabilities over eight relation types for a value pair. In this comparison, the Liberal and Protestant survey categories are paired with the Libertarian and Christianity discourse profiles, respectively. At 2,000 optimization steps, mean JSD is 0.094 for expression, 0.077 for intensity, and 0.098 for relations. Combining all three targets gives 0.074 ($p = 0.008$ against expression alone, paired across attributes). Each individual component supplies 600 prompts per attribute, whereas the combined condition uses all 1,800. Extending combined training to 6,000 steps gives 0.079.

D.6 Training Seeds and Layer Contributions

Across three training seeds, expression-based steering gives mean JSDs of 0.091, 0.086, and 0.087 for Qwen3-8B on the nine OpinionQA attributes. To measure the contribution of individual layers, we apply subsets of a jointly trained shift at inference time. Applying all four layers (8, 16, 24, and 32) gives the lowest JSD among the tested configurations (0.091). Layers 8 and 16 form the best tested two-layer subset (0.119), and the best single layer is layer 16 (0.176).

E Additional Analyses

E.1 Human Evaluation

We compare the sentence-level annotations used to construct the profiles with human judgments of expression modes, context-factor presence, and value relations. For each task, annotators see the target text, the relevant value item or items, and the candidate labels.

Evaluation Tasks. For expression modes, annotators read the text, the value definition, and the candidate expression states, including *Not Invoked*. They select the state that best describes how the value is expressed and may select more than one when multiple labels fit. In Table 6, an item counts as a match if at least one evaluator selects the model's label. For expression modes, the one-level-tolerant score (PM1) also counts a match when the model's label is within one numerical position of a selected label in the value-specific schema.

For context factors, annotators judge whether a candidate factor is present in the text, explicitly or implicitly, for the specified value. We pair factors identified by the parser with non-activated candidates sampled from the same value's factor inventory. Annotators answer yes or no.

For value relations, annotators see the target text, its surrounding context, and the source and target values. The four choices are association or co-occurrence, structural or directional support, tension or trade-off (including resolution), and no clear relationship. As in the expression task, annotators may select more than one label.

Annotator Recruitment and Eligibility. We recruit Prolific participants for the language-specific cultural samples. Participants are fluent in both English and the target language and hold at least a bachelor's degree. The instructions are in English, and the passages are in the target language. Before annotation, we filter examples for harmful, sensitive, or sexual content. We use the platform's eligibility filters and do not request additional identifying information. Estimated Prolific pay is £7.20–£8.00 per hour.

For the in-person component, English-speaking volunteers annotate the political and religious samples through an online interface. Prolific sessions are designed for approximately 10–20 questions and volunteer sessions for 30–60, depending on the task.

Annotation Scale. The study covers approximately 1.8K samples and 280 participants, with a target of at least three independent judgments per sample. The study records contain 1,376 distinct Prolific task items from 268 participants and 486 volunteer task items from 15 participants. In these records, 1,039 Prolific items and seven volunteer items have responses from at least three evaluators.

Annotation Interface and Instructions. Figure 10 shows the instructions for each task. Figure 11 shows how the text, candidate labels, and response options appear to annotators.

E.2 Profile Analyses

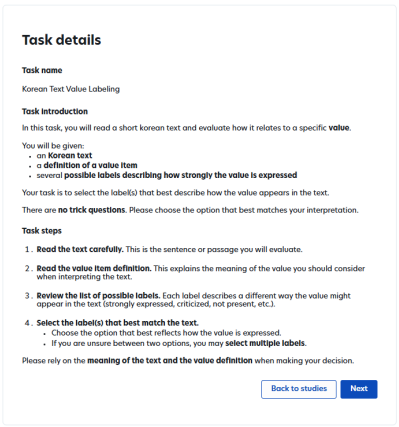
E.2.1 Profile Statistics

Table 3 summarizes the profiles by attribute category. The base-profile statistics cover seven cultural, six political, and seven religious attributes, for a total of 20. The relation summaries cover 21 attributes, including the Socialism corpus.

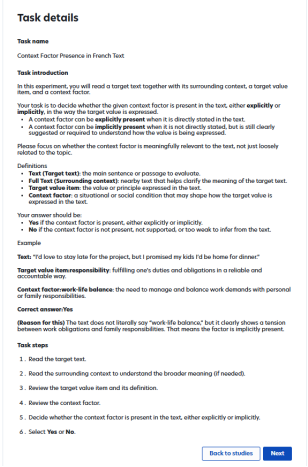
Value support and expression modes. Value support is the number of annotated sentence–value instances for a given value and attribute. *Value Supp.* in Table 3 is the mean of these counts. For an expression profile p, the dominant-mode probability is $\max_k p_k$. We compute entropy and the effective number of states from the full distribution as

$$H(p) = -\sum_{k:\, p_k>0} p_k \log_2 p_k, \qquad N_{\text{eff}}(p) = 2^{H(p)}.$$

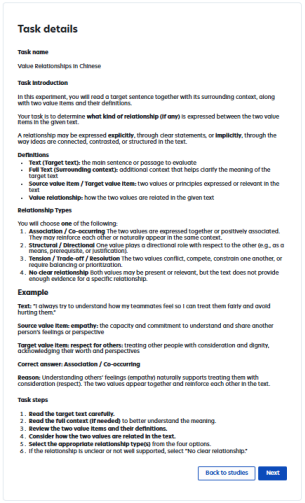
Entropy is measured in bits. The effective count is the number of equally probable modes that would give the same entropy. A profile concentrated in one mode has an effective count of one.



(a) Expression modes



(b) Context-factor presence



(c) Value relations

Figure 10: Prolific instructions for the three human-evaluation tasks.

Intensity distributions. We summarize the shape of the sentence-level intensity scores for each attribute–value pair. *Skew.* is the mean skewness across intensity distributions. The Kolmogorov–Smirnov (KS) distance is the largest absolute difference between the empirical cumulative distribution and that of the fitted Gaussian. *Gauss.* is the percentage of profiles with at least 20 scores, absolute skewness at most 0.5, absolute excess kurtosis at most 1.0, and KS distance at most 0.12.

Value relations. In the undirected co-occurrence graph, each node is a value and each edge connects a retained pair of distinct values observed together. Mean degree is the average number of neighbors per node. To measure how evenly a value's co-occurrences are spread across partners, we compute its effective partner count as $2^{H(p(v'|v,a))}$. We average this count over anchor values within each attribute, then give attributes equal weight within a category.

Contextual shifts. Discrete OT measures redistribution across the four invoked expression modes, using $|i - j|$ as the cost between modes in positions i and j. Continuous OT measures the one-dimensional Wasserstein distance between the empirical intensity distributions. Joint OT measures changes in partner values and relation labels using the greedy approximation in Section B.3.3.

The mean OT columns in Table 3 use all available distances for each component. The *Sig.* columns use comparisons with at least 30 observations per context state. They report the percentage for which the 95% interval for the signed change in mean expression-mode index or mean intensity excludes zero. We use a posterior interval for expression and a bootstrap interval for intensity.

Context factor groups. Table 16 compares shifts across the 13 factor groups defined in Table 13. Resource constraints have the largest mean intensity shift. Time pressure and psychological health share the next largest value at the reported precision. For expression modes, the largest mean shift occurs under role and identity.

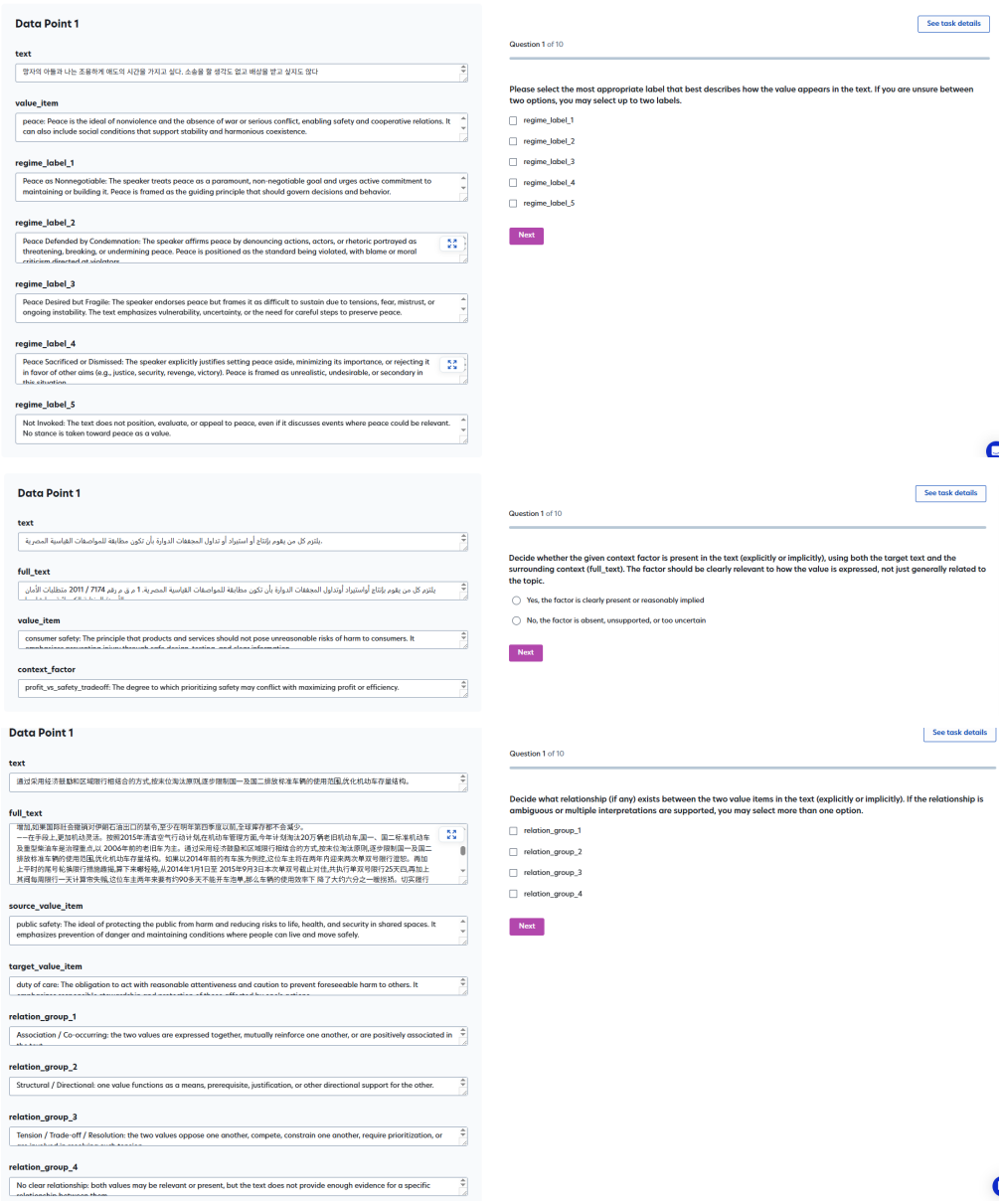


Figure 11: Representative Prolific task interfaces. From top to bottom, annotators label expression modes, judge context-factor presence, and label value relations.

Table 16: Mean contextual OT by factor group, ordered by mean intensity OT. The summary uses a minimum support of five observations per context state. Each component mean uses its available comparisons.

Factor group	Expression OT	Intensity OT
Resource constraints	0.306	0.131
Time pressure	0.333	0.125
Psychological health	0.315	0.125
Information environment	0.301	0.123
Institutional constraints	0.287	0.122
Social context	0.297	0.118
Cultural norms & sensitivity	0.311	0.114
Risk & uncertainty	0.302	0.112
Conflict/tradeoff structure	0.310	0.110
Moral salience & harm	0.305	0.103
Stakes	0.283	0.101
Role & identity	0.348	0.101
Temporal scope	0.264	0.100

E.2.2 Base Profile

Cross-attribute Divergence. We rank value items by the mean pairwise Jensen–Shannon divergence (JSD) of their base expression profiles. For each value, we average over attribute pairs for which the value is present and the profiles have the same number of bins. Figure 12 shows the 25 values with the largest means. *Corporate responsibility* and *leadership responsibility* rank first and second.

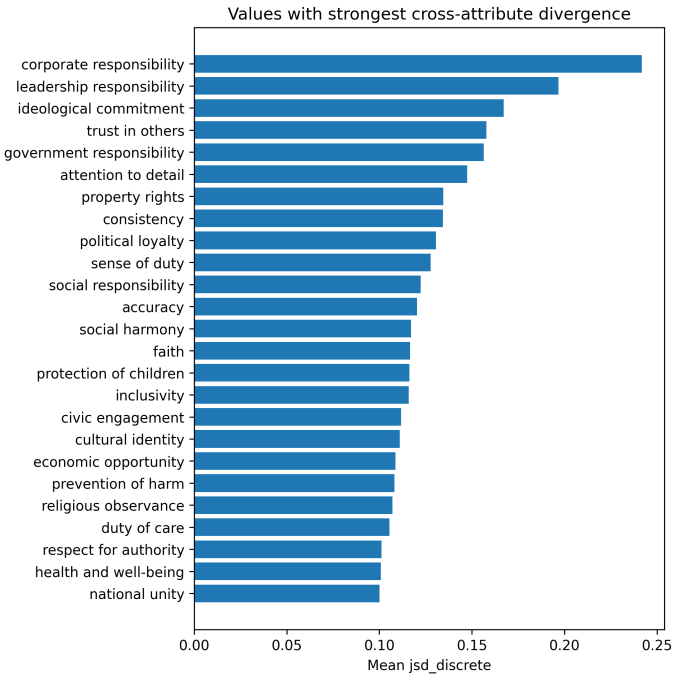


Figure 12: The 25 values with the highest mean pairwise JSD between base expression profiles. The available attribute pairs can differ by value.

Value Co-occurrence and Tension. Tables 17 and 18 rank value pairs by the number of annotated relation occurrences, counting both orientations of a pair together. The first table pools all relation labels, and the second retains only CONFLICT, TRADEOFF, and CONSTRAINT. *Honesty* and *truthfulness* form the most frequent pair overall. *Freedom of expression* and *respect for others* form the most frequent tension pair.

Examples. Figure 13 shows two values per attribute category. For *patience*, for example, the Arabic-language corpus places more weight on *Patience as Principle*, while the Japanese-language corpus places more weight on *Trying to Stay Patient*.

E.2.3 Contextual Shift

High-Shift Value Items. For *trust in others* in the Korean-language CulturaX corpus, the expression-mode OT distance between high and low *societal norms of trustworthiness* is 1.655. The corresponding distance between high and low *ethical consistency of others* is 1.625. Table 16 reports mean distances across comparisons for each factor group.

Relational Shifts under Context. For *corporate responsibility* in the Korean-language CulturaX corpus, the joint OT distance between minimal and severe stakeholder harm is 0.577, of which 0.571 is due to changes in partner frequencies. For *freedom of expression* in the Russian-language corpus, the distance between an unassigned retaliation-risk state (`NONE`) and a high-risk state is 0.608, with 0.595 due to changes in partner frequencies. In both comparisons, most of the distance comes from which values are linked to the anchor.

Table 17: The 15 value pairs with the most annotated relation occurrences, pooled across attributes and both directions.

Rank	Value Pair	Count
1	honesty ↔ truthfulness	15,039
2	integrity ↔ truthfulness	9,683
3	critical thinking ↔ intellectual honesty	8,526
4	honesty ↔ integrity	8,355
5	accountability ↔ truthfulness	7,375
6	critical thinking ↔ truthfulness	7,295
7	critical thinking ↔ honesty	6,528
8	intellectual honesty ↔ truthfulness	6,162
9	accountability ↔ honesty	5,545
10	honesty ↔ intellectual honesty	4,491
11	accountability ↔ transparency	4,439
12	accountability ↔ fairness	4,221
13	accountability ↔ integrity	4,188
14	accountability ↔ moral responsibility	3,789
15	critical thinking ↔ pursuit of knowledge	3,482

Table 18: The ten value pairs most frequently linked by conflict, trade-off, or constraint relations, pooled across attributes and both directions.

Rank	Value Pair	Tension Count
1	freedom of expression ↔ respect for others	909
2	fairness ↔ freedom of expression	862
3	accountability ↔ fairness	627
4	capitalism ↔ economic justice	456
5	freedom of expression ↔ religious tolerance	430
6	fairness ↔ truthfulness	416
7	freedom of expression ↔ religious freedom	344
8	fairness ↔ honesty	343
9	fairness ↔ justice	339
10	critical thinking ↔ faith	334

Examples. Figures 14, 15, and 16 show how the same context comparison appears in the intensity and expression profiles. For *democratic participation* in the Democrats corpus, difficult voting access is associated with a larger share of *Affirmed via Suppression Critique* than easy access, while the intensity curves overlap. For *quality of life* in the Japanese-language corpus, the most frequent mode changes from *Quality of Life Prioritized* with no long-term health issues to *Quality of Life Strained* with major issues.

Expression Mode and Intensity. Figure 7(b) shows mean intensity for affirmed, contested, and rejected expression, labeled *Support*, *Qualified*, and *Rejected*, respectively. Error bars show ± 1.96 standard errors of the mean. To compare the direction of contextual shifts, we use the same pair of context states for both components and require at least 30 observations per state in each component. We reverse the sign of the change in mean expression-mode index so that a positive change indicates stronger endorsement on both measures. Of the 74,325 comparisons with nonzero changes in both means, the two measures agree in direction in 55,894 (75.2%).

Sensitivity to expression-state costs and grouping. We recompute expression-mode OT using a common cost for any two different modes and, separately, costs based on the state descriptions. Compared with the original ordinal distances, the rank correlations are $\rho = 0.899$ for the categorical cost over 243,635 comparisons and $\rho = 0.891$ for the semantic cost on a 4,000-comparison subset.

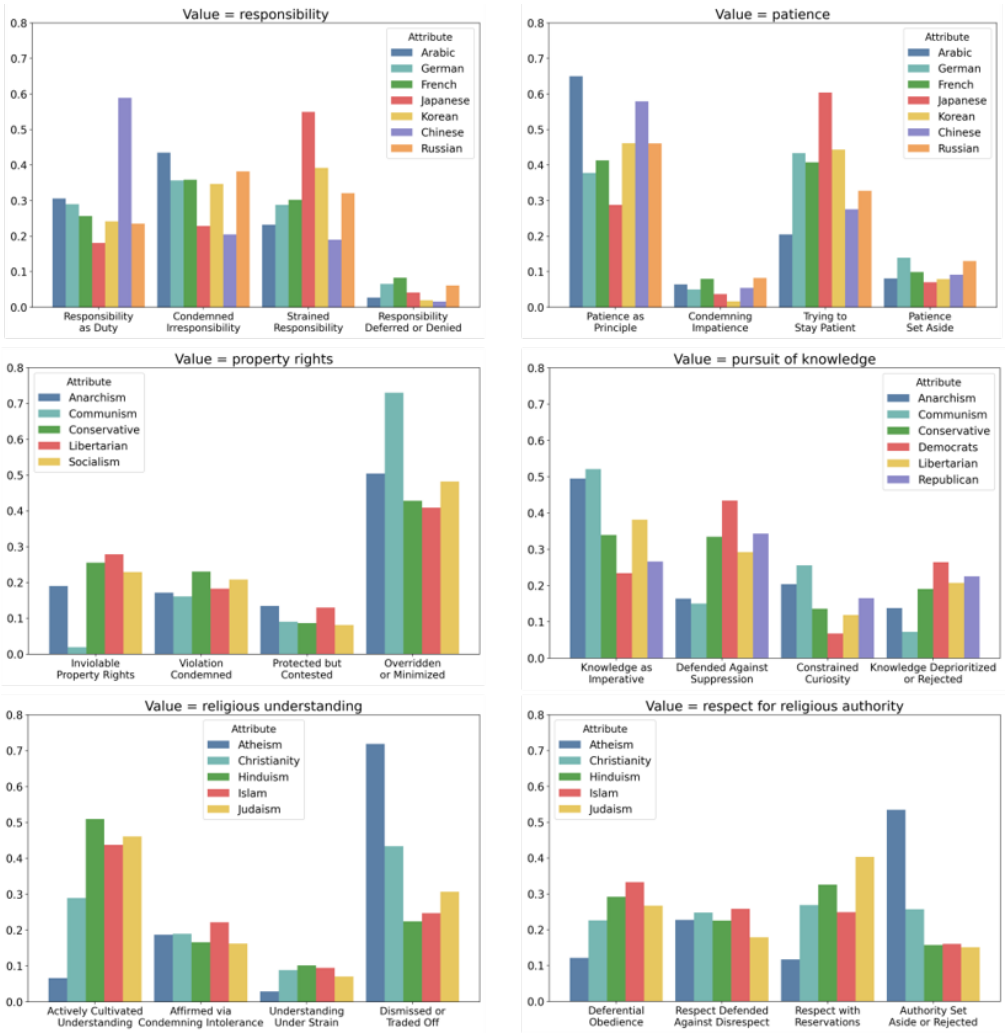


Figure 13: Base expression-mode distributions for two example values in each attribute category. Within a panel, each color denotes one corpus.

Individual rankings vary more at the upper end. In a separate comparison using empirical mode frequencies, the ordinal and categorical rankings share seven of their top 50 comparisons.

On a separate 23-attribute panel, we test how the rankings change when we merge the first two invoked modes or the middle two modes, leaving the others distinct. The rankings of contextual total variation distance correlate at $\rho = 0.532$ across the two groupings on 21,006 comparisons. When we hold the grouping fixed and split the data by parent document, correlations across halves are $\rho = 0.747$ and 0.797, respectively, on 20,999 comparisons.

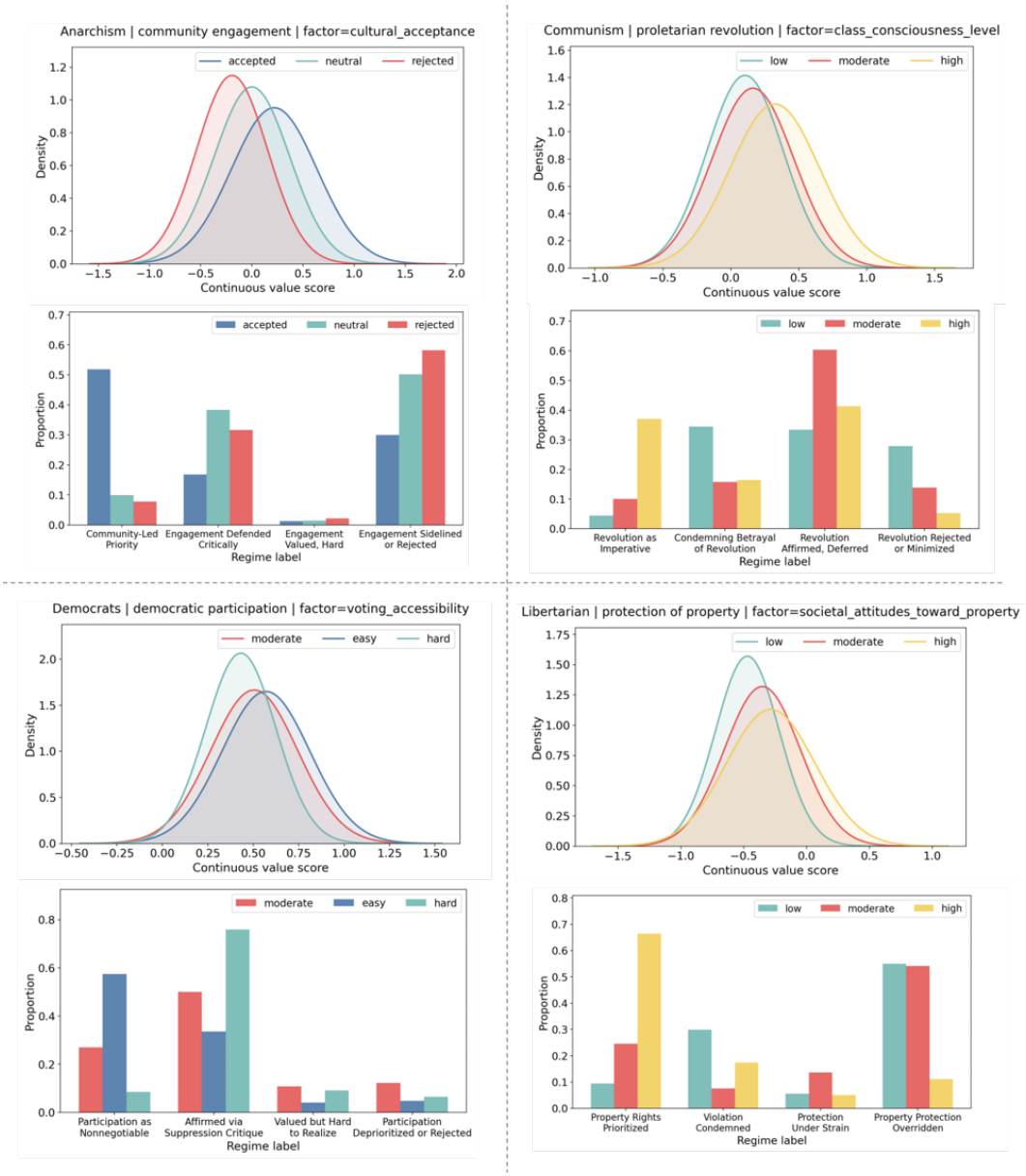


Figure 14: Contextual profiles in the Anarchism, Communism, Democrats, and Libertarianism corpora. Each example shows fitted intensity densities above and expression-mode probabilities below.

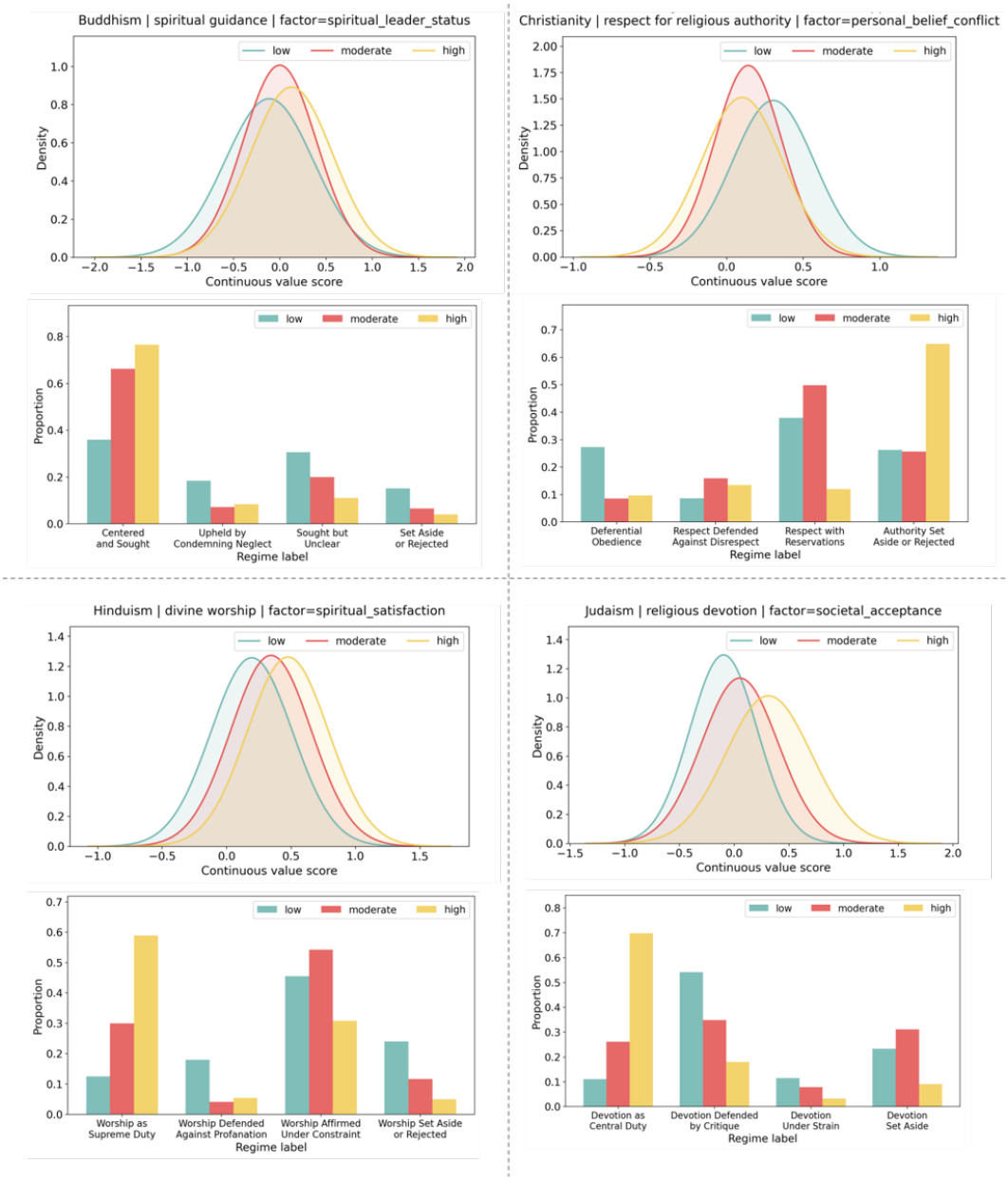


Figure 15: Contextual profiles in the Buddhism, Christianity, Hinduism, and Judaism corpora. Each example shows fitted intensity densities above and expression-mode probabilities below.

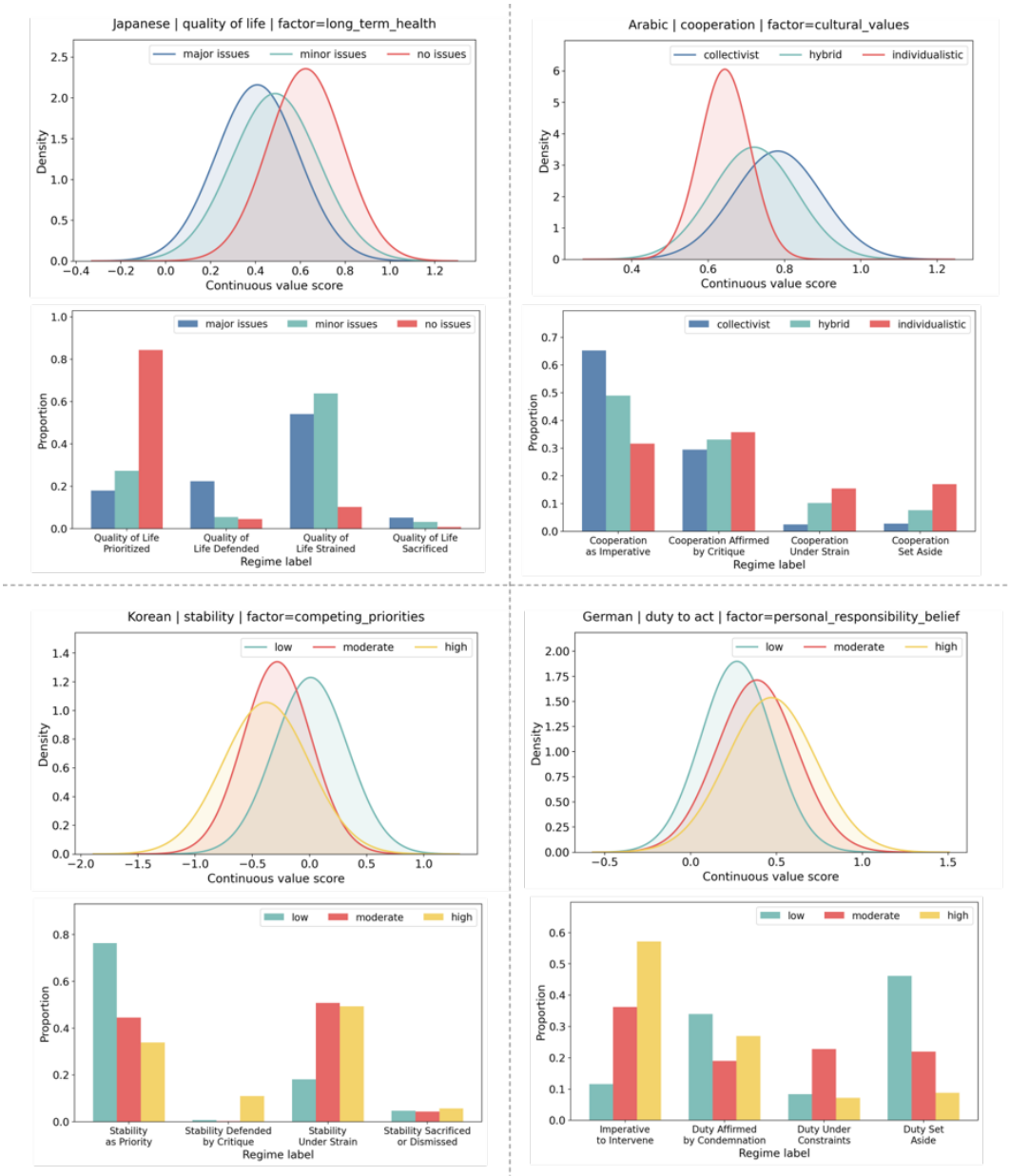


Figure 16: Contextual profiles in the Japanese-, Arabic-, Korean-, and German-language corpora. Each example shows fitted intensity densities above and expression-mode probabilities below.

E.3 Soft-Profile Sensitivity

Our primary profiles count one expression label per sentence-value pair. Here we examine how the profiles change when each label contributes the distribution of responses observed among human annotators. We use 300 multiply annotated items, 50 from each of the Arabic, Chinese, French, German, Japanese, and Korean corpora. These items received 910 responses from 87 annotators.

For this comparison, we map the five fine labels, in schema order, to *Support*, *Qualified*, *Qualified*, *Rejected*, and *Not Invoked*. For each model-assigned state, we estimate how often annotators selected each set of coarse labels, pooling across languages and values and giving each item equal weight. We pool model states 2 and 3 and add a half-count to each possible nonempty response set. If a set contains multiple categories, we divide its probability equally among them. To construct a soft profile, we average the resulting probabilities over all occurrences of a value in a language. The matched hard profile uses the same occurrences with their original labels mapped to the coarse categories. For both profiles, we exclude non-invocation and normalize over the three invoked categories.

We compare 50 values with at least 100 invoked occurrences from at least 10 parent texts in every language, with no single text contributing more than 25% of the occurrences. This gives 300 language–value profiles. One hard profile has a tie for the dominant mode. Among the remaining 299, 169 change from *Qualified* to *Support*, which is dominant in all 300 soft profiles. For each value, we also compare all 15 language pairs using total variation (TV) distance. Across the 750 comparisons, the median ratio of soft to hard TV is 0.264 (interquartile range 0.192–0.328). The shared response mapping thus reduces most cross-language distances, although 24 of the 750 distances increase.

E.4 Source Dependence

Group differences within Reddit. To examine group differences on a common platform, we compare six national Reddit communities associated with China, France, Germany, Japan, Korea, and Russia with six political and seven religious communities. For each pair of communities, we average the Jensen–Shannon divergence (JSD) between their expression profiles over shared values. The mean JSD is 0.017 higher for pairs from different categories than for pairs from the same category. Because communities from different categories tend to share fewer values, we regress pairwise JSD on the number of shared values and repeat the comparison on the residuals. The adjusted gap is 0.008 ($p = 0.010$, permuting category labels across the 19 communities).

The same values across sources. For the six countries, we compare expression profiles over the four invoked states in CulturaX and Reddit. The matched data cover 263 country–value pairs and 59 distinct values. We average same-value TV distances over the values available for each country pair, then give equal weight to the 15 country pairs. Mean TV is 0.269 for the same country across sources, 0.271 for different countries across sources, and 0.095 for different countries within a source (Table 7). Restricting the comparison to the 18 values observed for every country in both sources gives corresponding distances of 0.274, 0.273, and 0.098.

Source adjustment and country matching. On a separate panel of 38 values, we use 1,000 embedding-matched pairs of English Reddit records and local-language CulturaX records per country. We compare the original annotation procedure with an alternative annotator using either a regenerated expression schema or the original, fixed schema. For each pair of countries, we use the other four countries to estimate a pooled prior and source-specific adjustments to the expression probabilities. We compare uniform smoothing, smoothing with the pooled prior, and pooling with source adjustment, holding the prior mass fixed.

For each country pair, we ask whether a profile is closer in TV to the same country or the other country in the second source. We average over values observed at all four endpoints, use both countries as queries in both source directions, and give ties half credit. Table 19 shows that pooling reduces the same-country distance for all three annotation procedures. Adding the source adjustment leaves country matching unchanged for the fixed schema and slightly lowers it for the other two procedures.

E.5 Sensitivity to Profile Construction

Value extraction and expression labeling. We first test how much of the value vocabulary a different extractor recovers. Mistral-24B extracts value names from 540 sentences across three

Table 19: Source adjustment on the 38-value panel. TV is the mean same-country distance across sources. Match is the percentage of two-choice comparisons favoring the same country, with half credit for ties.

	Original		Regenerated schema		Fixed schema	
Method	TV	Match (%)	TV	Match (%)	TV	Match (%)
Uniform smoothing	0.167	55.0	0.150	48.3	0.196	45.0
Donor pooling	0.139	58.3	0.103	53.3	0.118	50.0
Pooling with adjustment	0.136	55.0	0.103	51.7	0.116	50.0

corpora. We embed the names with Qwen3-Embedding-0.6B. Of the 200 most frequent entries in the 1,515-item vocabulary used for this comparison, 94% have a match among the names extracted by Mistral at cosine similarity ≥ 0.80. We then compare expression labels assigned to the same sentence–value candidates by a direct Mistral classifier and a two-stage ValueLlama classifier. On 38,004 candidates from 16 groups, the classifiers agree in 40.3% of cases across the four labels *Support*, *Qualified*, *Rejected*, and *Not Invoked*. Mistral marks the value as invoked in 30.9% of cases and ValueLlama in 64.6%.

Confidence filtering and document aggregation. On a separate panel of 24 groups, we vary the expression-label confidence threshold on up to 40,000 sampled records per group. Within each group, we rank values by their mean index over the four invoked expression states and compare the ranking with the one obtained without filtering. At threshold 80, the median Spearman correlation across groups is 1.000. At threshold 90, it is 0.947, with 91.9% of labels retained on average.

We also compare sentence- and document-level aggregation using 5,001 parent texts from each of nine groups. Among document–value pairs with at least two sentences invoking the value, an average of 44% across groups contain more than one expression mode. Assigning each pair its most frequent invoked mode removes this within-document variation, although the pooled four-state distributions for each group change little (mean JSD approximately 10^{-4}).

Comparison with Moral Foundations. We map value items to five Moral Foundations (*Care*, *Fairness*, *Loyalty*, *Authority*, and *Sanctity*). We compare the resulting group rankings with rankings based on model judgments of stance, topic frequency, and scores from two dictionaries, the extended Moral Foundations Dictionary (eMFD) and the Moral Foundations Dictionary 2.0. Without seeing group identifiers, Mistral-Small-24B judges whether a passage affirms, is neutral toward, or rejects a foundation. Its 3,780 judgments cover 63 of the 65 group–foundation cells across 13 political and religious groups. Topic frequency and the two dictionaries cover all 65 cells.

For each foundation, we rank groups by each measure and compute its Spearman correlation with the profile ranking. We also compare the scores after subtracting foundation means and then group means from each matrix. Model judgments of stance give the highest mean within-foundation rank correlation. After centering, eMFD and topic frequency have the highest correlations with the profile scores (Table 20).

Table 20: Agreement between the profiles and alternative Moral Foundations measures. Spearman ρ is averaged across the five foundations. Pearson r is computed across matched cells after centering by foundation and group.

Measure	Mean Spearman ρ	Centered Pearson r
Model-based stance	0.491	0.007
Topic frequency	0.332	0.450
Extended Moral Foundations Dictionary	-0.256	0.498
Moral Foundations Dictionary 2.0	0.047	0.179

F Limitation and Impact Statement

F.1 Limitation

Reliance on LLMs. We use LLMs to identify value-relevant sentences, assign value items, and annotate expression modes, context states, and relations. Differences in how models interpret phrasing or rhetorical cues can affect the resulting profiles. Even on the same sentence–value pairs, different annotation procedures can assign different expression labels (Appendix E.5). Section 6.2 compares the annotations with human judgments on a subset of the data and examines how variation in those judgments affects the expression profiles.

Discourse-Level Representation Rather Than Belief Inference. The profiles describe how values are expressed in sampled, group-associated discourse. Cultural profiles draw on CulturaX web text, while political and religious profiles draw on Reddit discussions. These sources differ in audience and register, so differences across group categories can also reflect where the text was collected. In the six-country comparison, profiles for the same country differed more across sources, on average, than profiles for different countries within a source (Section 6.3). Comparisons across additional sources would help identify which patterns recur across settings.

Distributional Evaluation. GlobalOpinionQA and OpinionQA measure agreement with survey-response distributions over predefined options. In the free-text evaluation, the model answers without seeing the options, and a separate model maps each response back to an option (Appendix D.4). Answers assigned to the same option can differ in their wording or reasoning, which this evaluation does not capture.

Data Scale and Computational Requirements. After filtering, the corpora contain 141,353 to 1,249,808 sentences per attribute (Table 8). Only a subset of sentences invoke a given value, and dividing these observations by context further reduces the support for each comparison. This limits coverage of infrequent values and contexts in smaller corpora. Sentence filtering, value extraction, and annotation also require repeated model inference.

F.2 Broader Impacts

Distributional profiles preserve less frequent ways of expressing a value alongside the dominant mode. This can help alignment systems represent variation within a group. However, a system that applies a group profile as a default for every member can reinforce stereotypes and override individual preferences. Downstream applications should allow users to override group-level defaults.

The same profiles could be used to tailor persuasive messages or disinformation to a community, and steering could make generated text resemble that community's discourse. These risks extend to profiles that accurately describe the sampled text, since a description of how values are expressed can also be used to influence that expression.

F.3 Ethical Concerns & Safeguards

We annotate individual sentences from publicly available CulturaX documents and Reddit posts and aggregate the annotations into corpus-level profiles. We do not infer individual beliefs. Because each corpus covers a particular source and period, a profile should not be treated as a fixed characterization of a cultural, political, or religious group.

For the human evaluation, we recruited approximately 280 participants through Prolific and in-person annotation. We screened passages for harmful or sensitive content before annotation. Estimated Prolific pay was £7.20–£8.00 per hour. We used the platform's eligibility filters and requested no additional identifying information. Participants judged how values were expressed in the text without having to disclose their own views.

We plan to release the profiles and code with documentation describing the sources, the temporal scope of the corpora, and the intended use of the profiles as descriptions of group-associated discourse.

G License

We plan to release the constructed distributional value profiles, trained models, and code under the Apache 2.0 license. The CulturaX dataset [33] is distributed under the terms of the mC4 and OSCAR licenses, and the Reddit data [4] is subject to Reddit's Developer Terms (`https://redditinc.com/policies/developer-terms`). Since neither data source permits redistribution of raw text, we do not release the raw samples or their sentence-level annotations. For the profiling model, Reddit's policy restricts public release of models trained directly on raw post content; we therefore plan to release a version of the Stage 1 model trained without Reddit data, while the later stages that rely only on aggregated distributional targets are unaffected by this restriction.

H LLM Usage

Large language models were used in two capacities in this work. For writing, we used LLMs to polish prose and check code snippets. For methodology, LLMs serve as core components of the pipeline, including data filtering, value item extraction, expression state annotation, context factor parsing, relationship extraction, and multi-model auditing of the ontology, as described in the corresponding sections.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's scope: a distributional view of group-level value expression, profile construction across cultural, political, and religious groups, context-shift modeling, and downstream use for prediction and steering. The claims are matched to the reported empirical analyses and are framed with appropriate scope, including that text-derived profiles reflect attribute-associated discourse rather than direct measurements of individual beliefs.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are provided in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not make theoretical claims and formal proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed descriptions of the data collection, filtering, structured annotation pipeline, profile aggregation procedure, profiling model, steering method, baselines, and evaluation metrics. These details, together with the appendix descriptions of annotation, context modeling, and evaluation setup, provide sufficient information to reproduce or verify the main experimental results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We plan to release the code, data, and license information upon paper release, together with instructions for reproducing the main results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Details are provided in Appendix B, C, D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports the detailed statistics of the constructed profiles in Section 4.4, Appendix E.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the computation resources utilized in the paper in Appendix B and the data scale in Appendix B.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We certify that the paper follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader impacts are discussed in Appendix F.2 and Appendix F.3

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Ethical concerns and the safeguards are discussed in Appendix F.3

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Licenses are mentioned at Appendix G.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Details are mentioned at Appendix G.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We include the full details for human evaluation (instructions, compensation details) in Appendix E.1.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: While the paper includes evaluations conducted by human annotators, the annotation process does not collect any personal information and is used solely to assess text quality. The study does not involve human subjects as the object of analysis.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer:[Yes]

    Justification: The paper include the details of the LLM usage in Appendix H.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.