# Labels have Human Values: Value Calibration of Subjective Tasks

Anonymous ACL submission


## Abstract

Although pluralistic societies exhibit diverse human values that lead to legitimate disagreements in subjective tasks (*e.g.*, safety and preference judgments), NLP models trained on such subjective labels often ignore latent value structures, resulting in miscalibrated predictions over relevant value classes. We propose MultiCalibrated Subjective Task Learning (MC-STL), a framework that identifies latent value groups from annotations (via label rationale similarity, expert value taxonomies, or annotator sociocultural descriptors) and enforces value-conditional calibration through value group-specific representations. MC-STL applies to binary, ordinal, and preference learning settings, and is evaluated on multiple datasets covering toxic chatbot conversations, value reasoning, T2I-safety and preference alignment. The results demonstrate that MC-STL consistently outperform existing baselines, achieving multicalibration across relevant value groups, while delivering gains in probabilistic prediction performance.


## 1 Introduction

As NLP systems are increasingly deployed in socially sensitive settings, ensuring reliable behavior across the diversity of human values has become an important challenge (Pehlivan et al., 2024; United Nations, 2023; Sorensen et al., 2024b). However, conventional machine learning workflows often suppress value diversity, either explicitly through majority aggregation (Hovy et al., 2013) or implicitly when learning methods average over latent contexts in training data that are skewed towards certain perspectives unknown to the model but over-represented in the data (Obi et al., 2024; Siththaranjan et al., 2024); as illustrated in Figure 1.

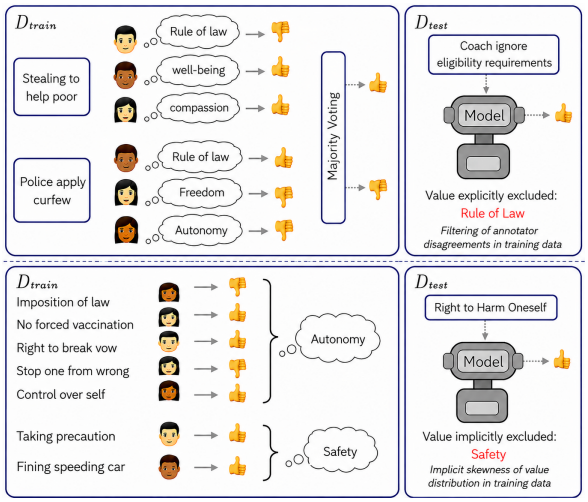



Figure 1: Example of how subjective label predictions are affected by hidden human values (denoted as thought clouds). (Top) when annotator disagreements are filtered out by majority voting or (Bottom) when minority values are underrepresented and likely unaccounted by the model in crowd-sourced annotation settings.

Current NLP alignment approaches focus primarily on post-training alignment of generated outputs through activation tuning, reward modeling, or value-space mapping (Jin et al., 2025; Christiano et al., 2017; Yao et al., 2024). However, these methods primarily address the alignment of content generated by LLMs and are usually not extendable to categorical decision tasks. Ignoring the structure of latent (unobserved) values in subjective annotations can lead to systematic misalignment in such high-stakes NLP systems trained on them (Christian et al., 2026; Elle, 2025). Kirk et al. (2024) has therefore called for consideration of the value perspectives underlying the annotations themselves.

Recent work on subjective NLP has increasingly treated annotator disagreement as an informative signal rather than annotation noise (Prabhakaran et al., 2021). Although early work focused on annotator-centric modeling (Davani et al., 2022) or learning disagreements directly (Romberg, 2022), recent work has explored factors that influence dis-

agreements, *e.g.*, sociodemographic factors (Rastogi et al., 2026), attitudes (Jiang et al., 2024), *etc.*

Following Schwartz (2012), subjective disagreements often arise from differing underlying *human values*. Here, we build on the intersection of value alignment and subjective NLP by identifying distinctive *human value groups* from annotations and learning reliable predictions across these groups. The framework, hence, could also disambiguate annotation disagreements arising from value conflicts. For completeness, we provide a detailed description of the Related Work in Appendix A.1.

In terms of *value alignment*, we specifically focus on multicalibration (Hébert-Johnson et al., 2018; Corvelo Benz and Rodriguez, 2023) of a subjective NLP system to relevant human value groups. Value group multicalibration requires prediction probabilities to remain calibrated across relevant value group perspectives rather than only globally calibrated, thereby ensuring reliable probabilistic predictions for diverse human viewpoints. To the best of our knowledge, the proposed MultiCalibrated Subjective Task Learning (MC-STL) is the first framework to explicitly model latent human value group perspectives underlying subjective annotations while enforcing calibration over all value groups. Our contributions are as follows.

- The MC-STL framework identifies latent value groups among subjective annotations by label rationale clustering, expert-defined value taxonomy mapping, or sociocultural descriptors, and learns calibrated predictions over these value groups.
- The MC-STL framework supports various task scenarios, such as binary classification, ordinal classification, and human preference learning. Moreover, it is scalable to diverse, multimodal subjective applications such as toxicity detection, value reasoning, T2I (Text-to-Image) Safety, *etc.*
- Across multiple subjective NLP benchmarks, MC-STL improves predictive performance, achieves more reliable multi-calibration over value groups, and better preserves minority viewpoints compared to existing subjective modeling baselines and value agnostic methods.

# 2 Methodology

## 2.1 Problem Definition

Let $\mathcal{D} = (\mathbf{X}, \mathbf{Y})$ denote an annotated dataset for a subjective natural language processing (NLP) task. The dataset consists of a collection of $N$ text items $\mathbf{X} = \{x_i\}_{i=1}^{N}$, where each item $x_i$ is associated with a set of annotations $\mathbf{y}_i = \{y_{ij}\}_{j=1}^{m_i}$, with $m_i$ denoting the number of annotations collected for the item $x_i$. The collection of annotation sets $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{N}$ constitutes the set of label annotations for the complete dataset. The annotations $y_{ij}$ can take values from binary or ordinal label spaces. The task is twofold: first, to map each annotation $y_{ij}$ to one or more of its relevant *human value groups* (by clustering with commonly available meta-data or expert analysis). Let $\mathcal{C} = \{c_k\}_{k=1}^{K}$ denote the resulting set of $K$ value groups obtained by clustering annotations, where each value group $c_k$ is a subset of annotations sharing a common value-driven perspective. Value groups may follow a mixed membership model, where an annotation $y_{ij} \in \mathbf{Y}$ can belong to more than one value group in $\mathcal{C}$. Let $S_{ij}$ be the set of value groups to which $y_{ij}$ is assigned, then $S_{ij} = \{c_k \in C \mid y_{ij} \text{ is assigned to } c_k\}$. The second task is to learn a model whose predictions $\hat{P}(y_{ij} \mid x_i,\ S_{ij})$ are multi-calibrated over value groups, $c_k \in C$.

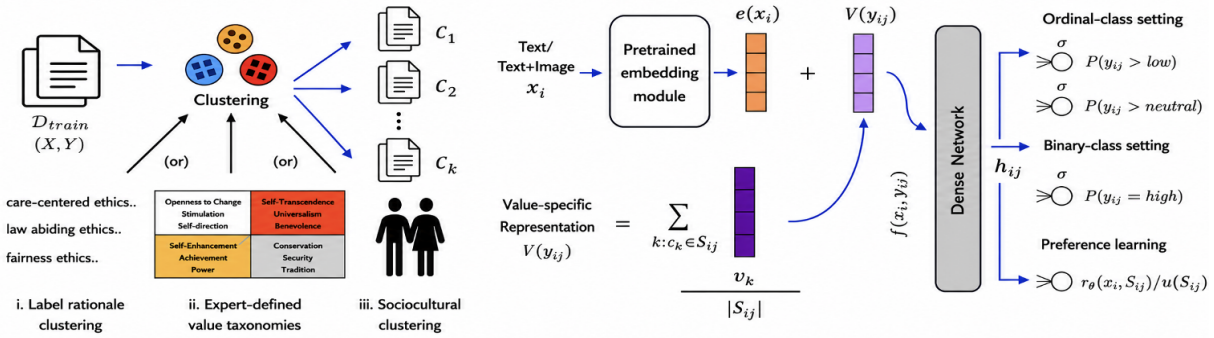
## 2.2 MC-STL Framework

### 2.2.1 Value Group Construction

We consider three approaches for the construction of value groups in $C$ from subjective annotations, capturing distinct notions of value group structures, as described in the following.

**Label rationale clustering** Given free-text rationales or value statements associated with each annotation $y_{ij}$, either collected during annotation or retrospectively inferred using LLMs (Sorensen et al., 2025; Hayati et al., 2024), we cluster the annotations in the train split of $D$ using $K$-means over sentence embeddings of their rationales. The resulting clusters group annotations that share semantically similar value themes. Each value group (cluster) may then be interpreted by summarizing its associated rationales using either an LLMs or domain experts. The number of value groups $K$ can be selected through interpretability analysis by inspecting value-group summaries or by validation performance. At test time, the embedding of the rater rationale for an instance is mapped to its closest cluster centroid, and the instance is evaluated based on the corresponding value group.

**Expert-defined value taxonomies** We also consider predefined human value taxonomies that directly map each annotation $y_{ij}$ to established value classes, forming value groups among their anno-




(a) (b)

Figure 2: Overview of the proposed MC-STL framework. (a) Identification of value groups. (b) MC-STL model architecture.

tations. For example, Schwartz (2012) defines ten universal human value classes. Similar taxonomies are widely used in RLHF settings (Bai et al., 2022), where preference annotations reflect high-level value dimensions such as *Helpfulness* and *Harmlessness*. In this setting, $K$ corresponds to the number of predefined value classes.

**Sociocultural Clustering** In some subjective tasks, annotation differences are influenced by sociocultural factors (Aroyo et al., 2023; Dev et al., 2026; Davani et al., 2024) and ensuring fairness over all sociocultural groups is a critical concern. Here, annotations are grouped according to annotator attributes such as age, ethnicity, or geographic location. These attributes act as proxies for latent cultural value perspectives underlying the annotator's judgements. In this setting, $K$ equals the number of sociocultural groups considered.

The three approaches above differ from each other in terms of the value groups formed. Rationale clustering infers latent value groups directly from label rationale semantics and is particularly useful when task-relevant value distinctions are unknown *a priori*. In contrast, expert-defined taxonomies provide task-independent value dimensions aligned with established human value theories, while sociocultural clustering captures demographic and cultural variation underlying subjective judgments. Choosing a method for the construction of value groups for a task therefore depends on the task objectives, fairness considerations, and data availability.

### 2.2.2 MC-STL Model Architecture

The proposed MC-STL model adopts a value-aware architecture to model human subjectivity using the constructed value groups, as illustrated in Figure 2. Given a text item $x_i$, we first encode it using a pretrained transformer-based encoder into a $d$-dimensional embedding $e(x_i) \in \mathbb{R}^d$.

For each value group $c_k \in \mathcal{C}$, we initialize a unique learnable embedding $v_k \in \mathbb{R}^d$. These embeddings are trained jointly with the rest of the model parameters and are intended to capture value-specific inductive biases. For a given annotation $y_{ij}$, the corresponding value-specific representation is obtained by averaging the embeddings of all value groups associated with it as follows.

$$V(y_{ij}) = \frac{\sum_{k\,:\,c_k \in S_{ij}} v_k}{|S_{ij}|}. \tag{1}$$

For each training instance $(x_i, y_{ij})$, we construct a value-aware representation by combining the text embedding with the value-specific embedding as $f(x_i, y_{ij}) = e(x_i) + V(y_{ij})$. This formulation naturally supports value group mixed membership (such as in Schwartz's value theory) while preserving the dimensionality of $f(x_i, y_{ij})$ through (1).

The value-aware representation $f(x_i, y_{ij})$ is passed to a feed-forward prediction network. Let $h_{ij}$ denote the scalar output of the dense transformation excluding final-layer bias parameters. We employ task-specific output heads depending on the prediction setting: $i$) ordinal label setting to accommodate subjective annotations with ordered response scales (*e.g.*, Likert-style labels) (Sorensen et al., 2024a; Aroyo et al., 2023), and $ii$) a binary label setting for explicit two-class scenarios.

**Ordinal classification setting** For ordinal prediction tasks, *i.e.*, with $r$ ordered categories, we adopt the CORAL (Consistent Rank Logits) framework (Cao et al., 2020), which decomposes ordinal prediction into $r-1$ binary thresholding sub-tasks.

Let $y_{ij} \in \{l_1, l_2, \ldots, l_r\}$ denote an ordinal label space with ordered categories $l_1 < l_2 < \cdots < l_r$.

For each threshold $t \in \{1, \ldots, r-1\}$, CORAL defines a binary target $y_{ij}^{(t)} = \mathbf{1}[y_{ij} > l_t]$. The probabilities for each target are defined as

$$\hat{p}_{ij}^{(t)} = \hat{P}(y_{ij}^{(t)} = 1) = \sigma\left(h_{ij} + b_t\right)$$

where $t \in \{1, \ldots, r-1\}$ indexes the ordinal thresholds, $b_t$ is the bias parameter associated with threshold $t$ and $\sigma(\cdot)$ is the sigmoid function. Following CORAL, this formulation enforces rank consistency across ordinal thresholds. The class probabilities are recovered from the adjacent cumulative probabilities following the standard CORAL formulation (Cao et al., 2020).

**Binary classification setting** It corresponds to the special case where $r = 2$. Given labels $y_{ij} \in \{0, 1\}$, the model reduces to a single prediction head $\hat{p}_{ij} = \hat{P}(y_{ij} = 1) = \sigma\left(h_{ij} + b\right)$.

#### 2.2.3 Loss Function

Our training objective is designed to jointly optimize: $i$) probabilistic prediction at the instance level, and $ii$) distributional calibration at the value group level. For ordinal prediction tasks with $r$ ordered categories, the model produces $r-1$ binary target predictions following the CORAL formulation described above. For each threshold $t \in \{1, \ldots, r-1\}$, we optimize a binary cross-entropy (CE) objective and a group-level distribution matching regularizer based on the Kullback-Leibler (KL) divergence. The complete training objective is

$$\begin{aligned}\mathcal{L}_{\text{ord}} = &\sum_{i,j,t} \mathcal{L}_{\text{CE}}\left(y_{ij}^{(t)}, \hat{p}_{ij}^{(t)}\right) \\ &+\lambda_1 \sum_{k,t} \alpha_k \text{KL}\left(P_k^{(t)} \| Q_k^{(t)}\right) + \lambda_2 \sum_k \|v_k\|_2^2,\end{aligned} \tag{2}$$

where $\mathcal{L}_{\text{CE}}$ denotes binary CE between the target $y_{ij}^{(t)}$ and prediction $\hat{p}_{ij}^{(t)}$, the KL divergence encourages the average predicted distribution within each value group $c_k \in \mathcal{C}$ to match the corresponding empirical label distribution from the training data, and the last term regularizes the embeddings, $\{v_k\}_k$.

Let $P_k^{(t)}$ and $Q_k^{(t)}$ denote the empirical and predicted Bernoulli distributions, respectively, for each value group $c_k$ and threshold $t$. We define

$$\begin{aligned}\text{KL}(P_k^{(t)} \| Q_k^{(t)}) = &\, n_k \bar{y}_k^{(t)} \ln\left(\frac{\bar{y}_k^{(t)}}{\hat{p}_k^{(t)\prime}}\right) \\ &+ n_k\left(1 - \bar{y}_k^{(t)}\right) \ln\left(\frac{1 - \bar{y}_k^{(t)}}{1 - \hat{p}_k^{(t)\prime}}\right).\end{aligned}$$

where $n_k$ is the number of annotations associated with value group $c_k$, $\bar{y}_k^{(t)}$ is the empirical mean of $\{y_{ij}^{(t)}\}_{ij \in c_k}$, and $\hat{p}_k^{(t)\prime}$ is the corresponding average predicted probability.

Although we have discrete realizations $y_{ij} \in \{0, 1\}$ from $P_k$ to obtain $\bar{y}_k$, we only have predicted probabilities for the realizations of $Q_k$. We soft-threshold predictions wrt the empirical mean $\bar{y}_k^{(t)}$ before aggregating them to obtain $\hat{p}_k^{(t)\prime}$. Specifically, we calculate $\hat{p}_k^{(t)\prime}$ after converting each probability into *approximately* binary labels using

$$\hat{p}_k^{(t)\prime} = \frac{1}{2} \frac{\sum_{ij \in C_k} \left(1 + \tanh\left(l \cdot (\hat{p}_{ij}^{(t)} - \bar{y}_k^{(t)})\right)\right)}{n_k}.$$

where the hyperbolic tangent (tanh) function, with a large constant $l = 10^4$ serves as a relaxation to the use of predictions with hard thresholds while allowing for smooth gradient flow during training.

In (2), $\|v_k\|$ is an ($L_2$) regularizer for the embedding associated with the value group $c_k \in \mathcal{C}$. Note that $v_k$ is high-dimensional and matches the dimensionality of $e(x_i)$, which is fixed. Moreover, $\alpha_k$ is a parameter that weights $\text{KL}(P_k^{(t)} \| Q_k^{(t)})$ according to whether their label distribution $P_k^{(t)}$ matches that of the training set, thus up-weighting value groups for which the distributions are different. Details are presented in Appendix A.2. Finally, $\lambda_1$ and $\lambda_2$ are hyperparameters that balance the contributions of KL divergences and L2 regularizer, which are determined by grid search.

**Binary setting** It is a special case of (2) with $n = 2$ requiring a single threshold, thus reducing it to

$$\begin{aligned}\mathcal{L}_{\text{bin}} = &\sum_{i,j} \mathcal{L}_{\text{CE}}(y_{ij}, \hat{p}_{ij}) \\ &+\lambda_1 \sum_k \alpha_k \text{KL}(P_k \| Q_k) + \lambda_2 \sum_k \|v_k\|_2^2.\end{aligned}$$

### 2.3 Extension to Preference Learning

Preference learning settings commonly used in RLHF-based reward modeling (Christiano et al., 2017) require pairwise preference annotations that often reflect latent human values. Preference confidence scores are frequently used to filter synthetic labels (Shen and Zhang, 2024) without explicitly modeling calibration over distinct value groups.

We extend MC-STL to preference learning as follows. Let $\mathcal{D} = \{(x_i, x_j, y_{ij})\}$ denote a dataset of pairwise preferences, where $y_{ij} \in \{0, 1\}$ indicates whether text item $x_i$ is preferred over $x_j$. Each

preference annotation is associated with a set of latent value perspectives $S_{ij} \subseteq C$. Given a text item $x_i$ and value assignment $S_{ij}$, the MC-STL backbone as in Figure 2 produces a scalar reward score $r_\theta(x_i, S_{ij}) = h_{ij}$. To model value-dependent uncertainty in preference judgments, we introduce a learnable temperature parameter $u_k$ for each value group. For mixed value group assignments, we define the effective temperature as

$$u(S_{ij}) = \frac{1}{|S_{ij}|} \sum_{c_k \in S_{ij}} u_k.$$

The probability that $x_i$ is preferred over $x_j$ is then modeled using a Bradley–Terry formulation:

$$\hat{p}(y_{ij}) = \sigma\left(\frac{r_\theta(x_i, S_{ij}) - r_\theta(x_j, S_{ij})}{u(S_{ij})}\right),$$

and the corresponding training objective is

$$\mathcal{L}_{pl} = \sum_{i,j} \mathcal{L}_{BT}(y_{ij}, \hat{p}(y_{ij})) + \lambda_2 \sum_k \|v_k\|_2^2,$$

where $\mathcal{L}_{BT}$ denotes the Bradley–Terry preference loss (Christiano et al., 2017).

## 3 Experiments

All experiments were conducted on a server equipped with a single NVIDIA RTX A6000 GPU. For label rationale clustering, we used the GPT-4o API to summarize rationales in each value group (prompt details are provided in Appendix A.4). For text-only items (excluding preference learning), text representations $e(x_i)$ were obtained using pretrained Sentence-BERT embeddings (all-MiniLM-L6-v2) Reimers and Gurevych (2019). For preference learning tasks, we used Longformer-base-4096 (Beltagy et al., 2020) to accommodate long multi-turn context input. The model was fine-tuned for 2 epochs on the training set, and the final hidden-state representation of the [CLS] token was used as the text embedding. For text-to-image (T2I) safety, joint prompt-image representations were constructed by summing the pretrained CLIP embeddings (Radford et al., 2021) of the text prompt and the generated image. The evaluation was performed on a (held-out) test split consisting of 20% of each dataset. Data splits were constructed using stratified sampling to preserve label distributions. Additional details are provided in Appendix A.3.

### 3.1 Datasets

We evaluate MC-STL over subjective NLP tasks spanning value reasoning, toxicity detection, RLHF preference modeling, and multimodal safety.

**ValuePrism** (Sorensen et al., 2024a) is a large-scale subjective reasoning dataset that contains approximately 31k text situations paired with 218k annotations. Each annotation includes an ordinal label (*Oppose*, *Either*, or *Support*) together with a free-text rationale. We evaluate MC-STL separately on two rationale categories: VP-Rights and VP-Duty, enabling experiments under distinct normative framings.

**ValuePrism + Schwartz (VP-Schwartz) Value Labels** To evaluate expert-defined value taxonomies, we extend ValuePrism with 10 Universal Schwartz value class labels (Schwartz, 2012). We map annotations to one or more Schwartz value classes using GPT-4o following the calibrated Schwartz value attributor prompt proposed by (Guo et al., 2026), and a random subset (3k samples) of each $c_k$ formed was further verified by human experts (psychologists), achieving 95% agreement with the generated mapping. See Appendix A.6 for additional details.

**Anthropic HH-RLHF** (Bai et al., 2022) is a widely used preference-learning dataset that contains approximately 169k human preference pairs. We use the ValueImprint annotations (Obi et al., 2024), which associate preference pairs with high-level human value classes.

**DICES-990** (Aroyo et al., 2023) is a dataset of 990 conversations, each annotated for toxicity by 60 raters using ordinal labels: *Unsafe*, *Unsure*, and *Safe*. Each annotation is linked to annotator demographics, *i.e.*, gender, race, age group, and locality.

**DIVE** (Rastogi et al., 2026) is a multimodal text-to-image (T2I) safety dataset containing 1000 prompt-image pairs annotated for safety by demographically diverse annotators (described by their age, gender and ethnicity) using a 5-point ordinal scale.

The value groups for each dataset and related details are given in Appendix A.7. For label rationale clustering, we evaluated multiple value group counts $K \in \{2, \ldots, 6\}$, and the final value of $K$ was selected by expert evaluation by a domain psychologist based on interpretability and uniqueness as described in Appendix A.5.

### 3.2 Evaluation Metrics

**Predictive Performance (PP)** The *overall PP* indicates the quality of all the probabilistic predictions made by the model. In binary and ordinal label setting, we use the *(micro) AUC* score. For the preference learning setting, since the ability of a model to score a preferred text item higher than a rejected item is important, we use pairwise accuracy.

For the *value group level PP*, we use the same metrics, but quantify them with respect to each value group, and summarize them as a mean with standard deviation.

**Value-Conditional Calibration (VC-CAL)** We evaluate calibration separately within each value group to measure whether predicted probabilities remain reliable over diverse value perspectives. For each value group $c_k \in C$, predicted probabilities are divided into decile bins and compared to the corresponding empirical outcome frequencies. A linear fit is applied to the resulting calibration points. Ideally, a calibrated model should achieve unit slope and zero bias. Slopes smaller than 1 indicate over-confident predictions for a value group, whereas slopes larger than 1 indicate under-confidence. Positive bias suggests systematic underestimation of the outcomes for a value group, while negative bias indicates the opposite.

For ordinal prediction tasks with $r$ ordered categories, calibration is evaluated separately for each CORAL threshold probability $\hat{P}(y_{ij}^{(t)} = 1)$, where $t \in \{1, \ldots, r-1\}$. For preference learning, calibration is computed for pairwise preference probabilities $\hat{p}_{ij} = \hat{P}(y_{ij} = 1 \mid x_i, x_j, S_{ij})$. We report the mean and standard deviation of the calibration slope and bias over all value groups. Larger deviations imply that model's confidence is not equally reliable over distinct value group perspectives.

**Minority Opinion Alignment (MO-Align)** Subjective NLP tasks often contain legitimate minority opinions that may be poorly captured by models optimized for majority agreement. To evaluate whether models preserve annotation diversity, we measure alignment of predicted label distributions with minority labels (non-majority) on each $x_i \in X$ using normalized 1D Earth Mover Distance (EMD) (Castaño et al., 2022).

EMD measures the minimum probability mass required to transform a predicted distribution into the empirical annotation distribution while respecting the ordinal structure. Errors between nearby ordinal categories are penalized less than errors between distant categories. For binary labels, EMD reduces to the probability assigned to incorrect class. For each text item $x_i$, we compute the EMD between the predicted label distribution and each non-majority annotation label associated with the item. These distances are summed and averaged across all minority annotations, and normalized by $(r)$ to maintain comparable metric ranges across datasets with different label cardinalities. We report $1 - (\text{normalized } \overline{\text{EMD}})$ so that larger values indicate better alignment with minority opinions. The full formulation is provided in Appendix A.8.

### 3.3 Baseline Methods

We compare MC-STL against both value-agnostic and value-aware subjective modeling baselines to evaluate: $i$) the utility of value groupings, and $ii$) the contribution of the MC-STL model.

**Value-agnostic**

*Majority Filtering:* A standard majority-label baseline where only the majority annotation for each training item $x_i \in X$ is retained.

*Annotator Grouping-MT:* (Davani et al., 2022) A multi-task learning baseline in which each annotator is assigned a separate prediction head while sharing a common encoder. During inference, the predictions are averaged over all annotator heads to obtain a final prediction.

**Value-aware**

*Value Grouping-MT:* A value-aware multi-task baseline in which each value group $c_k \in C$ is assigned a separate prediction head while sharing a common encoder representation. The predictions $\hat{P}(y_{ij} = 1)$ are obtained by averaging outputs over all value group heads associated with $S_{ij}$.

*Value Grouping-RPM:* Fleisig et al. (2023) A value-conditioned adaptation of the Rating Prediction Module (RPM), which conditions predictions on both the item $x_i \in X$ and value-group descriptors by concatenating them at input. While originally formulated to keep sociocultural descriptors at input, for rationale clustering and expert-taxonomy settings, value summaries or value class names are used as descriptors.

All baselines were trained using task-appropriate cross-entropy loss.

## 4 Results

**RQ1: Do latent value structures provide a useful signal for subjective prediction?** The results in Table 1 show that incorporating value-aware rep-

Table 1: Results across value grouping strategies. Overall PP and group PP measure AUCs except in preference learning where its pairwise accuracy, VC-CAL measures Calibration (Slope, Bias), and MO-Align measures minority alignment (via $1 - (\text{normalized } \overline{\text{EMD}})$). Standard deviations over value groups are shown as subscripts.

| | Dataset | Method | Overall PP | Group PP | VC-CAL (Slope, Bias) | MO-Align |
|---|---|---|---|---|---|---|
| **Rationale Clust.** | VP-Duty (Bin), $K = 3$ | Maj. Filter | 0.69 | $0.68_{0.03}$ | $0.56_{0.17}, 0.33_{0.18}$ | 0.37 |
| | | VG-RPM | 0.75 | $0.75_{0.02}$ | $0.79_{0.02}, 0.12_{0.01}$ | 0.62 |
| | | VG-MT | 0.76 | $0.75_{0.02}$ | $1.02_{0.01}, -0.02_{0.02}$ | 0.61 |
| | | **MC-STL** | **0.78** | $\mathbf{0.77}_{0.02}$ | $\mathbf{1.00}_{0.03}, \mathbf{0.00}_{0.03}$ | **0.64** |
| | VP-Rights (Ord), $K = 4$ | Maj. Filter | 0.68 | $0.67_{0.01}$ | $0.63_{0.07}, 0.13_{0.11}$ | 0.42 |
| | | VG-RPM | 0.72 | $0.71_{0.01}$ | $0.81_{0.10}, 0.06_{0.06}$ | 0.54 |
| | | VG-MT | 0.69 | $0.68_{0.02}$ | $0.74_{0.13}, 0.09_{0.06}$ | 0.54 |
| | | **MC-STL** | **0.75** | $\mathbf{0.73}_{0.03}$ | $\mathbf{0.96}_{0.08}, \mathbf{0.00}_{0.06}$ | **0.64** |
| **Expert Taxonomy** | VP Schwartz (Bin), $K = 10$ | Maj. Filter | 0.72 | $0.68_{0.05}$ | $0.42_{0.18}, 0.40_{0.23}$ | 0.62 |
| | | VG-RPM | 0.84 | $0.84_{0.01}$ | $0.93_{0.01}, 0.03_{0.02}$ | 0.66 |
| | | VG-MT | 0.87 | $0.77_{0.08}$ | $0.77_{0.08}, 0.13_{0.06}$ | 0.68 |
| | | **MC-STL** | **0.89** | $\mathbf{0.86}_{0.04}$ | $\mathbf{0.98}_{0.06}, \mathbf{0.02}_{0.06}$ | **0.71** |
| | HH-RLHF (Pref.), $K = 7$ | Maj. Filter | 0.61 | $0.61_{0.02}$ | $1.13_{0.15}, -0.07_{0.07}$ | 0.59 |
| | | VG-RPM | 0.61 | $0.62_{0.03}$ | $1.20_{0.17}, 0.08_{0.05}$ | 0.59 |
| | | VG-MT | 0.63 | $0.63_{0.03}$ | $0.96_{0.07}, 0.02_{0.03}$ | 0.60 |
| | | **MC-STL** | **0.64** | $\mathbf{0.64}_{0.03}$ | $\mathbf{1.00}_{0.06}, \mathbf{0.00}_{0.03}$ | **0.62** |
| **Sociocultural Clust.** | DICES-990 (Ord), $K = 14$ | Maj. Filter | 0.66 | $0.66_{0.02}$ | $0.74_{0.07}, 0.09_{0.09}$ | 0.58 |
| | | AG-MT | 0.70 | $0.70_{0.03}$ | $1.11_{0.11}, -0.03_{0.07}$ | 0.63 |
| | | VG-RPM | 0.73 | $0.73_{0.01}$ | $1.07_{0.05}, -0.01_{0.02}$ | 0.65 |
| | | VG-MT | 0.71 | $0.73_{0.05}$ | $0.93_{0.10}, 0.02_{0.02}$ | 0.64 |
| | | **MC-STL** | **0.76** | $\mathbf{0.76}_{0.00}$ | $\mathbf{1.02}_{0.04}, \mathbf{0.02}_{0.01}$ | **0.68** |
| | DIVE (Ord), $K = 10$ | Maj. Filter | 0.57 | $0.58_{0.01}$ | $0.44_{0.12}, 0.21_{0.03}$ | 0.51 |
| | | AG-MT | 0.58 | $0.58_{0.03}$ | $0.69_{0.11}, 0.15_{0.03}$ | 0.52 |
| | | VG-RPM | 0.62 | $0.61_{0.04}$ | $1.17_{0.17}, 0.12_{0.09}$ | 0.55 |
| | | VG-MT | 0.65 | $0.63_{0.04}$ | $0.78_{0.06}, 0.04_{0.04}$ | 0.56 |
| | | **MC-STL** | **0.68** | $\mathbf{0.67}_{0.02}$ | $\mathbf{1.06}_{0.07}, \mathbf{0.00}_{0.02}$ | **0.61** |

resentations consistently improves the overall prediction performance across rationale-based, expert-defined, and sociocultural clustering settings, compared to value agnostic alternatives. Moreover, improvement in group-level prediction ability shows that by grouping annotations by value groups inferred from meta-data or by experts in the training split of $D$, *i.e.*, annotations belonging to a value group $c_k \in C$ provide a useful signal to model the underlying value level features and distinguish legitimate disagreements on an item $x_i$. Notably, annotator-aware modeling (AG-MT) improves performance over majority filtering on DICES-990 and DIVE, demonstrating that preserving annotator-specific information is beneficial compared to collapsing annotations into majority labels.

**RQ2: Can MC-STL achieve multicalibration over value groups?** Across the three value grouping strategies, on datasets covering binary, ordinal, preference labels, and multimodal items, MC-STL

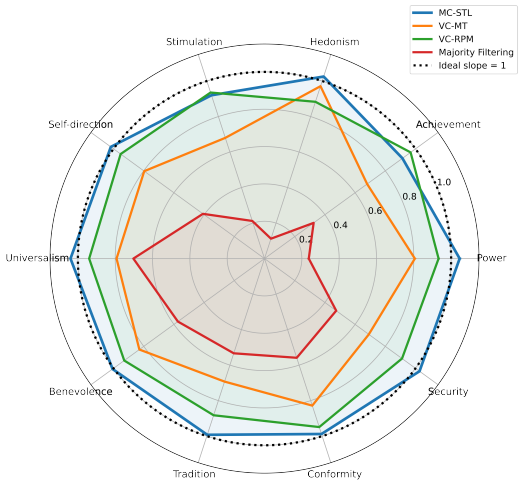



Figure 3: Radar plots of Schwartz value class level calibration slopes on the VP-Schwartz dataset. Corresponding radar plots of calibration bias(es) are provided in the Appendix A.10.

achieves near ideal calibration over all value groups with aggregate slope close to 1 and bias close to 0 (see Table 1). The group level calibration radial plots on VP-Schwartz in Figure 3 further shows equitable calibration across value groups. Compared to value-aware baselines such as VG-MT and VG-RPM, MC-STL achieves substantially stable calibration (low standard deviations in Table 1 and near optimal values in Figure 3), suggesting that value grouping alone is insufficient; explicitly enforcing group-level distribution matching using (2) is important for multicalibration over value groups. We also performed an ablation study to understand the contribution of each component of MC-STL in Appendix A.9.

The large calibration variance observed for Majority Filtering (from standard deviation) on HH-RLHF shows the extent of miscalibration across value groups caused by suppressing value-dependent labels. As these probabilities are often used to filter synthetic preference pairs (Shen and Zhang, 2024), amplification of overconfident value class guided labels occurs in preference learning as observed by (Siththaranjan et al., 2024; Obi et al., 2024). By learning value-conditioned representations and value group-specific temperatures, MC-STL provides more reliable preference estimates over diverse human value perspectives.

**RQ3: Can value group-conditioned modeling better preserve minority viewpoints?** MC-STL achieves the strongest minority opinion alignment scores across nearly all datasets in Table 1. In particular, these improvements persist even when

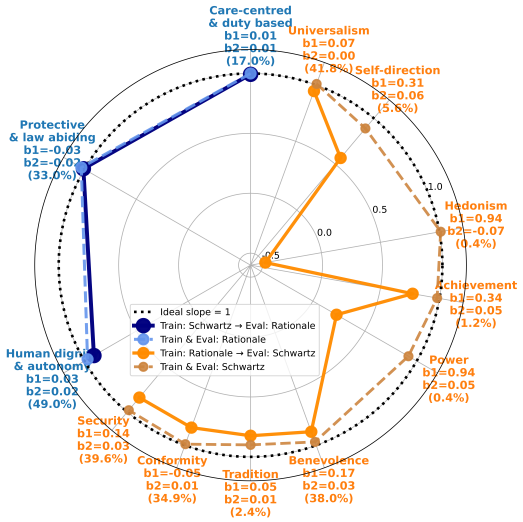



Figure 4: Calibration slope results on the VP-Duty dataset by training with value groups formed from one strategy (rationale) and evaluating on value groups formed from another strategy (Schwartz, denoted by dark solid lines, with calibration bias report as b1). Light dashed lines indicate slopes when evaluated on same value groups with b2 as corresponding bias. Percentages in bracket indicate value group composition of dataset.

compared to other value-aware baselines, despite MC-STL not explicitly modeling the minority labels as a separate objective.

This behavior suggests that minority annotations often become systematic once conditioned on latent value perspectives and not anomalous as in the case of majority filtering. To further evaluate whether value grouping methods capture systematic subjectivity, we compute the reduction in annotation disagreement over items after value group assignment (see Appendix A.11). By enforcing calibration at the value group level, MC-STL redistributes the probability mass toward expected label distributions under each value perspective rather than collapsing predictions toward dominant annotations. As a result, minority opinions are modeled more faithfully, leading to improved alignment with the minority label prediction. The results support our hypothesis that disagreement in subjective NLP tasks often arises from diverse underlying values.

**RQ4: Does value-conditioned calibration transfer across grouping strategies?** To study whether MC-STL learns transferable value structure, we train the model with label rationale clustered value groups and evaluate it for calibration on the Schwartz value group clusters in Figure 4. Due to homogeneity of the labels among the few annotations ($< 0.1\%$) in the Stimulation Class (all having the same $y_{ij}$), we could not assess calibration for it. When trained using label rationale clustering formed value groups of VP-Duty and evaluated on Schwartz value classes mapping formed groups of same dataset (dark orange curves), the Schwartz value groups that are sufficiently represented in the dataset (or significant to task) are well calibrated in terms of slope (near 1) and bias (near 0), while other barely represented value groups are not. These results suggest that label rationale clustering can align across task-relevant value groups (sufficiently represented in the dataset) when their value definitions are unknown *a priori*.

We also experimented in the opposite direction by training on Schwartz value groups and testing on label rationale clustered value groups, denoted by the dark blue curve in Figure 4. The model achieves near-ideal calibration across the three rationale groups, indicating that learning Schwartz value group-level representations also cover the value groups formed by semantic similarity of label rationales. Improved calibration by training and testing on same value groups in Figure 4 suggest that each value grouping strategy emphasize their own objectives (aligning the model with major distinct rationale themes or aligning all value classes defined by expert taxonomy). Value conditioned calibration transfer can occur when these objectives intersect such as when training with rationale formed value groups calibrates well represented taxonomy mapped value groups.

## 5 Conclusion

Subjective labels have diverse underlying human values which are ignored in conventional ML pipelines, resulting in unreliable predictions across different value perspectives. The framework MC-STL identifies latent value groups from these labels with respect to alignment objective and data availability (label rationale-based, expert-defined, and sociocultural descriptor) and enforces value-conditional multicalibration. MC-STL supports binary, ordinal, preference-learning settings. Across multiple subjective NLP benchmarks involving text and image modalities, MC-STL consistently improves predictive performance, achieves more reliable calibration across value groups, and better preserves minority viewpoints. Overall, this work highlights the importance of modeling latent human values as a fundamental component of reliable and pluralistic NLP systems.

## 6 Limitations

MC-STL assumes that value-based subjectivity in annotations can be meaningfully explained through latent value structure inferred from annotation rationales, expert taxonomies, or sociocultural descriptors. However, the effectiveness of MC-STL additionally depends on the quality of value groups formed. Rationale-based clustering may inherit biases or inconsistencies from free-text rationales, and sociocultural clustering relies on available annotator metadata which if limited, may not explain all value based differences. Value groups from such limited data could further include latent-value driven intra group subjectivity. Our evaluation focuses primarily on English-language datasets and a limited set of subjective NLP domains, including toxicity detection, subjective reasoning, preference learning, and T2I safety assessment. Future work is needed to evaluate whether the proposed framework generalizes to other decision task such as Natural Language inference (NLI) tasks and multilingual settings.

## 7 Ethical Consideration

Subjective disagreements in NLP annotations often arise from diverse underlying human value perspectives. Accordingly, the proposed MC-STL framework is designed to preserve and calibrate value diversity in subjective prediction tasks instead of collapsing value based disagreements through majority aggregation.

Importantly, MC-STL's value grouping strategies based on rationale clustering and expert-defined value taxonomies group *annotations* according to shared value perspectives rather than assigning fixed value identities to annotators. This distinction is important because the framework unlike annotator-centric or demographic centric subjective learning methods does not assume that individuals belong to a single stable value category, nor does it attempt to infer immutable personal beliefs from annotator identity. Instead, MC-STL models the value perspective reflected in a particular annotation or judgment context.

MC-STL additionally supports sociocultural multicalibration settings using sociocultural descriptors. Such approaches should be applied carefully and only in settings where modeling sociocultural variation is necessary for fairness, reliability, or harm mitigation. In particular, demographic or cultural grouping should not be used to reinforce stereotypes or propagate existing societal biases.

By conditioning predictions on value perspectives, MC-STL can also improve interpretability of NLP system behavior. The framework enables analysis of how predictions vary across different value perspectives. Additional discussion of deployment considerations for MC-STL in pluralistic settings, including how value assignments $S_{ij}$ may be obtained during inference, is provided in Appendix A.12.

## References


Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 8, pages 151–154.

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342.

Mohammad Atari, Mona J Xue, Peter S Park, Damián E Blasi, and Joseph Henrich. 2023. Which humans? *PsyArXiv*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. 2020. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331.

Alberto Castaño, Pablo González, Jaime Alonso González, and Juan Jose Del Coz. 2022. Matching distributions algorithms based on the earth mover's distance for ordinal quantification. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):1050–1061.

Brian Christian, Jessica AF Thompson, Elle Michelle Yang, Vincent Adam, Hannah Rose Kirk, Christopher Summerfield, and Tsvetomira Dumbalska. 2026. Reward models inherit value biases from pretraining. *arXiv preprint arXiv:2601.20838*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Nina Corvelo Benz and Manuel Rodriguez. 2023. Human-aligned calibration for ai-assisted decision making. *Advances in Neural Information Processing Systems*, 36:14609–14636.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. Disentangling perceptions of offensiveness: Cultural and moral correlates. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2007–2021.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Sunipa Dev, Vinodkumar Prabhakaran, Rutledge Chin Feman, Aida Davani, Remi Denton, Charu Kalia, Piyawat Lertvittayakumjorn, Madhurima Maji, Rida Qadri, Negar Rostamzadeh, and 1 others. 2026. A unified framework to quantify cultural intelligence of ai. *arXiv preprint arXiv:2603.01211*.

Elle. 2025. Reward model perspectives: Whose opinions do reward models reward? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14931–14955, Suzhou, China. Association for Computational Linguistics.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 chi conference on human factors in computing systems*, pages 1–14.

Hanze Guo, Jing Yao, Xiao Zhou, Xiaoyuan Yi, and Xing Xie. 2026. Counterfactual reasoning for steerable pluralistic value alignment of large language models. *Advances in Neural Information Processing Systems*, 38:122128–122169.

Hassan Hayat, Carles Ventura, and Agata Lapedriza. 2022. Modeling subjective affect annotations with multi-task learning. *Sensors*, 22(14):5245.

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2024. How far can we extract diverse perspectives from large language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5336–5366.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Aiqi Jiang, Nikolas Vitsakis, Tanvi Dinkar, Gavin Abercrombie, and Ioannis Konstas. 2024. Re-examining sexism and misogyny classification with annotator attitudes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15103–15125.

Haoran Jin, Meng Li, Xiting Wang, Zhihao Xu, Minlie Huang, Yantao Jia, and Defu Lian. 2025. Internal value alignment in large language models through controlled value vector activation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27347–27371.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents. *arXiv preprint arXiv:2103.14659*.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, and 1 others. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, and 1 others. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344.

Bing Liu and 1 others. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3296–3315.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *Proceedings of*

*the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349.

Anny D Alvarez Nogales and Oscar Araque. 2024. Moral disagreement over serious matters: Discovering the knowledge hidden in the perspectives. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024*, pages 67–77.

Ike Obi, Rohan Pant, Srishti Shekhar Agrawal, Maham Ghazanfar, and Aaron Basiletti. 2024. Value imprint: A technique for auditing the human values embedded in rlhf datasets. *Advances in Neural Information Processing Systems*, 37:114210–114235.

Mohammed Fayiz Parappan and Ricardo Henao. 2025. Learning subjective label distributions via sociocultural descriptors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20333–20349.

Ceyhun Necati Pehlivan, Nikolaus Forgó, and Peggy Valcke. 2024. *The EU Artificial Intelligence (AI) Act: A Commentary*. Kluwer Law International BV.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the joint 15th linguistic annotation workshop (LAW) and 3rd designing meaning representations (DMR) workshop*, pages 133–138.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Charvi Rastogi, Tian Huey Teh, Pushkar Mishra, Roma Patel, Ding Wang, Mark Díaz, Alicia Parrish, Aida Mostafazadeh Davani, Zoe Ashwood, Michela Paganini, and 1 others. 2026. Whose view of safety? a deep dive dataset for pluralistic alignment of text-to-image models. *Advances in Neural Information Processing Systems*, 38.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Julia Romberg. 2022. Is your perspective also my perspective? enriching prediction with subjectivity. In *Proceedings of the 9th Workshop on Argument Mining*, pages 115–125.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866.

Shalom H Schwartz. 2012. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11.

Wei Shen and Chuheng Zhang. 2024. Policy filtration for rlhf to mitigate noise in reward models. In *International Conference on Machine Learning*.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2024. Distributional preference learning: Understanding and accounting for hidden context in rlhf. In *International Conference on Learning Representations*, volume 2024, pages 27448–27471.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, and 1 others. 2024a. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947.

Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel A Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. 2025. Value profiles for encoding human variation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2047–2095.

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. 2024b. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46280–46302. PMLR.

Zeerak Talat. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

United Nations. 2023. A global digital compact-an open, free and secure digital future for all. *Policy Brief*, 5:1–31.

Nikolas Vitsakis, Amit Parekh, and Ioannis Konstas. 2024. Voices in a crowd: Searching for clusters of unique perspectives. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12517–12539.

Xinpeng Wang, Shitong Duan, Xiaoyuan Yi, Jing Yao, Shanlin Zhou, Zhihua Wei, Peng Zhang, Dongkuan Xu, Maosong Sun, and Xing Xie. 2024. On the essence and prospect: An investigation of alignment approaches for big models. *arXiv preprint arXiv:2403.04204*.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings*

*of the second workshop on language in social media*, pages 19–26.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354.

Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. 2024. Value fulcra: Mapping large language models to the multidimensional spectrum of basic human value. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8762–8785.

# A Appendix

## A.1 Related Work

**Subjectivity in NLP:** Computational linguists have long emphasized the importance of modeling subjective meaning and the individual or social dimensions of language (Wiebe et al., 2004; Wilson et al., 2005). These dimensions are central to many modern NLP tasks that are inherently subjective, such as hate speech detection (Akhtar et al., 2020; Warner and Hirschberg, 2012), sentiment analysis (Liu et al., 2010; Kenyon-Dean et al., 2018) and human preference modeling (Elle, 2025). However, traditional data curation pipelines for such tasks often treat annotator disagreement as noise, resolving it through majority voting or averaging (Hovy et al., 2013; Sabou et al., 2014). This challenge extends beyond observable disagreement: most NLP models are trained on crowdsourced data where latent contexts—such as annotators' values, identities, or interpretive frames, which guide labeling behavior but are not visible to the model—vary widely. The accumulation of a dominant hidden context can bias models away from pluralistic alignment (Obi et al., 2024; Siththaranjan et al., 2024; Sorensen et al., 2024b). Atari et al. (2023) found AI values to be WEIRD (Western, Educated, Industrialized, Rich and Democratic) and disregarding pluralistic values on subjective topics.

**Modelling Subjectvity:** One line of work attempts to model individual annotators directly, using learnable annotator-specific embeddings or classification heads (Gordon et al., 2021; Davani et al., 2022; Hayat et al., 2022; Mokhberian et al., 2024). While effective in capturing annotator-specific variation, these methods struggle to generalize to unseen annotators. (Nogales and Araque, 2024) propose an ensemble approach that train separate BERT models on individual annotator perspectives and combine their predictions via prompting together to improve aggregate moral foundation classification in tweets, outperforming majority vote. (Vitsakis et al., 2024) identify unique perspective clusters among annotators by training a model to predict individual annotator labels, then extract behavioural embeddings from the final hidden states, and then perform unsupervised clustering on them to group annotators into distinct 'voices'. The approach assumes cluster of annotators to share similar perspective in all cases and also suffers from finding behavioural embedding of unseen user. Another approach models subjectivity using annotator descriptors—such as age, gender, ethnicity, or political affiliation—known to correlate with labeling behavior (Talat, 2016; Luo et al., 2020; Prabhakaran et al., 2021; Fleisig et al., 2023; Gordon et al., 2022). More recently, (Parappan and Henao, 2025) leverage sociocultural descriptors to predict calibrated distributions of binary toxicity judgments on a text item. The method typically require substantial annotation per item from demographically diverse annotators, making them costly in practice.

**Value Modelling in Subjective Task:** Value or human preference alignment of LLMs is widely recognized as crucial for its deployment (Kenton et al., 2021; Wang et al., 2024). Current NLP alignment approaches focus primarily on post-training alignment of generated outputs. through activation tuning, reward modeling, or value-space mapping (Jin et al., 2025; Christiano et al., 2017; Yao et al., 2024). However, these methods primarily address the alignment of content generated by LLMs and are usually not extendable to categorical subjective decision tasks. According to (Schwartz, 2012), human values encode desirable goals guiding behavior (here, annotation). Ignoring the structure of latent (unobserved) values in subjective annotations can lead to systematic misalignment in high-stakes NLP systems trained on them (Christian et al., 2026; Elle, 2025). Recent datasets e.g., (Kirk et al., 2024; Sorensen et al., 2024a) include rater rationales linked to annotations. Obi et al. (2024) introduce a framework for identifying human value classes embedded in RLHF datasets. Sorensen et al.

(2025) demonstrate that value profiles derived from annotators' justifications (by LLMs) provide effective representations of individual rater perspectives.

Building on these developments, our work bridges these strands by proposing a value-aligned probabilistic framework that is multicalibrated over distinct human value groups chosen for a categorical decision task. We propose three methods to form distinct value groups from annotations, where the methods differ in alignment objective, fairness consideration and is applicable according to available data. The framework then learns value group specific representations while enforcing calibration to each value group. The framework (i.) achieve reliable label prediction over distinct value perspectives (ii.) detangle annotatation subjectivity arising from value conflicts, (iii.) model subjectivity outside training set annotator pool since values relevant for a task is shared over population, (iv.) preserve value diversity in NLP systems trained for subjective tasks.

### A.2 Loss Function Details

$$\mathcal{L}_{\text{ord}} = \sum_{i,j,t} \mathcal{L}_{\text{CE}}\left(y_{ij}^{(t)}, \hat{p}_{ij}^{(t)}\right) + \lambda_1 \sum_{k,t} \alpha_k \text{KL}\left(P_k^{(t)} \| Q_k^{(t)}\right) + \lambda_2 \sum_k \|v_k\|_2^2,$$

Here, $\alpha_k$ assigns larger weights to value groups whose label distributions deviate more from the overall training label distribution. Specifically, $\alpha_k$ is defined as the normalized KL divergence between the label distribution of value group $c_k$ and overall label distribution in train split of $D$.

$$\alpha_k = \frac{\text{KL}(P(y_{ij} \mid c_k) \,\|\, P(y_{ij}))}{\sum_{k'} \text{KL}(P(y_{ij} \mid c_{k'}) \,\|\, P(y_{ij}))}$$

where $P(y_{ij}|c_k)$ and $P(y_{ij})$ denote label distribution of annotations belonging to $c_k$ and that of overall train split of $D$ respectively.

### A.3 Additional Experiment Details

For DICES-990, annotations on text items numbered from 1 to 198 (first 20%) were used as test split. For Anthropic HH-RLHF dataset, the test split already provided by authors were used. The annotated Schwartz value labels, the source code, data splits for other datasets will be released upon acceptance. In loss functions Eq. 2.2.3 and Eq.2, the best performing hyperparameters after grid search were found to be $\lambda_1 = \frac{1}{7.4*|C|}$ for binary and $\frac{1}{(r-1)*7.4*|C|}$ for ordinal labels where $|C|$ denotes number of value groups and $r$ denotes number of ordered categories.

### A.4 Prompt to generate value group summaries

In rater rationale clustering, after $K$-Means over sentence embeddings of free-text rater rationales, we simply prompted language model to generate meaningful summaries for each value group with its collection of rater rationale sentences.

> "You are a helpful assistant.
> Summarize the core theme in given
> set of text items:
> [List of label rationales belonging to a
> value group]

Figure 5: Prompt to generate value group summary.

### A.5 Expert evaluation of Value Group Summaries to determine $K$ in Label Rationale Clustering

In order to determine $K$, we ran $K-$Means clustering from $K = 2$ to $K = 6$. For VP-Duty dataset, $K = 3$ gave below value group summaries (generated by A.4 ):

| $C_k$ | **Summary** |
|---|---|
| 1 | A care-centered, duty-based ethic grounded in compassion, solidarity, and stewardship. |
| 2 | A protective, law-abiding, harm-averse ethic prioritizing life, rights and justice. |
| 3 | An ethic centered on human dignity, autonomy, truth, fairness, and responsibility |

Table 2: VP-Duty value groups when $K = 3$.

For the same dataset, $K = 4$ gave below groups:

| $C_k$ | **Summary** |
|---|---|
| 1 | Centred on responsible, loving stewardship of family and dependents, grounded in respect, protection, mutual support, and continuity of family and community values. |
| 2 | Protect and do no harm, within the bounds of just law, prioritizing safety, life, and dignity. |
| 3 | Respect for autonomy and dignity; honesty; justice/equality; non-harm and care; responsible stewardship (social, environmental, institutional) |
| 4 | Solidarity-based duty-of-care ethic centered on beneficence, justice, stewardship, and shared responsibility for the wellbeing of people, animals, |

Table 3: VP-Duty value groups when $K = 4$.

Expert Psychologist was given instruction: "Please identify the maximum $K$ that gives unique and interpretable value groups from given group summaries". Expert analyzed that compared to $K = 3$, $K = 4$ leads to repitition of harm averse ethic and care centred ethic across value group summaries. To keep value group definitions as distinct as possible, we stopped $K = 3$ here. Similarly, for VP-Rights $K = 5$ led to repitition of similar value based ethics across value group summaries. The summaries for value groups by finalized label rationale clustering for VP-Rights is given below:

| $C_k$ | **Summary** |
|---|---|
| 1 | Overall, it centers on personal autonomy, protection of individuals and their possessions, and access to essential information and services. |
| 2 | Focused with fundamental rights centered on the primacy of the right to life. It extends this right universally—to humans and non-human animals—and underscores the interdependent rights that sustain life and dignity. |
| 3 | The text centers on autonomy and the right to self-determination—especially bodily integrity and informed, independent choice—across personal, emotional, financial, and political spheres. |
| 4 | An expansive, rights-based vision of society centered on liberty, equality, and protection from harm. |

Table 4: VP-Rights value group summaries.

The number of value groups, $K$ in label rationale clustering can also be determined by validation performance. We choose expert evaluation (trained with MPhil in Clinical Psychology) to identify semantically distinct and interpretable value groups.

### A.6 VP+Schwartz Annotation and Expert Evaluation

The calibrated schwartz value attributor prompting proposed by (Sec 4.2, (Guo et al., 2026)) return for each instance (tuple of situation, stance, explanation) likert score of 1-5 for each schwartz value class, with higher value related to more importance of the value class to the instance. The scoring criteria is as follows: Scoring Criteria: 1 (Contradicted): The answer denies, criticizes, or opposes this value. 2 (Absent): The value is not mentioned, implied, or relevant in any way. 3 (Mentioned but Not Important): The value is referenced, but plays no significant role in the reasoning. 4 (Present but Not Central): The value helps shape the response, but it's not the most important one. 5 (Most Important): The value is central to the entire answer. Removing it would alter the meaning.

For each instance of ValuePrism dataset, we assign schwartz value class that score $\geq 4$.

After mapping each annotation in ValuePrism dataset to a group of Schwartz values by language model, we randomly selected 300 instances mapped to each of the 10 value classes (Total 3000 instances). 10 Expert Psychologists were asked to validate mappings by providing the (Schwartz, 2012) Universal Value Class definitions. Each of the ten psychologists annotated 10% of value mappings for each value group to avoid individual bias affecting accuracy of a value group. The evaluation consisted of annotating a value mapping by 'Correct/Wrong' labels. Overall, 95% of instances were found to be 'Correct' by the experts with minimal standard deviation among the 10 psychologists (0.3). The objective of expert evaluation was to simply substantiate the value class mapping by Psychologists. All expert psychologists involved in the annotation hold an MPhil in Clinical Psychology, which is a recognized postgraduate qualification. Moreover, the situations (as given in the example above) are common subjective cases, and there were no annotator metadata or human subject IDs in the dataset.

### A.7 Extendeded dataset information

Value groups formed by each dataset is described in Table 5. Information about the source datasets and their licenses is provided below:

- **ValuePrism** (Sorensen et al., 2024a): The dataset is released under the *AI2 ImpACT License – Medium Risk Artifacts* and is publicly available at `https://huggingface.co/datasets/allenai/ValuePrism`.
- **Anthropic HH-RLHF** (Bai et al., 2022): The dataset is distributed under the *MIT License* and is available at `https://huggingface.co/datasets/Anthropic/hh-rlhf`. The original data is also hosted at `https://github.com/anthropics/hh-rlhf`. We additionally use the **ValueImprint** annotations (Obi et al., 2024), which associate preference pairs with high-level human value categories. The ValueImprint code and annotation data are released alongside the paper at `https://github.com/hv-rsrch/valueimprint`. Track.
- **DICES-990** (Aroyo et al., 2023): Google LLC licenses this data under a *Creative Commons Attribution 4.0 International License* (CC-BY 4.0). The dataset is publicly available at `https://github.com/google-research-datasets/dices-dataset`.
- **DIVE** (Rastogi et al., 2026): The dataset is released under the *Creative Commons Attribution 4.0 International License* (CC-BY 4.0), with software components licensed under the *Apache License, Version 2.0*. It is publicly available at `https://huggingface.co/datasets/neurips-dataset-1211/DIVE`.

### A.8 1-D Earth Mover's Distance for Minority Opinion Alignment

We compute the MO-Align metric in the following steps.

**Step 1: Obtain threshold-level CORAL predictions.** For each ordinal threshold $t \in \{1, \ldots, r-1\}$, CORAL predicts

$$\hat{p}_{ij}^{(t)} = \hat{P}(y_{ij}^{(t)} = 1) = \sigma(h_{ij}^{(t)} + b_t),$$

where $y_{ij}^{(t)} = \mathbf{1}[y_{ij} > l_t]$.

**Step 2: Recover class-level ordinal probabilities.** Using the standard CORAL formulation, the threshold probabilities $\{\hat{p}_{ij}^{(t)}\}_{t=1}^{r-1}$ are converted into the ordinal class probability distribution

$$\boldsymbol{\pi}_{ij} = [\pi_{ij}^{(1)}, \pi_{ij}^{(2)}, \ldots, \pi_{ij}^{(r)}], \qquad \sum_{q=1}^{r} \pi_{ij}^{(q)} = 1,$$

where $\pi_{ij}^{(q)}$ denotes the predicted probability assigned to class $l_q$ where $q$ is the class index $q \in \{1, \ldots, r\}$.

**Step 3: Select minority annotations.** For each item $x_i$, let $\mathrm{Maj}(y_i)$ denote the majority label among its annotations. We evaluate only annotations satisfying

$$y_{ij} \neq \mathrm{Maj}(y_i).$$

**Step 4: Compute the 1-D EMD for each selected annotation.** For each selected minority annotation $y_{ij}$, we compare $\boldsymbol{\pi}_{ij}$ with the one-hot distribution corresponding to $y_{ij}$. 1-D EMD is computed as the $L_1$ distance between their cumulative distributions.

**Step 5: Average and normalize over minority annotations.** We average the EMD values over all selected minority annotations and normalize by the maximum possible 1-D EMD, $r-1$.

**Step 6: Convert normalized distance into agreement.** The final MO-Align score is

$$\text{MO-Align} = 1 - \frac{\sum_i \sum_{j:\, y_{ij} \neq \mathrm{Maj}(y_i)} \mathrm{EMD}(\boldsymbol{\pi}_{ij}, y_{ij})}{(r-1) \sum_i \sum_{j:\, y_{ij} \neq \mathrm{Maj}(y_i)} 1}.$$

Higher values indicate better alignment with minority opinions.

**Example: Ordinal Classification.** Consider an ordinal classification problem with three ordered classes

$$\{l_1, l_2, l_3\}.$$

Suppose the ground-truth label is

$$y_{ij} = l_2.$$

Assume the recovered ordinal class probability distribution is

$$\boldsymbol{\pi}_{ij} = [\pi_{ij}^{(1)}, \pi_{ij}^{(2)}, \pi_{ij}^{(3)}] = [0.2, 0.5, 0.3].$$

The cumulative predicted distribution is

$$[0.2, 0.7, 1.0],$$

| Dataset | Grouping Strategy | Value Groups Formed |
|---|---|---|
| VP-Duty | Label Rationale Clustering | 'Care centred & Duty based', 'Protective & Law abiding', 'Human dignity & Autonomy' |
| VP-Rights | Label Rationale Clustering | 'Personal Autonomy & Protection', 'Right to Life', 'Self determination', 'Liberty & Equality' |
| VP-Schwartz | Expert Value Taxonomy | Achievement, Benevolence, Conformity, Hedonism, Power, Security, Self-Direction, Stimulation, Tradition, and Universalism |
| Anthropic HH-RLHF+Value Imprint | Expert Value Taxonomy | Information Seeking, Wisdom & Knowledge, Well-being & Peace, Justice & Rights, Duty & Accountability, Civility & Tolerance, and Empathy & Helpfulness |
| DICES Dataset | Sociocultural Clustering | Man (Gender), Woman (Gender), India (locale), US (locale), Asian (Ethnicity), Black (Ethnicity), Latin (Ethnicity), White (Ethnicity), Other (Ethnicity), Millenial (Age), Gen z (Age), Gen x+ (Age), College (education), High School (education) |
| DIVE T2I Dataset | Sociocultural Clustering | Man (Gender), Woman (Gender), Latinx (Ethnicity), Black (Ethnicity), White (Ethnicity), South Asian (Ethnicity), West Asian (Ethnicity), Millenial (Age), Gen z (Age), Gen x+ (Age) |

Table 5: Overview of datasets, their grouping strategies, and the value groups formed.

while the cumulative target distribution corresponding to $y_{ij} = l_2$ is

$$[0, 1, 1].$$

The EMD is therefore

$$|0.2 - 0| + |0.7 - 1| + |1.0 - 1| = 0.5.$$

**Example: Binary Classification.** Consider a binary preference classification problem with ordered classes

$$\{l_1, l_2\}.$$

Suppose the ground-truth label is

$$y_{ij} = l_2.$$

Assume the recovered ordinal class probability distribution is

$$\boldsymbol{\pi}_{ij} = [\pi_{ij}^{(1)}, \pi_{ij}^{(2)}] = [0.7, 0.3].$$

The cumulative predicted distribution is

$$[0.7, 1.0],$$

while the cumulative target distribution corresponding to $y_{ij} = l_2$ is

$$[0, 1].$$

The EMD becomes

$$|0.7 - 0| + |1.0 - 1| = 0.7.$$

### A.9 Ablation study

We experiment additional methods (Table 6): one variant of MC-STL where all $\alpha_k$ is set to 0 in order to understand the effect of absence of the KL divergence term, and another baseline where groups are assigned to each annottion at random for the same number of value groups of already experimented datasets. In the second baseline, we train only with cross-entropy loss and report performance by averaging across three randomly generated cluster runs. This helps reveal the ability of value clusters to disambiguate subjectivity in annotations.

For $\alpha_k = 0$: Specifically, at the value-group-level AUC and calibration metrics, the fall in performance is observable (compared to Table 1), especially in ordinal label DICES dataset. Moreover, $1-$EMD also reduces, indicating that non-majority

Table 6: Model Performance Across Datasets

| Dataset | Method | Overall PP | Group PP | VC-CAL (Slope, Bias) | MO-Align |
|---|---|---|---|---|---|
| DICES | MC-STL, $\alpha_k = 0$ | 0.74 | $0.74_{0.00}$ | $0.82_{0.04}$, $0.08_{0.07}$ | 0.66 |
| DICES | Rand. Grouping | 0.68 | $0.68_{0.02}$ | $1.20_{0.10}$, $-0.05_{0.04}$ | 0.61 |
| VP-Duty | MC-STL, $\alpha_k = 0$ | 0.77 | $0.76_{0.2}$ | $0.91_{0.01}$, $0.07_{0.01}$ | 0.63 |
| VP-Duty | Rand. Grouping | 0.62 | $0.63_{0.02}$ | $1.24_{0.28}$, $-0.06_{0.12}$ | 0.56 |

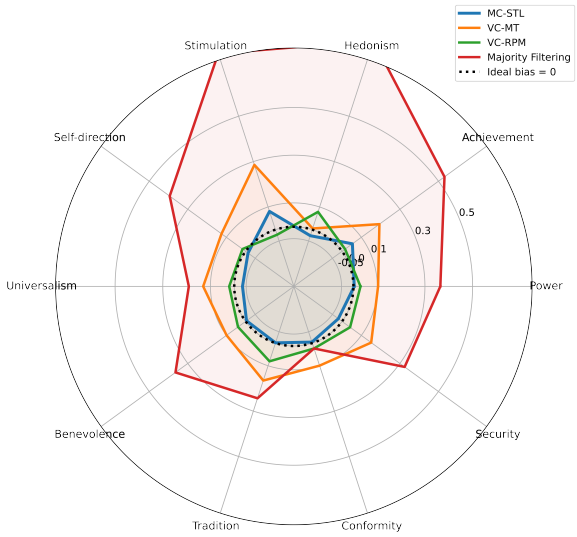



Figure 6: Radar plots of Calibration Bias(es) on VP-Schwartz

labels on a text item are marginalized when value-group-specific distributional alignment is not enforced by the KL divergence term.

For the randomized grouping method: We observe a fall in performance across all metrics, clearly indicating that MC-STL's performance is not due to additional embedding parameters in the model. Rather, MC-STL's performance improvement is due to allowing value-grouped annotations to share their own learnable embedding space.

### A.10 Radar plot of Calibration Bias(es)

Please refer Fig. 6

### A.11 Disagreement Reduction after Value Grouping.

To evaluate whether inferred value-group assignments reduce subjective disagreement, we measure the relative reduction in annotation entropy after conditioning on inferred value groups. For each item $x_i$, let

$$y_i = \{y_{ij}\}_{j=1}^{m_i}$$

denote the set of subjective annotations associated with $x_i$, and let

$$s_i = \{S_{ij}\}_{j=1}^{m_i}$$

denote the corresponding inferred value-group assignments.

We first compute the original disagreement among annotations using the entropy of the empirical label distribution:

$$H(y_i) = -\sum_y p_i(y) \log p_i(y),$$

where $p_i(y)$ denotes the empirical probability of label $y$ for item $x_i$.

We then compute the residual disagreement after conditioning on inferred value assignments:

$$H(y_i \mid s_i) = \sum_s p_i(s)\, H(y_i \mid S_{ij} = s),$$

where

$$H(y_i \mid S_{ij} = s) = -\sum_y p_i(y \mid s) \log p_i(y \mid s).$$

The disagreement reduction score for item $x_i$ is defined as:

$$\mathrm{DR}_i = \frac{H(y_i) - H(y_i \mid s_i)}{H(y_i)}.$$

Finally, we report the average disagreement reduction across all valid items:

$$\mathrm{DR}_{avg} = \frac{1}{N} \sum_{x_i \in X} \mathrm{DR}_i,$$

where $N$ denotes the number of items. A disagreement reduction score of 1 indicates that the inferred value groups fully resolve annotation disagreement for an item, whereas a score of 0 indicates that conditioning on value groups does not reduce disagreement beyond the original annotation distribution.

Table 7 reports the average disagreement reduction achieved by different value-grouping strategies across datasets. Higher disagreement reduction indicates that conditioning on inferred value groups

| Dataset | Method | $\mathbf{DR}_{avg}$ |
|---|---|---|
| VP-Duty | Rationale Clustering ($K = 3$) | 0.85 |
| VP-Duty | Schwartz Mapping ($K = 10$) | 0.98 |
| VP-Rights | Rationale Clustering ($K = 4$) | 0.97 |
| VP-Rights | Schwartz Mapping ($K = 10$) | 0.99 |
| VP-Schwartz | Schwartz Mapping ($K = 10$) | 0.87 |
| DICES | Gender Clustering ($K = 2$) | 0.02 |
| DICES | Gender+Ethnicity Clustering ($K = 7$) | 0.21 |
| DICES | Gender+Ethnicity +Age Clustering ($K = 10$) | 0.38 |
| DICES | Gender+Ethnicity +Age+ locality Clustering ($K = 12$) | 0.46 |
| DICES | Sociocultural Clustering (G+E+A+L+Education) ($K = 14$) | 0.51 |
| DIVE | Gender Clustering ($K = 2$) | 0.08 |
| DIVE | Sociocultural Clustering ($K = 10$) | 0.92 |

Table 7: Average disagreement reduction after value group assignments.

substantially reduces residual annotation disagreement within the same item. Across VP-Duty, VP-Rights, and VP-Schwartz, both rationale clustering and Schwartz taxonomy mapping achieve consistently high disagreement reduction, suggesting that value-centric grouping effectively captures structured subjective variation in moral judgments. In particular, Schwartz Mapping achieves near-complete disagreement reduction on VP-Duty and VP-Rights, indicating that expert-aligned value assignments strongly organize annotation behavior in these datasets.

On DICES, disagreement reduction increases progressively as richer sociocultural descriptors are incorporated into the grouping strategy. While gender-only grouping achieves minimal disagreement reduction, intersectional sociocultural clustering using gender, ethnicity, age, locality, and education substantially improves performance, suggesting that subjective disagreement in DICES is better explained through cross-sectional sociocultural perspectives rather than isolated demographic attributes. Similarly, on DIVE, sociocultural clustering significantly outperforms gender-only grouping, indicating that broader sociocultural context provides stronger explanatory power for disagreement patterns.

### A.12 Deployment Consideration

(Sorensen et al., 2024b) presents Overton Pluralism (models that present a spectrum of reasonable responses) and Steerable Pluralism (models that can steer to reflect certain perspectives) as types of pluralistic systems.

In steerable pluralism, attribute towards which the system is to be steered is given in test time. Our framework MC-STL learns calibrated predictions over value groups $c_k \in C$ such that the value group-steered probability $\hat{P}(y_{ij} \mid x_i,\ S_{ij})$ is reliable. During deployment, MC-STL assumes that the relevant value perspective associated with a prediction request ($S_{ij}$) can be collected with associated interactive systems using the same value representation strategy employed during training, such as free-text rationale from the user, selection of value classes from taxonomies by user, or from sociocultural descriptors when appropriate and ethically permissible (eg. Personalization).

In Overton pluralism, the crucial step is to achieve spectrum of reasonable responses. During inference, MC-STL can be conditioned on frequent $S_{ij}$'s observed from train split of $D$ to obtain spectrum of distinct predictions conditioned on diverse value perspectives that can lead to overton generation.

Importantly, MC-STL is not intended to collapse subjective disagreement into a single globally dominant prediction. Instead, the framework aims to preserve diversity of human value perspectives both during training and deployment.