Labels have Human Values: Value Calibration of Subjective Tasks

Anonymous ACL submission

Abstract

Building NLP systems for subjective tasks requires one to ensure their alignment to contrasting human values. We propose the MultiCalibrated Subjective Task Learner framework (MC-STL), which clusters annotations into identifiable human value clusters by three approaches (similarity of annotator rationales, expert-value taxonomies or rater's sociocultural descriptors) and calibrates predictions for each value cluster by learning cluster-specific embeddings. We demonstrate MC-STL on several subjective learning settings, including ordinal, binary and preference learning predictions, and evaluate it on multiple datasets covering toxic chatbot conversations, offensive social media posts and human preference alignment. The results show that MC-STL consistently outperforms the baselines that ignore the latent value structure of the annotations, delivering gains in discrimination, value-specific calibration, and disagreement-aware metrics.

1 Introduction

As NLP systems are widely deployed in content-sensitive domains, there is a growing call for AI value alignment (Pehlivan et al., 2024; United Nations, 2023). AI value alignment is the idea of designing AI systems so that their goals, decisions, and behaviors are consistent with pluralistic human values (Sorensen et al., 2024b). However, conventional machine learning practices often obscure the value differences present in the data. This occurs both explicitly, when annotator disagreement is treated as noise and removed through aggregation (majority voting) (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 overrepresented in the data (Obi et al., 2024; Siththaranjan et al., 2023); see Figure 1. Although fully capturing every latent context underlying each label is challenging, it is important to develop systems that remain aligned with the human value differences relevant to a given task or dataset.

This challenge is particularly important in subjective tasks—such as toxicity detection, sentiment analysis, and human preference modeling, where label variation often reflects genuine differences in task interpretation, sociocultural identity, and personal values, rather than annotation error. Existing work has addressed subjectivity by modeling each annotator's preferences independently (Davani et al., 2022), or by representing subjectivity through demographic or sociocultural descriptors of annotators (Fleisig et al., 2023). Sorensen et al. (2025) instead model individual annotators via the free-text value statements obtained from their annotations. We provide a detailed description of the Related Work in Appendix A.1.

In this work, we build on this direction by explicitly grouping crowd-sourced annotations in the training data from distinct human value perspectives reflected in the value justifications behind each label and then learning embeddings for each value group, enabling a richer modeling of subjectivity that aligns model behavior with task-relevant value groups. Moreover, we encourage multicalibration (Hébert-Johnson et al., 2018) for value groups by aligning uncertainty of the model predictions with the true label distribution for all value groups. In principle, a predictor is considered calibrated when its predicted probabilities match true likelihood of desired outcome, and multicalibration is considered a sufficient condition for human alignment (Corvelo Benz and Rodriguez, 2023). Although recent work in subjective tasks has started to incorporate calibration, (Parappan and Henao, 2025; Sorensen et al., 2025), multicalibration remains underexplored despite its relevance for subjective alignment.

To the best of our knowledge, the proposed MultiCalibrated Subjective Task Learner (MC-STL) is

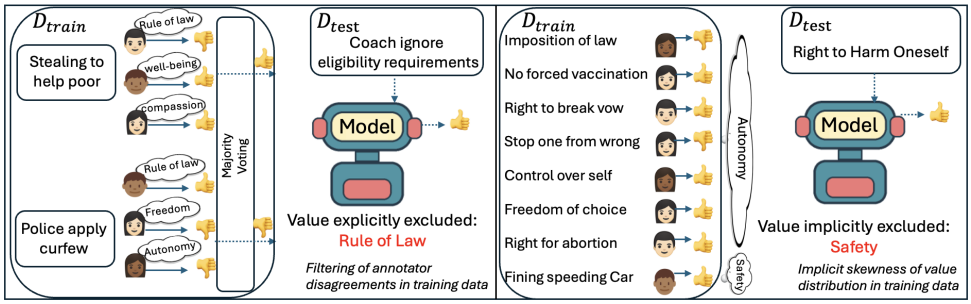

Figure 1: Example of how subjective label predictions are affected by hidden human values (denoted as thought clouds) (left) when annotator disagreements are filtered out by majority voting or (right) when minority values are underrepresented and likely unaccounted by the model in crowd-sourced annotation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

the first framework to model heterogeneity across human value groups that is responsible for subjectivity in labels for a given task or dataset, while encouraging calibrated predictions across value groups. Our contributions are as follows.

- We introduce a novel framework that models subjectivity for binary and ordinal labels while achieving multicalibration across human value cluster identities formed by: i) semantic similarity of free-text rater rationales behind annotations; ii) expert-defined human value taxonomies (*e.g.*, Schwartz); or iii) cross-sectional sociocultural identities of annotators.
- We present a thorough evaluation of overall and cluster-level discrimination, group-specific calibration, and instance-level disagreement alignment on diverse subjective datasets, including detection of conversational toxicity and offensiveness, and pairwise human preference datasets.

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 full dataset. Each annotation $y_{ij} \in \mathbf{Y}$ represents the j-th label assigned to x_i. We allow annotations y_{ij} to take values from an ordinal label space (*e.g.*, *Support*, *Neutral*, *Oppose*) to accommodate ambiguous or neutral judgments that commonly arise in subjective annotation tasks.

To model the latent subjectivity underlying these annotations, we draw on the principle of Schwartz (2012) that human values guide action and judgment. We assume that the observed subjectivity in the dataset arises from K latent human value groups, each reflecting distinct value-driven perspectives that influence annotation behavior. We explore multiple representational choices for modeling these value groups, where the often unknown number of groups K also varies depending on the chosen representation, described in the following.

Semantic similarity of free-text rater rationales Given free-text explanations or value statements underlying each annotation y_{ij} (either collected directly during annotation or retrospectively inferred by prompting LLMs with rating demonstrations (Sorensen et al., 2025; Hayati et al., 2024; Parappan and Henao, 2025), conventional semantic clustering methods can be applied to group annotations $y_{ij} \in \mathbf{Y}$ into K latent value groups. Each cluster can then be interpreted by summarizing the associated value statements, using either LLMs or domain experts, *e.g.*, psychologists. The number of value groups K can be determined by inspecting these summaries and identifying distinct, interpretable value profiles. It is worth noting that as rater rationales may change from dataset to dataset, thus clusters obtained from these rationales on a particular dataset may not be useful for others.

Expert-defined value taxonomies In this setting, we leverage established human value taxonomies and map annotations to predefined value categories, either during annotation or by mapping their associated free-text rater rationales. For example, Schwartz (2012) propose ten universal human-value categories, a subset of which may be used to guide annotation behavior and intent for a given task. Expert-defined value taxonomies are also widely adopted in human preference modeling, where annotators select preferred or ideal responses to train reward models that align LLMs with hu-

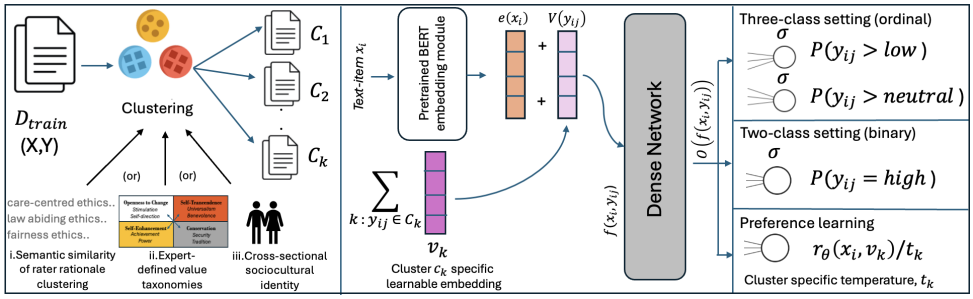

Figure 2: Overview of the proposed MC-STL framework. (left) Three methods for partitioning of label annotations into value-based clusters. (middle) The MC-STL backbone architecture. (right) Task-specific output heads.

man preferences. In such settings, high-level value dimensions (*e.g.*, *Helpfulness*, *Harmlessness*) are commonly used to structure labeling efforts.

Note that using expert defined value taxonomies to cluster annotations have the benefit of making such clusters enjoy generalizable meaning beyond the context of a specific task or dataset.

Cross-sectional sociocultural identity In certain subjective tasks, such as toxicity detection, ensuring fairness over demographic groups is a critical concern. Existing work has shown that in such settings sociocultural identities can substantially influence annotation behavior (Davani et al., 2024b). In sociocultural clustering, one pairs each annotation y_{ij} with cross-sectional sociocultural groups defined by annotator attributes, *e.g.*, age group, ethnicity and geographic location. These attributes act as proxy for the underlying value-driven subjectivity that shapes annotation decisions.

Using the three representations described above, we partition the label annotations $\mathbf{Y} \in \mathcal{D}_{\text{train}}$ in the training set $\mathcal{D}_{\text{train}}$ into K value-based clusters. Let $\mathcal{C} = \{C_k\}_{k=1}^{K}$ denote the resulting set of value clusters, where each C_k groups annotations sharing a common value-driven perspective, as illustrated in Figure 2(a). Clusters may follow a mixed membership model, where an annotation $y_{ij} \in \mathbf{Y}$ can belong to more than one cluster in $\mathcal{C}$. For example, an instance can be mapped to human value groups "Universalism" and "Benevolence".

Let S_{ij} be the set of value groups to which y_{ij} is assigned, then $S_{ij} = \{C_k \mid y_{ij} \in C_k\}$. We propose the MultiCalibrated Subjective Task Learner (MC-STL), a model that estimates $P(y_{ij} \mid x_i,\ S_{ij})$, and importantly is multicalibrated w.r.t. each value cluster in $\mathcal{C}$. At inference time, the model produces a probability distribution over label classes for a given text item, conditioned on the specified value perspective cluster(s).

2.2 MC-STL Architecture

The proposed MC-STL model adopts a simple yet novel architecture to model human subjectivity using value clustering, as described in Figure 2(b). 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 cluster $C_k \in \mathcal{C}$, 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 aggregating the embeddings of all value clusters associated with such annotation as

$$V(y_{ij}) = \textstyle\sum_{k:\, y_{ij} \in C_k} v_k \,. \qquad (1)$$

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 additive formulation preserves the dimensionality of the representation while allowing each annotation to be influenced by multiple value clusters. Importantly, in frameworks such as Schwartz's value theory, an annotation may be associated with more than one human value group, which our architecture naturally supports through (1).

The value-aware representation $f(x_i, y_{ij})$ is then passed to a feed-forward neural network. Specifically, let $o(f(x_i, y_{ij}))$ denote a dense network transformation applied to $f(x_i, y_{ij})$, but excluding the bias parameter(s) in its last layer. We employ task-specific output heads depending on the setting: i) a binary prediction head for two-class scenarios, and ii) a three-class prediction head for ordinal labels that include an ambiguous or neutral class.

Two-class setting For binary classification tasks, where labels $y_{ij} \in \{0, 1\}$, we apply a sigmoid activation and bias to the model output to obtain

class probabilities as

$$\hat{p}_{ij} = \hat{P}(y_{ij} = 1) = \sigma\big(o(f(x_i, y_{ij}) + b)\big),$$

where b denotes the bias and $\hat{p}_{ij}$ is introduced for notational simplicity.

Three-class setting The presence of a *Neutral* or *Ambiguous* label is common in subjective annotation tasks, as the personal perspective of annotators often fits between opposing binary judgments, *e.g.*, the "Either" or "Unsure" labels in (Sorensen et al., 2024a; Aroyo et al., 2023). Despite this, such labels and their ordinal relationship with the classes at the extremes are often ignored in the subjective learning literature. To model the ordinal structure, we adopt the consistent rank logits (CORAL) framework proposed by Cao et al. (2020).

Let $y_{ij} \in \mathbf{Y}$ take values in $\{\textit{Low}, \textit{Neutral}, \textit{High}\}$, which correspond to an ordinal label space with $n = 3$ ordered categories. CORAL reformulates this task as $n - 1$ binary sub-tasks defined as

$$y_{ij}^{(1)} = \mathbf{1}[y_{ij} > \textit{Low}]\,, \quad y_{ij}^{(2)} = \mathbf{1}[y_{ij} > \textit{Neutral}]\,,$$

where $\mathbf{1}[\cdot]$ denotes the indicator function. In our model, these sub-tasks share the same dense network transformation $o(f(x_i, y_{ij}))$, differing only in their bias terms at the final layer. Given the value-aware representation $f(x_i, y_{ij})$, the model outputs are

$$\hat{p}_{ij}^{(1)} = \hat{P}\left(y_{ij}^{(1)} = 1\right) = \sigma(o(f(x_i, y_{ij}) + b_1)\,,$$
$$\hat{p}_{ij}^{(2)} = \hat{P}\left(y_{ij}^{(2)} = 1\right) = \sigma(o(f(x_i, y_{ij}) + b_2)\,,$$

where b_1 and b_2 are sub-task-specific biases.

This formulation enforces rank monotonicity, such that $\hat{p}_{ij}^{(1)} \geq \hat{p}_{ij}^{(2)}$, as proved by Cao et al. (2020, Theorem 1). Note that we only consider ordinal settings with three classes, but our approach can be easily extended to higher-class ($n > 3$) settings.

Finally, class-level probabilities are recovered from the predictions as: $\hat{P}_{\textit{Low}} = 1 - \hat{p}_{ij}^{(1)}$, $\hat{P}_{\textit{Neutral}} = \hat{p}_{ij}^{(1)} - \hat{p}_{ij}^{(2)}$ and $\hat{P}_{\textit{High}} = \hat{p}_{ij}^{(2)}$. Also, by construction, these probabilities sum up to 1.

2.3 Loss Function

Our training objective is twofold: i) to ensure that the predicted probabilities, $\hat{p}_{ij}$ align with the corresponding ground truth labels, $y_{ij} \in \mathbf{Y}$, and ii) to ensure that the empirical distribution Q_k of the predicted probabilities for each value cluster $C_k \in \mathcal{C}$ reflects its corresponding true label distribution, P_k obtained from ground truth labels $\{y_{ij} \mid y_{ij} \in C_k\}$. To achieve this, correspondingly, we employ a composite loss function consisting of cross-entropy and the distribution-matching Kullback-Leibler (KL) divergence. To incorporate group-level matching, the KL-divergence for each value cluster $C_k \in \mathcal{C}$, aligns the average predicted distribution of annotations belonging to the cluster with the empirical label distribution observed for that cluster. Moreover, to control for overfitting due to the high-dimensional nature of v_k, we also consider an L2 regularizer for value embeddings. More specifically, in the two-class setting, the loss function is:

$$\begin{aligned}\mathcal{L}_2 &= \textstyle\sum_i \sum_j \mathcal{L}_{\text{CE}}(y_{ij}, \hat{p}_{ij}) \\ &\quad + \lambda_1 \textstyle\sum_k \alpha_k \text{KL}(P_k \| Q_k) + \lambda_2 \sum_k ||v_k||_2^2\,,\end{aligned} \tag{2}$$

where $\mathcal{L}_{\text{CE}}(y_{ij}, \hat{p}_{ij})$ is the binary cross-entropy loss between y_{ij} and $\hat{p}_{ij}$. $\text{KL}(P_k \| Q_k)$ is the Kullback-Leibler (KL) divergence between the two (empirical) binomial distributions, P_k formed by ground-truth ratings and Q_k formed from ratings from probabilistic predictions belonging to value cluster $C_k \in \mathcal{C}$. Specifically, we write

$$\begin{aligned}\text{KL}(P_k \| Q_k) &= n_k \bar{y_k} \cdot \ln\left(\frac{\bar{y_k}}{\hat{p}'_k}\right) \\ &\quad + n_k(1 - \bar{y_k}) \cdot \ln\left(\frac{1 - \bar{y_k}}{1 - \hat{p}'_k}\right)\,.\end{aligned}$$

where n_k is the number of annotations in value cluster C_k, and $\bar{y}_k$ and $\hat{p}'_k$ are the averages of $\{y_{ij}\}_{ij \in C_k}$ and $\{\hat{p}'_{ij}\}_{ij \in C_k}$, respectively.

Although we only have discrete realizations ($y_{ij} \in \{0, 1\}$) from P_k to obtain $\bar{y}_k$, we do have predicted probabilities for the realizations of Q_k. Borrowing from Parappan and Henao (2025), in which they soft-threshold predictions before aggregating them to obtain $\hat{p}'_k$. Specifically, we calculate $\hat{p}'_k$ after converting each predicted probability into *approximately* binary labels using the ground-truth value cluster-level mean $\bar{y}_k$ as a reference using

$$\hat{p}'_k = \frac{1}{2} \frac{\sum_{ij \in C_k} (1 + \tanh(l \cdot (\hat{p}_{ij} - \bar{y}_k))}{n_k}\,, \tag{3}$$

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.

Finally, in (2) $\|v_k\|$ is an L_2 regularizer for the value embedding associated with cluster $C_k \in \mathcal{C}$, α_k is a constant set to KL divergence between

the label distribution in C_k and that of the whole D_{train}. This allows the KL term for each cluster to have a weight proportional to its divergence from the overall distribution. Finally, λ_1 and λ_2 are hyperparameters that balance the contributions of KL divergence and L2 regularizer, which are determined by grid search.

In the three-class setting, we extend the loss function in (2) to accommodate the two required binary tasks. Specifically

$$\begin{aligned}\mathcal{L}_3 = \textstyle\sum_{i,j}(\mathcal{L}_{\text{CE}}(y_{ij}^{(1)}, \hat{p}_{ij}^{(1)}) + \mathcal{L}_{\text{CE}}(y_{ij}^{(2)}, \hat{p}_{ij}^{(2)}) \\ + \lambda_1 \textstyle\sum_k \alpha_k (\text{KL}(P_k^{(1)} \| Q_k^{(1)}) + \text{KL}(P_k^{(2)} \| Q_k^{(2)})) \\ + \lambda_2 \textstyle\sum_k |v_k\|_2^2 \,, \qquad (4)\end{aligned}$$

where P_k^1 and P_k^2 denote the distributions formed with the ground truth ratings within the value cluster $C_k \in \mathcal{C}$ for prediction tasks that produce $y_{ij}^{(1)}$ and $y_{ij}^{(2)}$, respectively. Similarly, Q_k^1 and Q_k^2 represent the distributions of probabilistic predictions on the corresponding tasks obtained using (3).

2.4 Extension to Preference Learning

We further extend the MC-STL framework to the preference learning setting, where the model learns from what a human prefers from a given pair of text items. This approach is commonly used to train reward models for reinforcement learning from human feedback (Christiano et al., 2017). In this case, the dataset is given by $\mathcal{D} = \{(x_i, x_j, y_{ij})\}$, where $y_{ij} \in \{0, 1\}$ indicates whether an annotator prefers text item x_i over x_j ($y_{ij} = 1$ if x_i is preferred and $y_{ij} = 0$ otherwise). Each preference is assumed to be driven by an underlying set of value perspectives $S_{ij} = \{\, C_k \mid y_{ij} \in C_k \,\}$.

The objective is to learn a scalar scoring function $r_\theta(\cdot)$ such that the preferred text item receives a higher score than the rejected one when conditioned on its associated value perspective(s) S_{ij}. Specifically, a forward pass over a single text item x_i and its value cluster assignment S_{ij} produces a scalar score $r_\theta(x_i, S_{ij})$. Although the backbone architecture of MC-STL remains the same here, *i.e.*, $r_\theta(x_i, S_{ij}) = o(f(x_i, y_{ij}))$, we introduce a learnable temperature parameter t_k for each value cluster C_k. This temperature term regulates the scale of the output logits, preventing the reward scores from becoming too extreme. Given a preference annotation comparing text items x_i and x_j with value perspectives S_{ij}, the probability that x_i is preferred over x_j is modeled as

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

and the corresponding loss function is

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

where $\mathcal{L}_{\text{BT}}(y_{ij}, \hat{p}_{ij})$ is the Bradley-Terry loss for a preference pair (Christiano et al., 2017) that is consistent with the cross entropy loss in our setup, *i.e.*, $\mathcal{L}_{\text{BT}} = -\big[y_{ij} \log \hat{p}_{ij} + (1 - y_{ij}) \log(1 - \hat{p}_{ij})\big]$.

3 Experiments

Our experiments were performed in server with a single NVIDIA RTX A6000 GPU. For the free-text rater rationale clustering, we used the K-Means clustering algorithm, although other clustering methods are also supported by the MC-STL framework. We used the GPT-4o API to summarize rater rationale clusters, and to obtain meaningful summaries of the value concepts addressed in each cluster (see Appendix A.3 for the prompt used). We performed clustering for multiple values of $K \in \{2, \ldots, 6\}$ and an expert psychologist[1] determined the best K by examining the uniqueness of the resulting clusters. All text encodings (except for preference modeling) were produced using a pretrained sentence-BERT (all-MiniLML6-v2) Reimers and Gurevych (2019). For preference modeling, the text encodings were generated by longformer-base-4096 (Beltagy et al., 2020) fine-tuned on training dataset for 1 epoch, to accommodate the context window needed for multi-turn conversation. The final hidden-state embedding of the [CLS] token was used to represent each text item. Evaluation was done on a test set consisting of 20% of the combined dataset for every dataset, which is created by stratified sampling to preserve the label distribution of D_{train}. Additional experiment details are given in Appendix A.2.

3.1 Datasets

ValuePrism (Sorensen et al., 2024a) is a large-scale dataset consisting of roughly 31k subjective situations (text items) paired with 218k annotations. Each annotation assigns one of three ordinal labels: *Oppose*, *Either*, or *Support*, and is accompanied by a free-text rationale explaining the annotator's judgment. These rationales are categorized into

[1] Expert annotation details are omitted for anonymity.

three types that reflect different normative frames: *Human Rights*, *Human Values*, and *Human Duties*.

We benchmark MC-STL separately on the three corresponding subsets: **VP-Rights** (49k annotations), **VP-Values** (97k annotations) and **VP-Duties** (71k annotations). Evaluating these subsets independently allows us to study value clustering within distinct rationale categories, which differ in linguistic structure and framing. For example, *Human Rights* rationales frequently begin with expressions such as "Right to . . . ". Separating categories prevents clustering methods from trivially grouping instances by rationale type rather than by the finer-grained semantic differences in value topics.

ValuePrism + Schwartz Value Labels Schwartz's theory of basic human values (Schwartz, 2012) defines ten universal value categories: *Achievement*, *Benevolence*, *Conformity*, *Hedonism*, *Power*, *Security*, *Self-Direction*, *Stimulation*, *Tradition*, and *Universalism*. Using the ValuePrism dataset, we map each instance, namely the subjective situation, the annotator's free-text value rationale, and the explanation of the label, to one or more Schwartz value categories. This mapping is performed by prompting an LLM (GPT-4o) to assign Schwartz value labels based on free-text rationales, as discussed in Appendix A.4. Although prior work has shown that LLMs can map complex social situations to Schwartz values (Yao et al., 2024; Senthilkumar et al., 2024), our setting involves a simpler task: mapping explicitly stated value rationales to fixed value categories. To validate the quality of these mappings, 100 randomly selected instances per value category were evaluated by experts with training in psychology, achieving an agreement of 93%, as described in Appendix A.5.

Anthropic HH-RLHF is a dataset (Bai et al., 2022) is a widely used preference-learning corpus consisting of roughly 169k pairs of human-preferred and rejected model-generated responses. Each pair indicates which response is preferred by a human rater. We use the **ValueImprint** annotations introduced by Obi et al. (2024), which consist of seven expert-defined high-level value categories: *Information Seeking*, *Wisdom & Knowledge*, *Well-being & Peace*, *Justice & Rights*, *Duty & Accountability*, *Civility & Tolerance*, and *Empathy & Helpfulness*.

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.

D3CODE (Davani et al., 2024b) contains 4.5k social media posts annotated for offensiveness by 4,309 participants from 8 geo-cultural regions. The posts were rated on a 5-point Likert scale and later binarized by authors (Davani et al., 2024a).

3.2 Evaluation Metrics

Overall & Value group-level Performance The overall performance indicates the quality of all probabilistic predictions made by the model. In binary and ordinal label setting, we use the *macro-AUC* score. For the preference learning setting, the ability of a model to score a preferred text item higher than a rejected item is important, thus we use pairwise accuracy to measure instance level performance.

In value group level performance, we use the same metrics, but quantify them with respect to each value cluster and summarize them as mean with standard deviation.

Value Calibration This metric seeks to measure model calibration w.r.t. each value group. In binary label settings where $y_{ij} \in [0, 1]$, for each value group $C_k \in \mathcal{C}$, we bin the probabilistic predictions for $P(y_{ij} = 1)_{ij \in C_k}$ by deciles. For all bins, we show the average of predictions in each bin (x axis) against the average of ground truth label for instances belonging to each bin (y axis). Then we estimate the calibration for a value group by fitting a line to these points. The ideal calibration must have unit slope and zero bias.

For the ordinal label scenario (three classes), we extend this by plotting a calibration plot for two binary tasks $P(y_{ij}^{(1)} = 1)_{ij \in C_k}$ and $P(y_{ij}^{(2)} = 1)_{ij \in C_k}$ for different value groups. For the preference label scenario, we estimate the calibration of $P(y_{ij} = 1 | x_i, x_j, S_{ij})_{ij \in C_k}$ in the same manner. This is a common practice in the preference learning setup (Shen et al., 2024). In all cases, we record the mean and standard deviations of the calibration slope and bias over all value clusters (and also over both binary tasks in the ordinal setting).

Item-level minority predictability On text items with multiple, contrasting annotations, it is important for a subjective NLP model to be able to account for minority opinions. To measure this in the ordinal setting, in which one should penalize a wrong prediction the more it is farther from the correct label, we use the normalized Earth Mover Distance (EMD) (Castaño et al., 2022). EMD computes the probability mass that must be shifted to convert a predicted distribution to its label distri-

bution. This makes moving probability mass from ordinal classes near to true label cheaper than from those classes farther away from true label. In 1D, EMD has a simple closed form where it is equal to L_1 norm of the two cumulative distributions (Vallender, 1974), and it ranges from 0 to $L-1$ where L is the number of classes. In the binary case, EMD reduces to the absolute difference between the probability and label, since there is only one boundary separating both classes. We obtain EMD for predictions to each non-majority label $y_i \in \mathbf{Y}$ and record the $1 - (\text{normalized EMD})$ to keep metric in [0, 1] across setups and higher values indicating better performance. Further details are explained in Appendix A.6.

3.3 Baseline Methods

We mainly compare our work against two conventional baselines where latent values behind annotations are either ignored or are considered as noise. **∅**: No information about value clusters $\{S_{ij}\}$. In this case, $P(\mathbf{Y} \mid \mathbf{X}, \varnothing(S_{ij})) = P(\mathbf{Y} \mid \mathbf{X})$. This is the case described in Figure 1(right) where hidden values behind annotations are ignored in modeling. **Maj**$(\mathbf{Y})$: Majority filtering of annotator disagreements. For each $\mathbf{y}_i \in \mathbf{Y}$ in D_{train}, we remove all annotations y_{ij} that are not majority in $\mathbf{y}_i$ because they are considered noise. Consequently, we model $P(\mathbf{Maj}(\mathbf{y_i}) \mid \mathbf{x_i})$. This is the case described in Figure 1(left) where annotations undergo majority voting, with hidden values explicitly ignored.

Both baselines are trained using the cross-entropy loss and the same architecture of MC-STL, excluding value group specific learnable embedding. Note that MC-STL model also do not consider any additional input besides x_i, but it learns independent embeddings for each value group obtained through clustering. The closest work we could find related to sociocultural profiling that supports the three-class setting is Fleisig et al. (2023). We compare to their approach in Appendix A.12.

4 Results

Rater Rationale Clustering We clustered annotations in D_{train} using K-means over sentence embeddings of rater rationales for different values of K and the experts preferred $K=5$, $K=4$, and $K=3$ clusters for VP-Value, VP-Right and VP-Duty datasets, respectively, after reviewing summarized value perspective of each cluster generated by language model, as explained in Appendix A.8.

Table 1: Results on rater rationale clustering. AUCs, Slope/Bias and 1-EMD indicate discrimination, calibration and minority predictability, respectively. Figures are averages with standard deviations.

Method	**Overall** (AUC)	**Group** (AUC)	**Calibration (Slope, Bias)**	**Item** (1-EMD)
VP-Duty (Binary)				
MC-STL	**0.76**	$\mathbf{0.74}_{0.02}$	$\mathbf{0.99}_{0.04}, \mathbf{0.00}_{0.03}$	**0.58**
ϕ	0.71	$0.69_{0.02}$	$0.79_{0.16}, 0.15_{0.19}$	0.51
Maj(Y)	0.69	$0.68_{0.03}$	$0.56_{0.17}, 0.33_{0.18}$	0.37
VP-Right (Ordinal)				
MC-STL	**0.75**	$\mathbf{0.73}_{0.03}$	$\mathbf{0.96}_{0.08}, \mathbf{0.00}_{0.06}$	**0.64**
ϕ	0.67	$0.67_{0.02}$	$0.81_{0.08}, 0.06_{0.07}$	0.55
Maj(Y)	0.68	$0.67_{0.01}$	$0.63_{0.07}, 0.13_{0.11}$	0.42

Value cluster summaries for VP-Duty are C_1: *A care-centered, duty-based ethic grounded in compassion, solidarity, and stewardship*; C_2: *A protective, law-abiding, harm-averse ethic prioritizing life, rights and justice*; and C_3: *An ethic centered on human dignity, autonomy, truth, fairness, and responsibility*; and for the other two (VP-Rights and VP-Values) in Appendix A.7. At test time, embedding of the rater rationale for an instance is mapped to the closest cluster w.r.t. their centroids and is evaluated based on the mapped cluster perspective.

The Results in Table 1 highlight the superiority of the MC-STL framework for rationale clustered value groups over all metrics on both datasets. The highest AUC scores of MC-STL indicate the ability of annotation-rationale grouping to model latent subjectivity. The higher (1-EMD) scores of 0.58 and 0.64 on VP-Duty and VP-Rights respectively showcase improvement in minority vote predictability. The comparatively poor performance of Maj(y) in calibration and minority predictability substantiates how latent human values can be suppressed by filtering annotator disagreements through methods such as majority voting. MC-STL also achieve near ideal calibration scores across rationale clusters with minimal deviation as shown in Figure 3(row 1). We also present performance metrics for VP-Duty(Ordinal), VP-Rights(Binary), VP-Values (Ordinal and Binary) in Appendix A.13, all of which are consistent with the results here.

Expert Taxonomy Clustering Performance in expert taxonomy clustering also depends on the ability of taxonomies to represent the subjectivity in the dataset. The change in AUC scores of MC-STL (Table 2) on the VP+Schwartz dataset relative to ϕ by 0.07 reveal that segregating the annotations by the Schwartz Universal value taxonomy provides a useful signal to model the latent subjectivity. In the

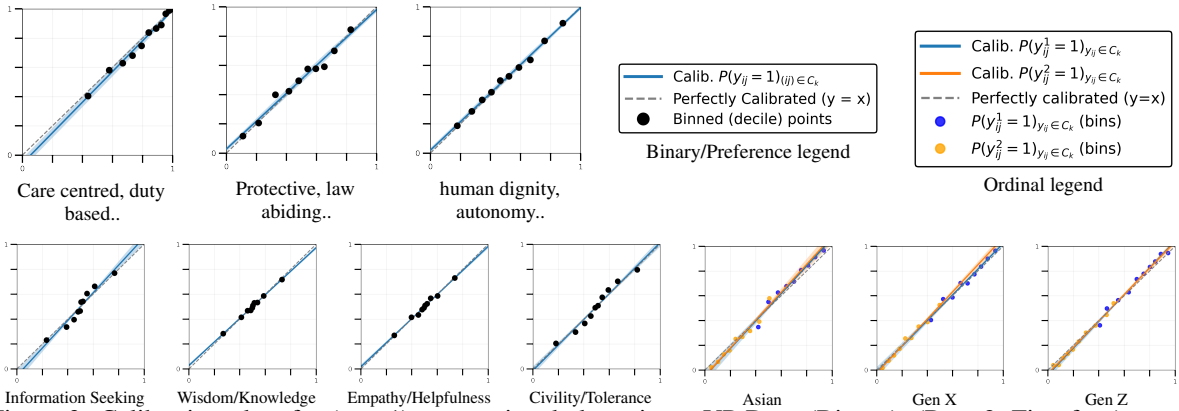

Figure 3: Calibration plots for (row 1) rater rational clustering – VP Duty (Binary); (Row 2, First four) expert taxonomy clustering – Anthropic+VI (Preference); (Row 2, Last three) Sociocultural Clustering - DICES 990 and rest of the calibration plots of MC-STL are displayed in Appendix A.9 and that of $\varnothing$ are shown in Appendix A.10

Table 2: Results on expert taxonomy clustering. The preference dataset uses pairwise accuracy not AUC.

Method	**Overall** (AUC/Acc)	**Group** (AUC/Acc)	**Calibration (Slope, Bias)**	**Item** (1-EMD)
VP+Schwartz (Binary)				
MC-STL	**0.80**	**0.76**$_{0.07}$	**0.98**$_{0.10}$, **0.02**$_{0.06}$	**0.65**
ϕ	0.73	0.69$_{0.08}$	0.53$_{0.30}$, 0.31$_{0.30}$	0.50
Maj(Y)	0.72	0.68$_{0.06}$	0.40$_{0.19}$, 0.42$_{0.24}$	0.49
Anthropic HH-RLHF+VI (Preference)				
MC-STL	**0.64**	**0.64**$_{0.05}$	**1.02**$_{0.05}$, **-0.01**$_{0.02}$	**0.61**
ϕ	0.60	0.59$_{0.04}$	0.94$_{0.55}$, 0.03$_{0.25}$	0.58
Maj(Y)	0.59	0.56$_{0.05}$	0.94$_{0.68}$, 0.02$_{0.34}$	0.58

Table 3: Results on Sociocultural clustering.

Method	**Overall** (AUC)	**Group** (AUC)	**Calibration (Slope, Bias)**	**Item** (1-EMD)
DICES-990 (Ordinal)				
MC-STL	**0.75**	**0.74**$_{0.01}$	**1.03**$_{0.03}$, **-0.03**$_{0.01}$	**0.44**
ϕ	0.68	0.69$_{0.01}$	0.95$_{0.07}$, 0.03$_{0.04}$	0.34
Maj(Y)	0.66	0.66$_{0.02}$	0.74$_{0.07}$, 0.09$_{0.09}$	0.22
D3 (Binary)				
MC-STL	**0.63**	**0.62**$_{0.02}$	**1.00**$_{0.15}$, **0.06**$_{0.04}$	0.26
ϕ	0.61	0.60$_{0.01}$	0.72$_{0.14}$, 0.05$_{0.04}$	**0.31**
Maj(Y)	0.59	0.58$_{0.01}$	0.44$_{0.12}$, 0.21$_{0.03}$	0.15

case of the pairwise preference dataset Anthropic HH-RLHF, the modeling w.r.t. value taxonomies proposed by Obi et al. (2024) also improves pairwise accuracy scores as shown in Table 2. Ideal calibration scores of MC-STL on both datasets indicate the alignment over each value taxonomy classes, as presented in Figure 3(row 2). Although MC-STL improves item level minority predictability in both cases, the modest difference between $\varnothing$ and Maj(Y) in Anthropic HH-RLHF (both having a score of 0.58) can be attributed to the relatively small number of paired items with multiple contrasting annotations (74 instances) in the dataset.

Sociocultural Clustering In Table 3, MC-STL outperforms the baselines in overall and group level AUC scores along with improved item-level minority predictability on DICES-990. While prior works have modeled subjectivity by annotator demographics, we focus on calibrating model to each distinct sociocultural group, so as to align model with perspectives from every unique group. MC-STL preserves equitable calibration across identity group with minimal deviation on the DICES-990 Dataset (Figure 3). On the D3 dataset, we observe comparatively lower performance by all methods that can be attributed to limited annotator descriptors in D3 (only gender, age, and region) that do not include rater ethnicity or rater educational backgrounds, as in DICES-990. While MC-STL attempts to calibrate available limited descriptors, it averages over distinct behaviors within these high dimensional descriptors causing it to have lower $1 - \text{EMD}$ relative to ϕ on the D3 Dataset, as shown in the figure comparing aggregated calibration *vs.* $1 - \text{EMD}$ over epochs in Appendix A.11. Still, if the objective is merely to calibrate against available descriptors, MC-STL is a good choice as it achieves close to ideal calibration on D3.

5 Conclusion

Human values underlying annotations are often ignored while building NLP systems. MC-STL constitutes a simple approach for modeling such value diversity by grouping annotations using three representations and calibrating model predictions to each resulting value group through cluster-specific representations, achieving pluralistic alignment of model outputs across diverse value settings. Results show ability of MC-STL to learn subjective labels while preserving underlying value pluralism.

6 Limitations

While we have presented three possible representations to group annotations, there could be other potential representations inspired from social science theory to extract latent value clusters from annotations. Moreover, while our work focused on treating each of three representations as alternatives, the effects of extracting latent clusters at the intersection of them have not been studied. Also, previous works have shown the language model's ability to extract annotation rationales from rating demonstrations, which would help us to group annotations without rationales collected at annotation time. We have not attempted to understand the effects when language models are deployed to generate such rationales. Our work focuses on multicalibration as a group-fairness measure and have not studied fairness measures such as error rates across groups. There is a growing need to extend universal value alignment methods to further subjective NLP task setups such as text generation, which have not been studied in this work.

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.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, and 1 others. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Mohammad Atari, Mona Xue, Peter Park, Damián Blasi, and Joseph Henrich. 2023. Which humans?

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.

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 Diaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024a. D3code: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024b. 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.

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. *arXiv preprint arXiv:2305.06626*.

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.

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.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. *arXiv preprint arXiv:2010.15149*.

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.

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. *arXiv preprint arXiv:2110.05699*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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.

Rithik Appachi Senthilkumar, Amir Homayounirad, and Luciano Cavalcante Siebert. 2024. Leveraging large language models to identify the values behind arguments. In *International Workshop on Value Engineering in AI*, pages 87–103. Springer.

Judy Hanwen Shen, Archit Sharma, and Jun Qin. 2024. Towards data-centric rlhf: Simple metrics for preference dataset comparison. *arXiv preprint arXiv:2409.09603*.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2023. Distributional preference learning: Understanding and accounting for hidden context in rlhf. *arXiv preprint arXiv:2312.08358*.

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 Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, and 1 others. 2024b. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

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.

SS Vallender. 1974. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786.

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., 2023; 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. 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. However, these methods typically require substantial annotation per item from demographically diverse annotators, making them costly in practice.

Value Modelling in Subjective Task: While value or human preference alignment of LLMs is widely recognized as crucial for its deployment (Kenton et al., 2021; Wang et al., 2024), how to

define such values remains an open question. The commonly used principle of *helpful, honest and harmless* "HHH" (Askell et al., 2021), in its adoption is identified to be ambiguous (Yao et al., 2024). Moreover, limited works have explored value modelling beyond text generation setup. According to (Schwartz, 2012), human values encode desirable goals guiding behavior or action. 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 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 across distinct human value groups. We form these value groups over annotations in a subjective dataset through three possible representations and let the model learn value-group-specific embedding to model subjectivity for instances belonging to each group explicitly. Additionally, we model the ordinal structure of intermediate labels (e.g., "ambiguous" that lies between opinionated extremes), a common yet under-explored feature of subjective tasks where annotators choose neutral option between binary extremes.

A.2 Additional Experiment Details

The source code, annotated dataset, and data splits will be released upon acceptance. In loss functions Eq. 2 and Eq.4, the best performing hyperparameters after grid search were found to be $\lambda_1 = \frac{1}{7.4*m_i}$ for binary and $\frac{1}{2*7.4*m_i}$ for ordinal labels where m_i denotes number of annotations for text item x_i in the dataset.

A.3 Prompt to generate cluster 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 cluster with its collection of rater rationale sentences.

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

Figure 4: Prompt to generate cluster summary.

A.4 Prompt to generate Schwartz value annotation from value rationales

Free text rationales for each annotation in ValuePrism Dataset (Sorensen et al., 2024a) was provided as "input_text" and associated explanation statement was provided as "explanation". Along with value class definitions, we also provided keywords associated with each class in (Schwartz, 2012).

```
SYSTEM_PROMPT = """
You are an expert annotator that maps
    short input human value text
    descriptions to Schwartz 10 value
    classes. Use the 10-class Schwartz
    taxonomy of 10 classes as listed:

["Power","Achievement","Hedonism","
    Stimulation","Self-direction","
    Universalism","Benevolence","
    Tradition","Conformity","Security"]


Schwartz 10 value classes definitions:

Power: social status and prestige,
    control or dominance over people and
     resources-authority; wealth,
    authority, recognition

Achievement: personal success through
    demonstrating competence according
    to social standards-success;
    influential, successful, ambitious,
    capable, intelligent

Hedonism: pleasure or sensuous
    gratification for oneself-pleasure;
    pleasure, enjoy, indulgent

Stimulation: excitement, novelty and
    challenge in life-daring activities;
     daring, variation, excitement

Self-direction: independent thought and
    action choosing, creating, and
    exploring-creativity; freedom,
    curious, independent, goal, privacy,
     respect

Universalism: understanding,
    appreciation, tolerance, and
    protection for the welfare of all
    people and for nature-
    broadmindedness; broadminded,
```

```
    equality, unity, protection, harmony
    , justice, wisdom, beauty

Benevolence: preserving and enhancing
    the welfare of those with whom one
    is in frequent personal contact-
    helpfulness; love, spiritual,
    helpful, friendship, forgiving,
    responsible, loyal

Tradition: respect, commitment, and
    acceptance of the customs and ideas
    that one's culture or religion
    provides-accepting one's portion in
    life; humble, respect, devout,
    moderate

Conformity: restraint of actions,
    inclinations, and impulses likely to
     upset or harm others and violate
    social expectations or norms- self-
    discipline; discipline, politeness,
    obedient

Security: safety, harmony, and stability
     of society, of relationships, and
    of self-cleanliness; family security
    ; healthy, family, order, clean,
    safety, belonging

Hard rules (must follow exactly):

1. Determine which Schwartz value(s)
    apply to the input human value text
    description.

   - If exactly only one value class
       clearly applies, return only that
        single label in a list.

   - If more than one value class
       clearly applies, return all such
       labels as a comma-separated list.

2. Return only a single python list
    object per input (no additional text
     or commentary).

3. The python list must match this
    schema exactly:
["<label1>","<label2>",...]

4. Output constraints:
- No extra commentary, no trailing
    commas, valid list only.

Input human value text description:  "{
    input_text}"

Explanation of the human value context:
    "{explanation}"
"""
```

A.5 Expert evaluation of VP+Schwartz

After mapping each annotation in ValuePrism dataset to a group of Schwartz values by language model, we randomly selected 100 instances mapped to each of the 10 value classes (Total 1000 instances). 10 Expert Psychologists validated mappings of distinct 100 instances with 'Correct/Wrong' labels. Overall, 93% of instances were found to be 'Correct' by the experts.

A.6 1-D EMD Calculation for Discrete Distribution

For a discrete one-dimensional class distribution, the Earth Mover's Distance (EMD) measures the minimum amount of "work" required to transform a predicted probability distribution into the ground-truth label distribution. In the 1-D case, EMD admits a closed-form solution and can be expressed as the L_1 norm of the difference between their cumulative distribution functions (CDFs).

Let p_{ij} denote the predicted probability distribution over L ordered classes for sample ij, and let y_{ij} denote the corresponding ground-truth label distribution, represented as a one-hot vector. The cumulative distributions are defined as

$$P_{ijl} = \sum_{l=1}^{L} p_{ijl}, \quad Y_{ijl} = \sum_{l=1}^{L} y_{il}, \quad l = 1, \ldots, L.$$

The 1-D Earth Mover's Distance is then given by

$$\mathrm{EMD}(p_i, y_i) = \sum_{k=1}^{K} |P_{ik} - Y_{ik}| .$$

Thus, the 1-D EMD is equivalent to the L_1 distance between the cumulative distributions of the predicted and true labels.

Consider an ordinal classification problem with three ordered classes $\{1, 2, 3\}$. Let the ground-truth class be class 2, so the one-hot label distribution is

$$y_{ij} = [0,\ 1,\ 0].$$

Assume the model predicts the probability distribution

$$p_{ij} = [0.2,\ 0.5,\ 0.3].$$

The cumulative distributions are

$$P_{ij} = [0.2,\ 0.7,\ 1.0], \quad Y_{ij} = [0,\ 1,\ 1].$$

The EMD is computed as

$$\mathrm{EMD}(p_{ij}, y_{ij}) = |0.2-0|+|0.7-1|+|1.0-1| = 0.5.$$

For a binary/ preference classification problem with classes $\in \{0, 1\}$, suppose the ground-truth label is class 1, yielding

$$y_{ij} = [0,\ 1].$$

Let the predicted distribution be

$$p_{ij} = [0.7,\ 0.3].$$

The cumulative distributions are

$$P_{ij} = [0.7,\ 1.0], \quad Y_{ij} = [0,\ 1].$$

The EMD is therefore

$$\mathrm{EMD}(p_i, y_i) = |0.7 - 0| + |1.0 - 1| = 0.7.$$

EMD ranges from $[0, L-1]$. Therefore to keep EMD in same range across ordinal and binary/preference setting, we normalize EMD by $L-1$. For all instances $y_{ij} \in y_i$ such that y_{ij} is different from the majority label in y_i, we measure the *EMD* between their predicted distribution p_{ij} and the corresponding label distribution by calculating the L_1 norm of their cumulative distributions.

We obtain and record the $1 - (normalized\ \mathrm{EMD})$ by

$$1 - \frac{\sum_i \sum_{j:\, y_{ij} \neq \mathrm{Maj}(y_i)} \mathrm{EMD}(y_{ij}, p_{ij})}{(L-1) * (\sum_i \sum_{j: y_{ij} \neq \mathrm{Maj}(y_i)} 1)},$$

to keep the metric in the $[0, 1]$ range with higher values indicating better performance.

A.7 Value Cluster Summaries

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 cluster summaries.

C_k	**Summary**
1	The corpus heavily prioritizes autonomy and freedoms, balanced by strong commitments to safety/security, lawful order, property rights, economic performance, and democratic norms.
2	The list is overwhelmingly centered on well-being, happiness, and health, supported by honesty/integrity, altruism, and personal growth, with strong attention to relationships/-family, emotional health, and balancing work.
3	Centered on compassion-driven ethics, extending strongly to animals and people, prioritizing loving relationships, social cohesion, and altruistic generosity.
4	The list overwhelmingly centers on justice, respect, responsibility, equality, fairness, and the rule of law.
5	The text overwhelmingly emphasizes valuing and protecting life in all forms, coupled with stewardship of the environment, respect for human dignity, and preservation of cultural and religious heritage

Table 5: VP-Values value cluster summaries.

A.8 Expert evaluation of Value Summaries to determine K in Rater Rationale Clustering

In order to determine K, we ran $K-$Means clustering from $K = 2$. For VP-Duty dataset, $K = 4$ gave below clusters:

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 6: VP-Rights value clusters when $K = 4$.

Expert Psychologist analyzed that compared to $K = 3$ (given in main script), $K = 4$ leads to repitition of harm averse ethic and care centred ethic across clusters summaries. To keep cluster definitions as distint as possible, we stopped $K = 3$ here. Similarly, for VP-Rights, $K = 5$ and for for VP-Values, $K = 6$ led to repitition of similar value based ethics across cluster summaries.

A.9 Calibration plots of MC-STL

Please refer figure 5.

A.10 Calibration Plots of ∅

Please refer figures 6 and 7.

A.11 Aggregate Calibration vs. (1-EMD) on D3 dataset

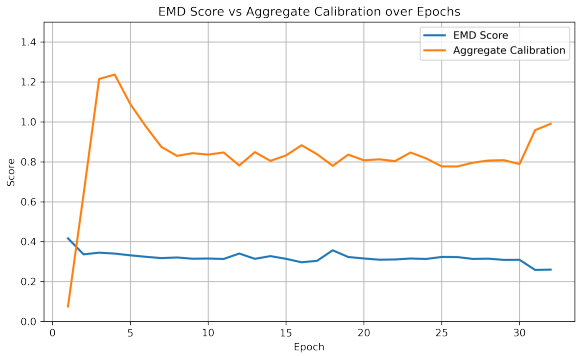

Figure 8: Aggregate Calibration vs (1-EMD) over epochs of MC-STL on D3

It is visible in the plot that as MC-STL gets calibrated to limited number of annotator descritpors in D3, (1-EMD) score gradually descreases. The limited number of descriptors cause their calibration to aggregate minority opinions within them, thereby decreading (1-EMD) score. For instance, the rater group "Asian" can represent multiple contrasting cultures, and calibrating to this group could wash away differences within the group in the absence of further differentiating attributes.

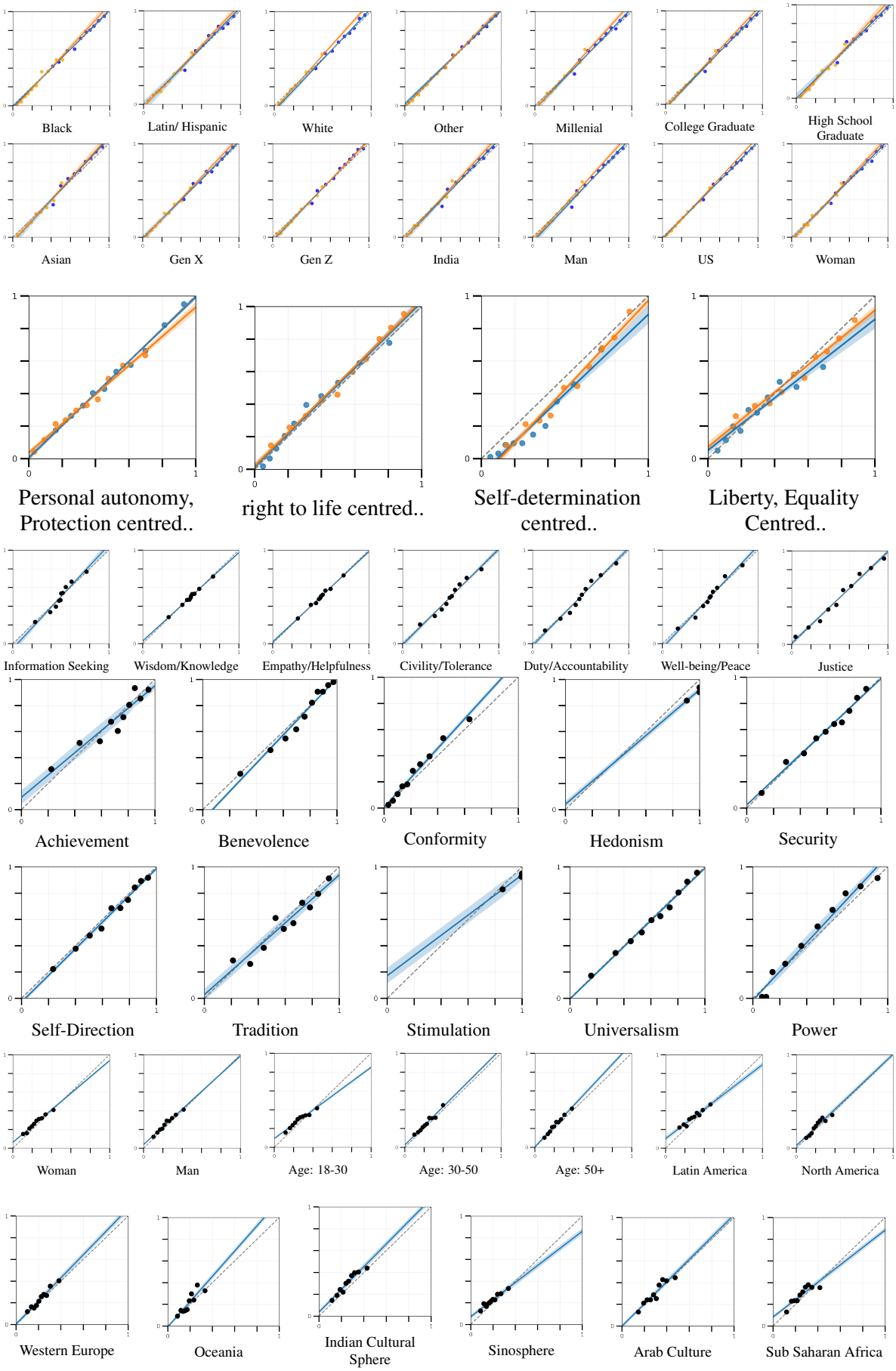

Figure 5: Calibration plots of MC-STL on (Row 1 & 2):DICES-990 (Ordinal), (Row 3): VP-Right (Ordinal), (Row 4): Anthropic (Preference), (Row 5 & 6): VP+Schwartz (Binary), (Row 7 & 8): D3 (Binary)

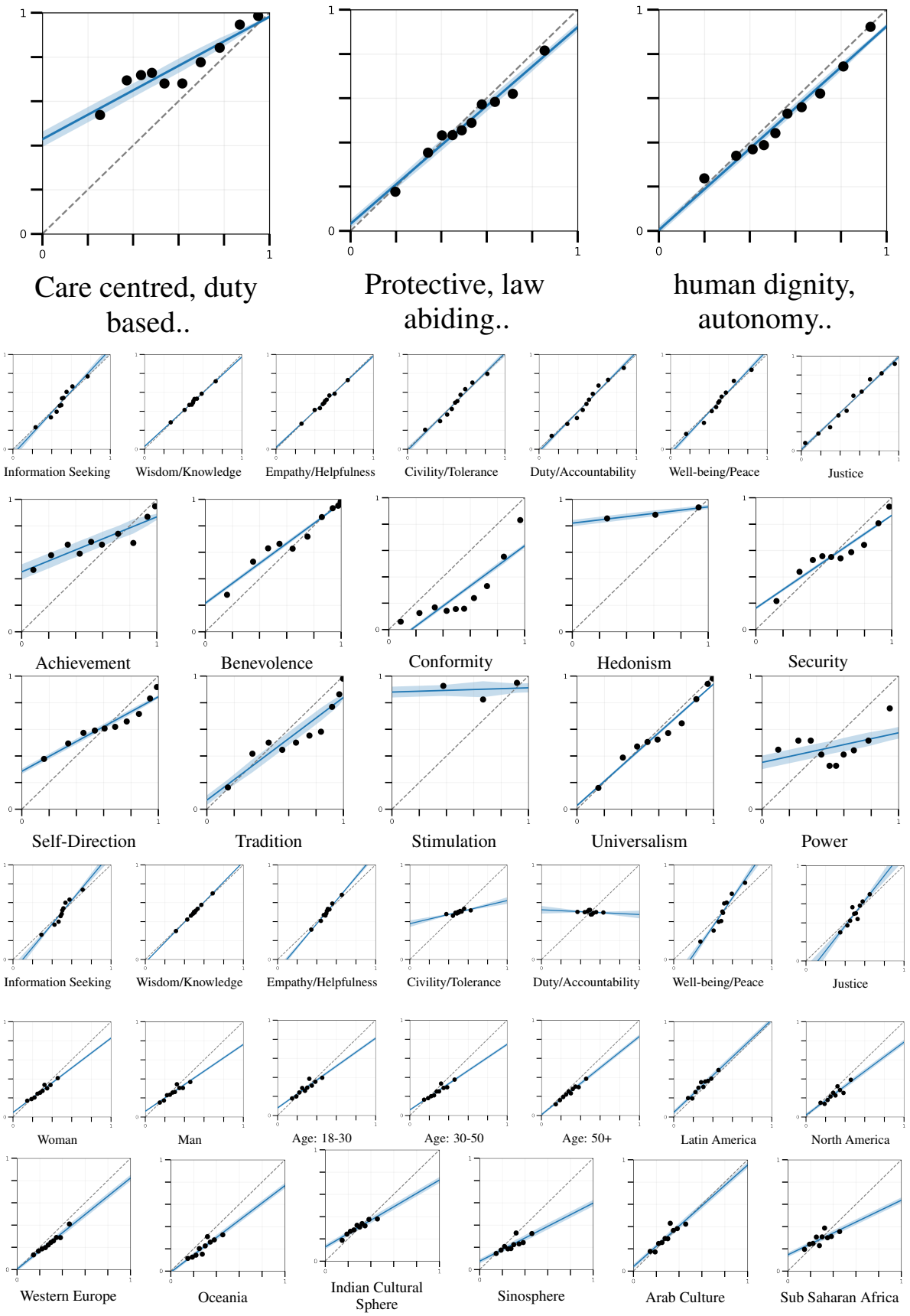

Figure 6: Calibration plots of ∅ on Binary- (Row 1):VP-Duty, (Row 2 & 3): VP+Schwartz, (Row 4): Anthropic HH-RLHF(Preference), (Row 5 & 6): D3

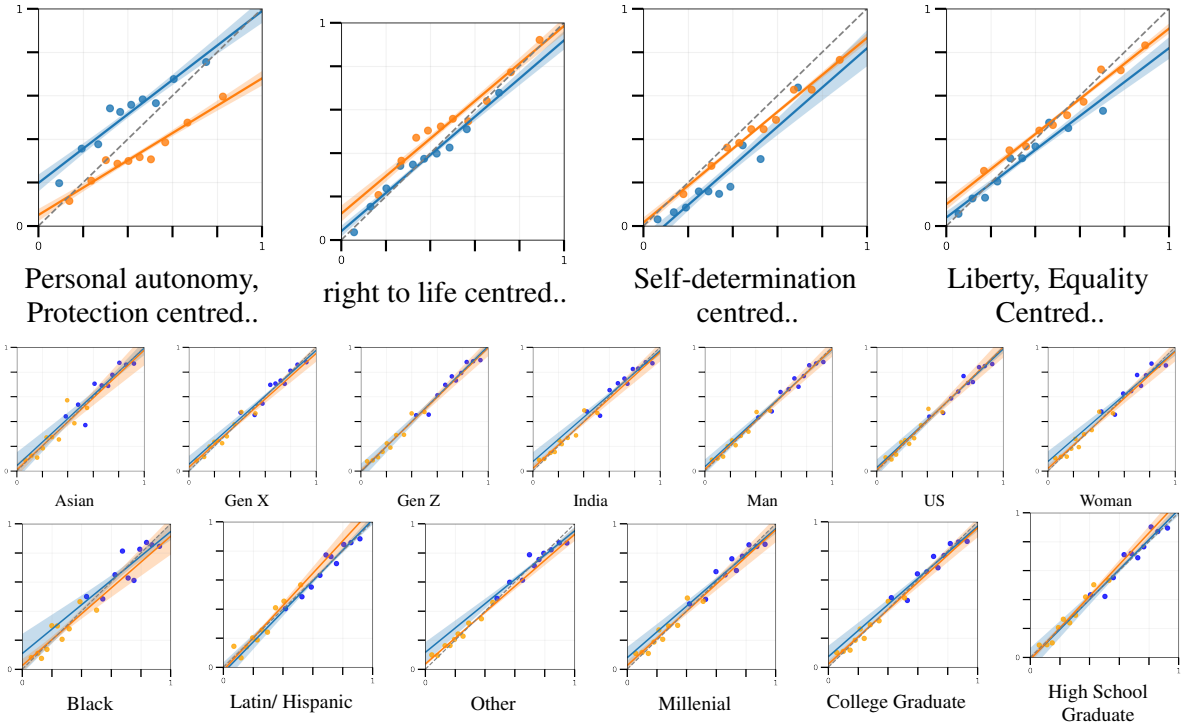

Figure 7: Calibration plots of ∅ on Ordinal- (Row 1):VP-Rights, (Row 2 & 3): DICES-990

A.12 Comparing MC-STL against IRPM

The Individual Rating prediction module in Fleisig et al. (2023) predicts annotator-specific ratings by conditioning on both sociocultural descriptors of the annotator and the target text instance. These inputs are encoded using a pretrained RoBERTa model (Liu et al., 2019). Annotator descriptors are combined with the text via a template-based input format, "[t_j] [SEP] x_i," where t_j denotes the annotator profile and x_i the text item. The model is trained using cross-entropy loss for both binary and ordinal prediction settings.

Table 7: Results on IRPM.

Method	**Overall** (AUC)	**Group** (AUC)	**Calibration** **(Slope, Bias)**	**Item** (1-EMD)
DICES-990 (Ordinal)				
MC-STL	**0.75**	$\mathbf{0.74}_{0.01}$	$\mathbf{1.03}_{0.03}$, $\mathbf{-0.03}_{0.01}$	**0.44**
IRPM	0.63	$0.65_{0.00}$	$2.47_{0.23}$, $-0.61_{0.09}$	0.39
D3 (Binary)				
MC-STL	**0.63**	$\mathbf{0.62}_{0.02}$	$\mathbf{1.00}_{0.15}$, $\mathbf{0.06}_{0.04}$	0.26
IRPM	0.60	$0.61_{0.01}$	$1.21_{0.21}$, $0.42_{0.05}$	**0.27**

A.13 Additional Results

Table 8: Rater Rationale Clustering

Method	**Overall** (AUC)	**Group** (AUC)	**Calibration** **(Slope, Bias)**	**Item** (1-EMD)
VP-Duty (Ordinal)				
MC-STL	**0.70**	$\mathbf{0.69}_{0.02}$	$\mathbf{0.99}_{0.06}$, $\mathbf{0.00}_{0.03}$	**0.59**
ϕ	0.65	$0.65_{0.00}$	$0.78_{0.13}$, $0.09_{0.11}$	0.55
Maj(Y)	0.64	$0.64_{0.02}$	$0.59_{0.15}$, $0.17_{0.06}$	0.43
VP-Right (Binary)				
MC-STL	**0.81**	$\mathbf{0.80}_{0.02}$	$\mathbf{0.97}_{0.02}$, $\mathbf{0.00}_{0.01}$	**0.64**
ϕ	0.73	$0.73_{0.03}$	$0.82_{0.05}$, $0.11_{0.05}$	0.50
Maj(Y)	0.72	$0.72_{0.04}$	$0.64_{0.07}$, $0.27_{0.08}$	0.35
VP-Value (Binary)				
MC-STL	**0.80**	$\mathbf{0.79}_{0.03}$	$\mathbf{0.96}_{0.03}$, $\mathbf{0.01}_{0.02}$	**0.67**
ϕ	0.70	$0.71_{0.05}$	$0.89_{0.19}$, $0.07_{0.19}$	0.57
Maj(Y)	0.69	$0.71_{0.05}$	$0.58_{0.14}$, $0.30_{0.15}$	0.51
VP-Value (Ordinal)				
MC-STL	**0.70**	$\mathbf{0.70}_{0.04}$	$\mathbf{0.96}_{0.09}$, $\mathbf{0.03}_{0.04}$	**0.65**
ϕ	0.63	$0.64_{0.05}$	$0.81_{0.22}$, $0.08_{0.13}$	0.58
Maj(Y)	0.63	$0.64_{0.04}$	$0.57_{0.15}$, $0.17_{0.13}$	0.54