

Annotators of a Feather FLOCKE Together: A Framework to Analyze Latent Group Patterns Based on Annotator Behavior

Giovanni Valer    Elisa Leonardelli  Camilla Casula 
Sara Tonelli  Michele Lombardi 

 Fondazione Bruno Kessler  University of Bologna  University of Pisa
{gvaler,eleonardelli,ccasula,satonelli}@fbk.eu, michele.lombardi2@unibo.it

Abstract

Accounting for disagreement among annotators is important for a variety of NLP applications. However, understanding what insights can be drawn from disaggregated annotations remains an open question. We introduce FLOCKE, a framework that jointly learns a task and models latent groups of annotators with similar behavior. We first formalize six analyses enabled by our framework, including annotator-group discovery, group- and individual-level label prediction, and finding the items most prone to cause disagreement. We then validate FLOCKE with a controlled synthetic benchmark, creating pre-defined annotator groups and labeling patterns, and showing that it can recover group structure with very high accuracy. We apply FLOCKE to two real-world datasets, where it predicts individual annotations, discovers interpretable annotator groups, and shows that groups are not explained by socio-demographic traits alone.

1 Introduction

The long-standing paradigm of a single “ground truth” in annotated data has been challenged in the last years by studies proposing human label variation (Plank, 2022) as a way to accommodate multiple plausible annotators’ views on the same linguistic item. In fact, not all disagreement stems from sloppy annotation, but may be due to multiple valid views that cannot be reconciled (Frenda et al., 2025) because they legitimately represent people with diverse values (Sorensen et al., 2024), especially when dealing with subjective annotation tasks. This new pluralistic paradigm poses challenges that cover the whole pipeline of machine learning systems, from the annotation of training data (Leonardelli et al., 2025) to the implementation of perspectivist systems (Romberg, 2022) and the development of novel metrics for evaluation (Rizzi et al., 2024; Lo et al., 2025).

Following initiatives such as the *Learning with Disagreement* (LeWiDi) shared tasks (Leonardelli

et al., 2023, 2025) and *Perspectivist Approaches to NLP* workshop series (Abercrombie et al., 2022, 2023, 2024), an increasing number of benchmarks, frameworks and datasets with disaggregated annotation has been made available to the NLP research community (Lo et al., 2025; Meister et al., 2025). However, understanding what insights can be drawn about annotator behavior and data characteristics from this kind of resources is still an open challenge.

We address this issue by proposing FLOCKE, a novel framework for the analysis of disaggregated datasets. Beyond predicting labels, FLOCKE enables several downstream analyses of annotation behavior, such as identifying latent groups of annotators that exhibit similar labeling patterns, estimating group-specific label distributions and detecting ambiguous or subjective examples. In doing so, FLOCKE provides a compact and interpretable intermediate solution between majority-vote aggregation and more fine-grained alternatives, such as soft-label distributions or perspectivist modeling.

This work proposes a novel validation paradigm based on synthetic annotations and groups derived from HateCheck (Röttger et al., 2021), a template-based hate speech test suite. This paradigm allows the assessment of group quality and disagreement patterns identified by FLOCKE under controlled conditions. Unlike evaluations based only on real-world annotation datasets, where latent group structure and disagreement-driving items are not directly observable, our setup provides reference points to evaluate model outputs. After validating FLOCKE in this controlled setup, we further apply it to real-world data and compare our insights with those obtained by existing methods.

FLOCKE is available on PyPI¹; the code for validation and experiments is released on GitHub².

¹  flocke

²  dhfbk/flocke-annotator-latent-groups

Method ↓	Capabilities						
	Latent groups discovery	Group number detection	Annotator group distribution	Group label prediction	Annotator label prediction	Generalize to unseen items	Most relevant items
FLOCKE (ours)	✓	✓	✓	✓	✓	✓	✓
Venanzi et al. (2014)	✓	✓	✗	✓*	✓	✗	✗
Fell et al. (2021)	✓	✗	✗	✗	✗	✗	✓*
Lo and Basile (2023)	✓	✓*	✗	✓	✗	✓	✗
Li et al. (2024)	✓	✓*	✓	✓	✓	✗	✗
Vitsakis et al. (2024)	✓	✓*	✗	✗	✓	✓*	✗
Sorensen et al. (2025)	✓	✗	✓	✓	✓	✓	✓*

Table 1: Comparison of methods that perform latent group discovery, across seven different tasks. FLOCKE, thanks to its joint probability modeling, integrates all capabilities. ✓* identifies capabilities that are only partially supported, as well as those that may be supported but were not considered by their authors.

2 Related Work

Work on annotation modeling can be broadly grouped into individual, group-level, and distributional approaches (Kocoń et al., 2021). Individual modeling predicts annotator-specific labels through multi-task architectures, annotator embeddings, or value-conditioned inference (Mostafazadeh Davani et al., 2022; Deng et al., 2023; Mokhberian et al., 2024; Homayounirad et al., 2025; Sorensen et al., 2025), though these methods are often limited by annotation sparsity (Mokhberian et al., 2024). Group-level approaches instead model shared biases or latent clusters, either using demographic information (Liu et al., 2022; Fleisig et al., 2023; Prabhakaran et al., 2024) or unsupervised perspective discovery (Venanzi et al., 2014; Lo and Basile, 2023; Li et al., 2024; Vitsakis et al., 2024; Sorensen et al., 2025). Distributional approaches aim to predict full soft-label distributions rather than single labels (Khurana et al., 2024; Benito-Santos and Ghajari, 2025), while other works focus on predicting which instances are likely to induce annotator disagreement (Rizzi et al., 2025). Lo et al. (2025) combine several of these ideas through user- and population-conditioned adaptation strategies. While a relatively small number of works perform latent group discovery, they differ from FLOCKE in joint-task modeling by adopting separate steps. Table 1 summarizes these differences, highlighting how FLOCKE integrates all considered tasks, and is capable of generalization to new annotators and items without recomputing steps, unlike existing pipeline-based approaches. Finally, our framework is validated with ground-truth latent groups, while previous works use internal validity measures or socio-demographics.

Few approaches move towards calibration and

steering. Khurana et al. (2024), Benito-Santos and Ghajari (2025), and Kurniawan et al. (2026) show that training on soft or disagreement-aware label distributions can improve calibration, accuracy, and fairness. Beyond calibration, Hwang et al. (2023) model users’ prior opinions to align LLM outputs with individual perspectives, and Li et al. (2024) introduce data-driven persona embeddings that steer LLMs toward specific viewpoints without predefined demographic categories. Recent work also shows that LLMs cannot reliably replace human annotators in subjective settings (Orlikowski et al., 2025; Ni et al., 2026). Overall, these methods focus on decoding control and calibration of LLMs, while FLOCKE performs annotation prediction.

Finally, research on annotator variation also spans analyses and taxonomies of disagreement (Plank, 2022; Sandri et al., 2023; Jiang and de Marneffe, 2022). FLOCKE differs from this body of work because it addresses a complementary question: it models behavioral patterns across annotations to discover latent annotator groups, rather than focusing only on the presence or source of disagreement.

3 Methodology

Our approach is based on modeling *latent group behaviors*, which can be jointly learned together with a task model, given a dataset with disaggregated annotations. While our method is general in principle, in this paper we will focus on binary classification as the target task. A schematic overview of our methodology is presented in Fig. 1.

3.1 Formulation

We consider a Machine Learning task that involves associating labels (e.g., numeric values or classes)

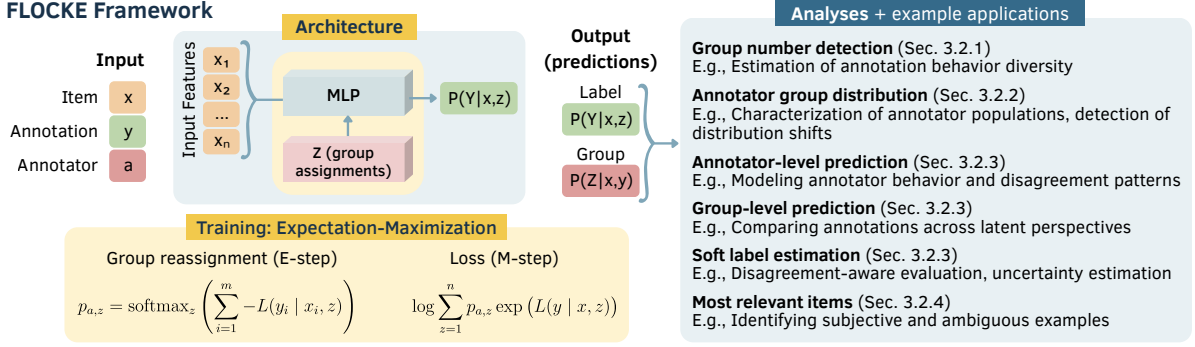


Figure 1: A schematic overview of FLOCKE with the six types of analysis it enables (L refers here to log-likelihood).

to contextual information (e.g., text, demographics, user profiles based on question-answering). In our empirical evaluation, the task consists of binary classification of text embeddings. We assume the training data to include identifiers for the annotators themselves. We use random variables X, A, Y to represent the contextual information, annotator ID, and label. The training data then consists of a sample $\mathcal{S} = \{x_i, a_i, y_i\}_{i=1}^m \sim P(X, A, Y)$, where x_i is the contextual information, a_i the annotator, and y_i the annotation for the i -th example.

We model group membership via a (latent) random variable Z with support $\{1..n\}$, where n is the number of groups. We also assume independence between groups and contextual information, i.e., no bias for having certain inputs annotated by specific groups. As discussed in Appendix A, this enables factoring the full distribution $P(X, A, Y, Z)$ as:

$$P(Y | X, Z)P(Z | A)P(X, A) \quad (1)$$

where $P(Y | X, Z)$ corresponds to the group behavior and $P(Z | A)$ to the group membership distribution for the annotator. By introducing an estimator $\hat{P}_\theta(Y | X, Z)$, with parameter vector θ , for the group behavior, we can define the sample likelihood by marginalizing out the Z variable:

$$L_S(\theta) = \prod_{i=1}^m \mathbb{E}_{z \sim \hat{P}_{\theta,S}(Z|a_i)} \left[\hat{P}_\theta(y_i | x_i, z) \right] \quad (2)$$

$$= \prod_{i=1}^m \sum_{z=1}^n \hat{P}_{\theta,S}(z | a_i) \hat{P}_\theta(y_i | x_i, z) \quad (3)$$

where $\hat{P}_{\theta,S}(z | a_i)$ represents here a *posterior* estimate for $P(Z | A)$, i.e., one performed by taking advantage of the available data. As discussed in Appendix A, we use the Bayesian estimate:

$$\hat{P}_{\theta,S}(z | a_i) = \frac{1}{Z} \prod_{\substack{j=1 \\ a_j=a_i}}^m \hat{P}_\theta(y_j | x_j, z) \hat{P}(z) \quad (4)$$

Algorithm 1 TRAIN($\{x_i, a_i, y_i\}_{i=1}^m$)

```

init  $\hat{P}(Z)$  as uniform, and init  $\theta$ 
for  $k = 0..n_{\text{iter}} - 1$  do
  if  $k \bmod n_{\text{reassign}} = 0$  then
    compute  $\hat{P}_\theta(z | a)$  for all annotators
    perform a gradient descent step
  update the prior  $\hat{P}(Z)$  for later inference

```

where Z is the partition function, and $\hat{P}(z)$ is a prior on the group distribution – which can be assumed to be uniform in case of lack of information. We use a Neural Network, specifically a Multi-Layer Perceptron (MLP), for the estimator $\hat{P}_\theta(Y | X, Z)$, and we train it via the Expectation-Maximization process described in Algorithm 1. The procedure alternates between n_{reassign} gradient descent steps where we optimize the group behavior estimator, and the recomputation of the group assignment posterior $\hat{P}_{\theta,S}(Z | A)$. During training, we use a uniform prior $\hat{P}(z)$, which is updated after the process to reflect the inferred group distribution. Additional details can be found in Appendix B, which also covers tempering techniques that can be applied to mitigate sampling noise.

3.2 Analyses Enabled by FLOCKE

By relying on the learned group behavior estimator, FLOCKE enables several types of analysis on datasets with disaggregated annotations, presented in this section. We assume a classification task, but in most cases the described procedures can be easily adapted to regression. An overview of the analyses enabled by FLOCKE is shown in Fig. 1.

3.2.1 Group Number Detection

Since our groups are behavior-based, and therefore implicit, their number is not known *a-priori*. FLOCKE allows inferring group number via Maximum Likelihood Estimation: one can test multiple

numbers of groups n , and assess the corresponding likelihood values, computed as in Eq. (2). Once n matches the number of groups naturally occurring in the data, the likelihood should plateau or decrease, so that an optimal number of groups can be estimated via the elbow method commonly used in clustering. Such evaluation can be performed either on the training data (as in conventional clustering), or on a held-out dataset.

As a performance metric for this analysis, we use the *calibration gain* $\Delta_\ell L$, defined as the difference between the likelihood for a reference number of groups n_{gt} (e.g., the elbow method estimate, or a known ground truth in a controlled setting) and that for $n_{gt} - \ell$ groups. For example, if $n_{gt} = 3$, $\Delta_{-1}L$ is the likelihood score gain between $n = 3$ and $n = 2$ and $\Delta_{-2}L$ is the likelihood score gain between $n = 3$ and $n = 1$. When the calibration gain $\Delta_\ell L > 0$, then assuming more than one group is leading to a better model for the annotation behavior. Conversely, $\Delta_\ell L \leq 0$ indicates that introducing groups does not provide further explanatory power.

This task is a prerequisite for all subsequent types of analysis. It can also be used to gauge the level of behavioral diversity in a dataset.

3.2.2 Annotator Group Distribution

Given examples for a single annotator, FLOCKE can be used to estimate their **group membership**, based on the annotation pattern. This is done by computing $\hat{P}_{\theta,S}(z | a_i)$ as described in Eq. (4) with fixed (pre-trained) model weights. This task can be understood as a form of soft-clustering (if we use the group probabilities) or a hard-clustering (if we assign annotators to their most likely group). The analysis can also be performed on a novel dataset, possibly annotated by different individuals, i.e., our approach can predict the group distribution for annotators not seen during model training.

If examples are available for multiple annotators, the **level of representation of each group** z can be estimated by aggregation, i.e.,

$$\hat{P}_{\theta,S}(z) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \hat{P}_{\theta,S}(z | a_i) \quad (5)$$

where $\mathcal{A}$ is the annotator set and $|\mathcal{A}|$ its cardinality.

One can use this analysis to better quantify the behavioral diversity in a pool of annotators, or to compare different annotator pools. By recruiting annotators based on their estimated group member-

ship, one could also improve the level of annotation diversity, or build a fairer annotator pool.

3.2.3 Annotation Prediction

The group behavior estimator $\hat{P}_{\theta,S}(Y | X, Z)$, can be directly used to perform **group-level annotation prediction**. Specifically, we can either estimate the group response distribution (by straightforward evaluation), or a single label prediction (by taking the output with the largest estimated probability).

We can also perform **individual annotation prediction**, by averaging the group predictions over the group distribution for a single annotator a :

$$\hat{P}_{\theta,S}(y | x, a) = \sum_{z=1}^n \hat{P}_{\theta,S}(z | a) \hat{P}_{\theta}(y | x, z) \quad (6)$$

Note that this task requires having estimated the group distribution for the annotator, based on a set of examples $\mathcal{S}$. As in the previous case, we can use Eq. (6) either to estimate the label probabilities or to predict the most likely label. Finally, given a set $\mathcal{A}$ of annotators, we can estimate the aggregated label distribution (i.e., the **soft label**):

$$\hat{P}_{\theta,S}(y | x, \mathcal{A}) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \hat{P}_{\theta,S}(y | x, a) \quad (7)$$

with the requirements already discussed.

Inspecting group predictions may highlight fairness issues, or just to compare groups. Group predictions can also provide deeper insight into group behavior, for example via explainability methods such as SHAP (Lundberg and Lee, 2017), if the contextual information x is interpretable. Individual predictions and soft labels can assist in the investigation of disagreement, or be used to optimize the assignment of examples to annotators in empirical user studies.

3.2.4 Most Relevant Items

Group-level annotation prediction can be employed to identify the **most relevant items** in terms of their ability to elicit group disagreement. In particular, given two groups z', z'' , the relevance of an example x can be measured in terms of the difference between the two predicted label distributions, e.g., using the Jensen-Shannon divergence:

$$\text{JSD}(\hat{P}_{\theta,S}(Y | x, z') \parallel \hat{P}_{\theta,S}(Y | x, z'')) \quad (8)$$

Any statistical distance can be used in place of JSD. The relevance for multiple groups can be assessed

by averaging over all pairwise relevance values, at the cost of a less interpretable outcome.

Isolating the most relevant items can provide a form of explainability when the contextual information x is not easily interpretable (as in our experiments, where we use text embeddings). It can also play a role in the design of experiments to investigate disagreement, or in assessing how much of the observed disagreement is due to noise, as opposed to actual subjective point of views.

4 Method Validation

We validate FLOCKE through experiments based on synthetic data, allowing us to simulate annotator disagreement in a controlled setting. Since the synthetic setup provides access to the ground-truth annotator groups, we can directly evaluate whether FLOCKE is able to recover the underlying disagreement structure.

4.1 Synthetic Annotator Groups

To validate our framework, we rely on HateCheck (Röttger et al., 2021), a functional test set for hate speech detection models composed of hateful and non-hateful synthetic texts created using 406 templates (e.g., “I hate/love [target identity]”), including 7 identity groups (e.g., *immigrant*, *woman*) and 29 functionalities (e.g., *hate expressed using slur*) to cover distinct expressions of hate and non-hate. It includes 3,728 examples annotated with a binary hate speech label.

Given the template-based structure of HateCheck, the underlying template semantics provide us with ground-truth labels. We exploit this feature to create different sets of artificially controlled annotations, simulating different annotator behaviors. Specifically, we introduce variations along the *Target identity* dimension to simulate groups of annotators following specific annotation patterns.³

Annotator group patterns consist in systematically labeling as non-hateful all examples mentioning one or more specific target identities, while examples not involving those identities retain their original HateCheck labels.⁴ Figure 2 shows an example of synthetic labels. The target identities associated with the *non-hateful* labeling pattern, as

Women are the absolute worst. $\left\{ \begin{array}{l} \text{Group 0: Hate} \\ \text{Group 1: Non-hate} \end{array} \right.$
Ground truth: Hate

Figure 2: Example of synthetic labels with 2 annotator groups and the *women* target identity for Group 1.

well as their possible combinations, are assigned randomly across groups. This construction provides a ground truth number of groups n_{gt} and a labeling pattern for each annotator group. We can therefore systematically vary the number of annotator groups, as well as the number of annotators per group, and the number of annotations collected for each item. Annotators are then randomly assigned to annotate items, such that a given item may be annotated either by annotators from different groups or by multiple annotators belonging to the same group. Despite this randomized annotation process, the underlying ground-truth label associated with each group-item pair remains known.

4.2 Validation Design

We evaluate FLOCKE across multiple synthetic configurations by varying the number of ground-truth n_{gt} annotator groups generated through the procedure described above. Since our objective is to identify a small number of coherent annotator perspectives, for the purpose of validation we adopt $n_{gt} \in [2, 5]$, as in real scenarios a larger number of groups would reduce interpretability.

To isolate the effect of each factor and avoid introducing excessive variability at once, for each value of n_{gt} we repeat the experiments with different configurations by varying the number of annotators per group (10, 20, 50, and 100) and the number of annotations collected per item (2, 5, and 10). This results in a total of 12 experimental configurations for each choice of artificial number of groups, that range from a minimum number of synthetic annotations per annotator⁵ of 74 to a maximum of 3,728. To account for stochasticity, each configuration is evaluated across 5 independent runs with different randomly generated group annotation patterns. The goal of this synthetic setup is not to reproduce the full complexity of naturally occurring disagreement, but to create a controlled benchmark in which the latent group structure is known and the method can be directly evaluated.

The input features for FLOCKE consist of the

³The *Functionality* dimension is instead used to introduce individual-level biases independent of group membership (see Appendix C.2).

⁴The pattern introduced is unidirectional with systematic shifts toward non-hateful labels, whereas no meaningful opposite shift is possible, given the semantic of the templates.

⁵Note that these figures refer to the “average annotator”. Since (a, x) pairs are assigned randomly, one annotator can have more annotations than another.

sentence representation of the item embedded with multilingual BERT (Devlin et al., 2019).⁶ This model has been selected due to its widespread adoption in the NLP community, its native multilingual capabilities, and the computationally efficient embedding size of 768.

4.3 Validation Tasks and Results

We validate FLOCKE along the analyses detailed in Section 3.2.

4.3.1 Group Number Detection

We first aim to validate the ability of the algorithm to recover the ground-truth number of annotator groups. For each experimental configuration, we fit FLOCKE varying n as mentioned in Section 3.2.1. Fig. 3 shows the likelihood score aggregating results across different n with respect to n_{gt} , e.g., if we define $n_{gt} = 3$ groups in the synthetic data and set $n = 3$ in FLOCKE, then we are in the correct case; while if we set $n = 4$, then we are in the +1 case. In Fig. 3 we observe the likelihood increasing until the correct number of groups is reached, plateauing afterwards. Across the tested n_{gt} , we observe an average $\Delta_{-1}L = 0.13$ and $\Delta_{-2}L = 0.17$. This validates the ability of FLOCKE to recover the ground-truth number of groups.

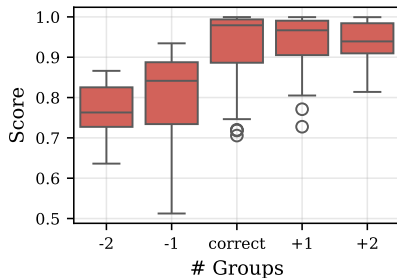


Figure 3: Calibration of the number of groups on validation data. The box-plots show the likelihood score distribution across different configurations, with median scores (middle lines) and outliers (dots).

4.3.2 Annotator Group Distribution

We validate FLOCKE’s ability to correctly cluster annotators according to their simulated annotation behavior. To evaluate clustering quality, we measure *Accuracy*, *Purity* (“does each predicted group mostly contain annotators from a single true group?”), and *Completeness* (“does each true group mostly fall into a single predicted group?”). This assessment utilizes a hard clustering approach from

FLOCKE, assigning each annotator to their most probable group. The resulting clusters are then evaluated against the ground-truth groupings to compute performance metrics. For this analysis, we fix n to the correct number of groups, after having demonstrated how this can be identified in Section 4.3.1, in order to focus on the ability to correctly cluster annotators. We observe an average accuracy of 0.93 ($\sigma = 0.16$, purity = 0.93; completeness = 0.99), validating the clustering capabilities of our approach (more details in Appendix C). In particular, the high completeness score demonstrates that each true group mostly falls into a single predicted group. Despite variability, we observe no consistent trend with respect to the number of annotators or annotations/item, showing that FLOCKE remains effective with as few as two annotations/item.

4.3.3 Annotation Prediction

We validate the ability to predict group-level annotations on the training set and on a held-out set. In order to assess FLOCKE performance in modeling group behaviors, we measure the macro F1 score in predicting group-level annotations, as described in Section 3.2.3.

First, in each configuration, we validate the performance of FLOCKE in predicting labels on the training set. This gives us a reference of the method’s ability to *reconstruct* group annotation patterns. Then, we split training and testing data (80% and 20% respectively), and measure FLOCKE performance in predicting labels on the test set when trained only on the first split. The results, presented in Fig. 4b, show the annotation prediction performance across all settings. We observe an average F1 score of 0.82 (compared to 0.94 obtained on the training set). Fig. 4a visually compares the

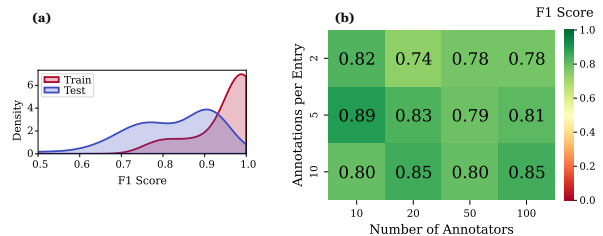


Figure 4: Validation results of FLOCKE ability in predicting group-level annotations on a held-out test set. (a) Comparison of density functions of F1 scores for different experiments, on train- vs test-set. (b) Macro-F1 group-level label prediction averaged across all experimental configurations and runs.

⁶huggingface.co/google-bert/bert-base-multilingual-cased

annotation prediction performance achieved on the training and the held-out set for all configurations. We observe that, despite a generally lower F1 score on the test set, FLOCKE correctly predicts specific group behavior in the large majority of cases.

4.3.4 Most Relevant Items

We validate the ability of FLOCKE to identify items with disagreement across groups, even when the observed annotations, randomly assigned, originate from annotators belonging to the same group.

For any two groups g_1, g_2 , the ground-truth set of most relevant items, i.e., examples on which the two groups would assign different labels, is known and has size m_{g_1, g_2} . The predicted relevance of an item can be computed using the Jensen-Shannon divergence as explained in Eq. (8). An example of relevance-based item ranking is presented in Fig. 5a. For each configuration, we evaluate retrieval performance independently for every pair of groups (g_1, g_2). The prediction of the model is evaluated by computing the F1 score between the predicted and ground-truth top- m_{g_1, g_2} items for that pair. The final score for a configuration is obtained by macro-averaging the F1 scores across all group pairs.

We study how this performance metric varies across different numbers of annotators and annotations (Fig. 5b), observing that even in the most challenging configurations (i.e., 2 annotations/item) the F1 score is consistently higher than 0.74, thus confirming the ability of our method to retrieve group-level relevant items.

4.3.5 Robustness Checks

To further assess the stability of the validation results, we test FLOCKE also with imbalanced anno-

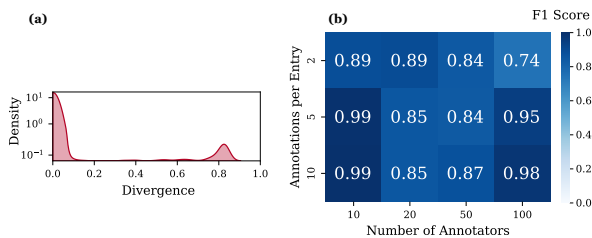


Figure 5: Validation results of FLOCKE’s ability in retrieving most relevant items. (a) Distribution of Jensen-Shannon divergence values (between group predictions) across examples. The most relevant items predicted by FLOCKE are clearly separated and clustered around JSD = 0.80. (b) Prediction of most relevant items (F1 Score); for each configuration, the score is computed by averaging across group-pair F1 scores.

tator groups, in which group sizes are randomly varied while keeping the total number of annotators fixed. Then, we introduce individual-level biases, modeled as annotator-specific variations independent of group membership. In both cases, the results are consistent with the main validation experiments, showing that FLOCKE can still recover group structure and identify disagreement-relevant items under more challenging conditions. Detailed results are reported in Appendix C. Moreover, we perform the same analyses with different language models other than multilingual BERT (namely, RoBERTa,⁷ RoBERTa-large⁸ and SBERT⁹), in order to assess robustness across vector spaces. The models yield largely consistent results, despite minor variations in specific outcomes. In particular, as embedding size increases, we observe weaker calibration gain (SBERT: $\Delta_{-2}L = 0.16$; RoBERTa-base: $\Delta_{-2}L = 0.13$; RoBERTa-large: $\Delta_{-2}L = 0.11$) and lower F1 scores in retrieving most relevant items (SBERT: $F_1 = 0.88$; RoBERTa-base: $F_1 = 0.74$; RoBERTa-large: $F_1 = 0.73$). A possible explanation is that lower-dimensional representations encourage the model to capture behaviors shared across groups, while higher-dimensional embeddings encode finer item-specific semantic distinctions, so that FLOCKE no longer needs latent groups to explain annotation variation. A complete summary of results is provided in Appendix C.3.

5 Analyses on Real-World Data

After validating FLOCKE in a controlled synthetic setting, we show how it can be used to analyze real datasets with disaggregated annotations.

5.1 MD-Agreement

We apply FLOCKE to the MD-Agreement dataset (Leonardelli et al., 2021), a dataset for offensiveness detection containing 10,753 tweets, with 5 annotations per item, annotated by 819 sparsely represented annotators.

Groups characterization We first estimate the number of annotator groups by varying the number of latent groups n , as described in Section 3.2.1. The likelihood scores plateau at $n = 3$, suggesting the presence of three main annotator groups in the dataset, with a calibration gain of $\Delta_{-2}L = 0.06$.¹⁰

⁷huggingface.co/FacebookAI/roberta-base

⁸huggingface.co/FacebookAI/roberta-large

⁹huggingface.co/sentence-transformers/all-MiniLM-L12-v2

¹⁰Not only do we observe a positive gain relative to $n = 1$, but we also see negative values for $n > 3$ ($\Delta_{+1}L = -0.02$)

After fixing $n = 3$, we inspect the distribution of annotators across the inferred groups in the training set. The resulting distribution is moderately imbalanced: 277 annotators are assigned to Group 0, 188 to Group 1, and 205 to Group 2. This suggests that the dataset contains a dominant annotation behavior, together with smaller groups capturing alternative labeling patterns.

We then use the trained group-level predictors to analyze the held-out test set. Performing the analysis on the test set allows us to assess whether the inferred group structure generalizes to unseen items, rather than only capturing patterns present in the training data. For each test item, we compute the average pairwise JSD between group-level predicted label distributions. Groups 0 and 2 emerge as the most distant pair, with a median JSD of 0.289, compared to 0.170 for Groups 1 and 2 and 0.123 for Groups 0 and 1. This suggests that Groups 0 and 1 capture more similar annotation patterns, while Group 2 is more clearly separated.

Most-relevant items To further characterize the differences between inferred groups on unseen data, we rank test items according to their average pairwise JSD across group-level predictions. We observe that items in the upper part of the ranking are generally associated with lower annotator agreement, while full-agreement items become more frequent toward the bottom of the ranking.¹¹ This suggests that FLOCKE’s most-relevant items capture meaningful variation in annotator behavior: items with lower observed agreement tend to be more group-discriminative, while full-agreement items are more often assigned lower divergence. The ranking can therefore be used to identify examples that are likely to reveal differences between inferred annotator groups. We manually inspect the 100 most relevant disagreement items by comparing the text of each instance with the labels assigned by the groups. Sources of disagreement are not defined a-priori, but rather discovered during observation. The main sources of disagreement seem to concern how sensitive each group is to politically charged content, direct insults, profanity,

and $\Delta_{+2}L = -0.01$), suggesting that additional groups do not improve the modeling of annotation behavior beyond $n = 3$.

¹¹The test set includes low-agreement, medium-agreement, and full-agreement items in comparable proportions. When dividing the ranked test items into three equally sized portions, full-agreement items increase from 38.7% in the first third to 54.8% in the last third. Conversely, low-agreement items decrease from 30.5% to 20.1%, while medium-agreement items decrease from 30.8% to 25.1%.

Item	G0	G1	G2
<user> <user> Flying in the face of science logic and common sense. People are dying and you don’t give a shit	0	0	1

Table 2: Top-ranked item for group-level (G0, G1, G2) disagreement and labels predicted by each annotator group (“0” = non-offensive).

and identity-related discourse. In particular, group 2 appears to adopt a broader criterion for defining offensive language, while group 0 appears to be the least sensitive. Table 2 illustrates this pattern with the top-ranked disagreement item.

Soft-label prediction. Finally, we evaluate FLOCKE on the prediction of soft annotation distributions on the held-out test set. We first compute the cross-entropy obtained by each group-specific predictor, corresponding to 0.782 for Group 0, 0.699 for Group 1, and 0.801 for Group 2. These values reflect the performance of each inferred group when considered independently.

We then exploit the estimated group-membership distribution of each annotator to obtain annotator-level predictions. For the 109 test annotators observed during training, predictions are conditioned on their learned group-membership distribution. The test split also includes 137 annotators not observed during training. For these unseen annotators, FLOCKE estimates their behavior as a weighted combination of the learned group-level distributions. When annotators are represented through these group weights, the overall cross-entropy decreases to 0.67.¹² This indicates that combining group-level predictors according to annotator-specific group membership provides a better approximation of soft annotation distributions than considering each group independently, also in a setting with annotators that were not observed during training.

5.2 MultiPICo

We further apply FLOCKE to MultiPICo (Casola et al., 2024), a dataset for irony detection consisting of 18,778 post-reply pairs in 9 languages, annotated by 506 individuals. This dataset also includes socio-

¹²In terms of the LeWiDi shared-task ranking, this score would correspond to 22nd place out of 26 submitted systems. However, FLOCKE is not optimized as a leaderboard system: our objective is to learn a compact and interpretable set of latent annotator groups. Using more groups or tuning the model primarily for predictive performance may improve CE, but would make the resulting group structure less interpretable.

demographic traits of annotators (age, gender, ethnicity, student status, employment status, country of birth, country of residence, and nationality), which allows us to investigate associations between latent groups and socio-demographics. While [Cassola et al. \(2024\)](#) found that age and student status are the most polarizing dimensions, they did not perform a statistical test of independence with annotator groups, as they did not explore annotator clustering.

For our analysis, we focus on the subset of English conversations and annotators. We estimate the number of latent groups (with the method presented in Section 3.2.1), finding a likelihood plateau after three groups ($\Delta_{-2}L = 0.07$ and $\Delta_{+1}L = -0.01$). Statistical association between learned latent groups of annotators and their socio-demographic data is then tested, performing a Pearson chi-square test. For variables with fine-grained, sparse categories, coarser groupings are adopted: age is grouped into five bins (18–24, 25–34, 35–44, 45–54, 55+), while continent is used instead of country of birth/residence and nationality. We find no statistically significant associations either for individual socio-demographic variables or for their pair-wise intersections. In particular, the smallest p-values are observed for country of residence ($p = 0.229$) and the interaction $\langle \text{gender}, \text{nationality} \rangle$ ($p = 0.111$), respectively. These findings indicate that learned latent groups emerge independently from the considered socio-demographic traits.

5.3 Discussion

Our framework is a probabilistic model that jointly learns the task at hand and the latent groups of annotators. This is not just a mere convenience of having multiple analyses in a unique framework, but rather a paradigm with practical implications: FLOCKE enables both individual and group-level prediction, the characterization of latent groups through the most relevant items (unavailable in [Venanzi et al., 2014](#) and [Fell et al., 2021](#)), and supports the prediction of items most prone to disagreement, differently from [Vitsakis et al. \(2024\)](#) and [Li et al. \(2024\)](#).

The MultiPICO case study illustrates the possibility of testing, instead of assuming, a link between latent groups and socio-demographics. We find no significant associations, suggesting that what differentiates annotators is not well captured by approaches using socio-demographics as a proxy of

perspective ([Fell et al., 2021](#); [Lo and Basile, 2023](#); [Vitsakis et al., 2024](#); [Li et al., 2024](#)).

Finally, we do not assume a fixed number of latent groups, but rather treat it as an empirical (likelihood-based) question, in order to observe how annotators are genuinely divided. This is in contrast to approaches adopting a fixed number of groups ([Fell et al., 2021](#)), relying on internal metrics ([Lo and Basile, 2023](#); [Li et al., 2024](#)), or assessing validity against socio-demographic variables ([Vitsakis et al., 2024](#)).

6 Conclusion

We introduced FLOCKE, a framework for analyzing disaggregated annotations by jointly learning group-level annotation behavior and latent annotator group membership. We also formalized six kinds of analyses enabled by our method, such as group-level annotation prediction and most relevant items retrieval (which require finding the optimal number of groups). Additionally, we introduced a novel validation method using synthetic data.

FLOCKE offers an intermediate solution between majority-vote aggregation and more fine-grained alternatives, such as perspectivist modeling and soft-label distributions, as promoted in the LeWiDi shared tasks. While these approaches move beyond single-label supervision, they can be difficult to operationalize in real-world scenarios with sparse annotator-level data or pipelines designed around discrete labels. FLOCKE instead captures meaningful variation through latent annotator groups, keeping the representation compact and interpretable.

The framework can be applied and customized to any existing dataset with disaggregated annotations. FLOCKE can also be used beyond retrospective analysis: given a small set of annotations from an unseen annotator, the model can estimate their group distribution and predict their labels on new items as a weighted combination of group-specific predictions. The contribution of each group can also be reweighted to simulate different annotator populations or to give more importance to under-represented perspectives. Finally, FLOCKE can be straightforwardly applied to model LLMs as annotators and analyze their latent group quality and disagreement patterns, or in a mixed scenario with human and LLM annotations to create datasets with different perspectives.

Limitations

A limitation of our work is that the validation is based mostly on synthetic data. This choice is actually necessary from a methodological point of view, as existing datasets with disaggregated annotations typically provide annotators’ socio-demographic attributes, but none provide different annotator group behaviors. As our goal was not that of coupling behavior with sociodemographic attributes (as done in previous work), we constructed a synthetic dataset with ground-truth annotator behavior structure in order to validate our method. To mitigate this weakness, we also test FLOCKE on real-world data. Although this means that we do not have ground-truth annotator behavior groups for real data, it allows us to test our framework in a real-world scenario, to understand its potential usefulness.

Furthermore, FLOCKE validation and the analyses on real data are performed on binary classification tasks of short texts in similar scenarios, i.e., hate speech detection for HateCheck, offensiveness detection for MD-Agreement, and irony detection for MultiPICO. Due to space constraints we could not show results for other tasks. However, FLOCKE is task-agnostic and could indeed be applied to any dataset with disaggregated annotations, both with textual and non-textual input. Also, its adaptation to multi-label classification or regression tasks would be possible, requiring minimal effort. We leave these additional analyses and implementations to future work.

Additionally, our analysis with MultiPICO is restricted to the English subset. While this is justified by the fact that there is a distinct set of annotators for each language, it prevents us from drawing conclusions on how latent groups relate to socio-demographics in languages other than English.

Finally, FLOCKE relies on observing sufficiently informative annotation patterns to infer stable latent annotator groups. In settings with very sparse annotations per annotator, limited overlap between annotators and items, or very low number of annotators, group assignments may become unstable or less identifiable.

Ethics Statement

We do not envision any specific ethical risk related to this work. On the contrary, one of the applications of FLOCKE is the identification of annotator groups to ensure that different perspectives are taken into account in a dataset with disaggregated

annotations. Furthermore, FLOCKE does not require any socio-demographic information to model annotators and no explicit profiling is performed, keeping predicted groups agnostic with regards to socio-demographic characteristics. Moreover, we respected the intended use of all artifacts used in this work.

FLOCKE does not require extensive computational resources, making it a feasible choice in low-resource scenarios, as it is possible to run it on CPU. All experiments of this paper were conducted using in-house computational resources, totaling around 150 hours on a Nvidia Titan Xp GPU.

References

- Gavin Abercrombie, Valerio Basile, Davide Bernardi, Shiran Dudy, Simona Frenda, Lucy Havens, Elisa Leonardelli, and Sara Tonelli, editors. 2023. *Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP co-located with 26th European Conference on Artificial Intelligence (ECAI 2023), Kraków, Poland, September 30th, 2023*, volume 3494 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Gavin Abercrombie, Valerio Basile, Davide Bernardi, Shiran Dudy, Simona Frenda, Lucy Havens, and Sara Tonelli, editors. 2024. *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP, NLPerspectives@LREC-COLING 2024, Torino, Italy, May 21, 2024*. ELRA and ICCL.
- Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma, editors. 2022. *Proceedings of the 1st Workshop on Perspectivist Approaches to NLPerspectives@LREC 2022, Marseille, France, 20th June 2022*. European Language Resources Association.
- Alejandro Benito-Santos and Adrian Ghajari. 2025. *Beyond averages: Learning with annotator disagreement in STS*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35551–35557, Suzhou, China. Association for Computational Linguistics.
- Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. *MultiPICO: Multilingual perspectivist irony corpus*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.
- Naihao Deng, Xinliang Frederick Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. *You are what you annotate: Towards better models through annotator representations*. In *Findings of the Association for Computational Linguistics: EMNLP*

- 2023, pages 12475–12498, Singapore. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Michael Fell, Sohail Akhtar, and Valerio Basile. 2021. [Mining annotator perspectives from hate speech corpora](#). In *Proceedings of the Fifth Workshop on Natural Language for Artificial Intelligence (NL4AI)*.
- 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, Singapore. Association for Computational Linguistics.
- Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. [Perspectivist approaches to natural language processing: a survey](#). *Language Resources and Evaluation*, 59(2):1719–1746.
- Amir Homayounirad, Enrico Liscio, Tong Wang, Catholijn M Jonker, and Luciano Cavalcante Siebert. 2025. [Will annotators disagree? Identifying subjectivity in value-laden arguments](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15237–15252, Suzhou, China. Association for Computational Linguistics.
- EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. [Aligning language models to user opinions](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore. Association for Computational Linguistics.
- Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. [Investigating reasons for disagreement in natural language inference](#). *Transactions of the Association for Computational Linguistics*, 10:1357–1374.
- Urja Khurana, Eric Nalisnick, Antske Fokkens, and Swabha Swayamdipta. 2024. [Crowd-calibrator: Can annotator disagreement inform calibration in subjective tasks?](#) *Preprint*, arXiv:2408.14141.
- Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. [Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach](#). *Information Processing & Management*, 58(5):102643.
- Kemal Kurniawan, Meladel Mistica, Timothy Baldwin, and Jey Han Lau. 2026. [On the interplay between human label variation and model fairness](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 967–976, Rabat, Morocco. Association for Computational Linguistics.
- Elisa Leonardelli, Gavin Abercrombie, Dina Almane, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. [Semeval-2023 task 11: Learning with disagreements \(LeWiDi\)](#). In *Proceedings of the The 17th International Workshop on Semantic Evaluation, SemEval@ACL 2023, Toronto, Canada, 13-14 July 2023*, pages 2304–2318. Association for Computational Linguistics.
- Elisa Leonardelli, Silvia Casola, Siyao Peng, Giulia Rizzi, Valerio Basile, Elisabetta Fersini, Diego Frassinelli, Hyewon Jang, Maja Pavlovic, Barbara Plank, and Massimo Poesio. 2025. [LeWiDi-2025 at NLPerspectives: Third edition of the learning with disagreements shared task](#). In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 182–195, Suzhou, China. Association for Computational Linguistics.
- Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. [Agreeing to disagree: Annotating offensive language datasets with annotators’ disagreement](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Junyi Li, Charith Peris, Ninareh Mehrabi, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2024. [The steerability of large language models toward data-driven personas](#). 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 7290–7305, Mexico City, Mexico. Association for Computational Linguistics.
- Haochen Liu, Joseph Thekinen, Sinem Mollaoglu, Da Tang, Ji Yang, Youlong Cheng, Hui Liu, and Jiliang Tang. 2022. [Toward annotator group bias in crowdsourcing](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1797–1806, Dublin, Ireland. Association for Computational Linguistics.
- Soda Marem Lo and Valerio Basile. 2023. [Hierarchical clustering of label-based annotator representations for mining perspectives](#). In *Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP*, volume 3494, pages 1–10. CEUR-WS.
- Soda Marem Lo, Silvia Casola, Erhan Sezerer, Valerio Basile, Franco Sansonetti, Antonio Uva, and Davide Bernardi. 2025. [PERSEVAL: A framework for perspectivist classification evaluation](#). In *Proceedings of the 2025 Conference on Empirical Methods in*

- Natural Language Processing*, pages 22334–22359, Suzhou, China. Association for Computational Linguistics.
- Scott M Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2025. [Benchmarking distributional alignment of large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 24–49, Albuquerque, New Mexico. Association for Computational Linguistics.
- Negar Mokherian, 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, Mexico City, Mexico. Association for Computational Linguistics.
- 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.
- Jingwei Ni, Yu Fan, Vilém Zouhar, Donya Rooein, Alexander Miserlis Hoyle, Mrinmaya Sachan, Markus Leippold, Dirk Hovy, and Elliott Ash. 2026. [Can reasoning help large language models capture human annotator disagreement?](#) In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–54, Rabat, Morocco. Association for Computational Linguistics.
- Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. [Beyond demographics: Fine-tuning large language models to predict individuals’ subjective text perceptions](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2111, Vienna, Austria. Association for Computational Linguistics.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Aida Mostafazadeh Davani, Alicia Parrish, Alex Taylor, Mark Diaz, Ding Wang, and Gregory Serapio-García. 2024. [GRASP: A disagreement analysis framework to assess group associations in perspectives](#). 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 3473–3492, Mexico City, Mexico. Association for Computational Linguistics.
- Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. [Soft metrics for evaluation with disagreements: an assessment](#). In *Proceedings of the 3rd Workshop on Perspective Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94, Torino, Italia. ELRA and ICCL.
- Giulia Rizzi, Paolo Rosso, and Elisabetta Fersini. 2025. [Is a bunch of words enough to detect disagreement in hateful content?](#) In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 1–11, Abu Dhabi, UAE. International Committee on Computational Linguistics.
- 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, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.
- Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. [HateCheck: Functional tests for hate speech detection models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.
- Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. [Why don’t you do it right? Analysing annotators’ disagreement in subjective tasks](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.
- 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, Suzhou, China. Association for Computational Linguistics.
- Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. [Position: a roadmap to pluralistic alignment](#). In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.

Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. 2014. [Community-based bayesian aggregation models for crowdsourcing](#). In *Proceedings of the 23rd international conference on World wide web*, pages 155–164.

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, Miami, Florida, USA. Association for Computational Linguistics.

A Probabilistic Model and Assumptions

Overall, our probabilistic model involves 4 random variables: X , representing the contextual information; Y , representing the label/output; A , representing the annotator; Z , representing the group. We refactor the joint distribution as follows:

$$P(X, Y, Z, A) \quad (9)$$

$$= P(Y | X, Z, A)P(X, Z, A) \quad (10)$$

$$= P(Y | X, Z)P(Z | X, A)P(X, A) \quad (11)$$

$$= P(Y | X, Z)P(Z | A)P(X, A) \quad (12)$$

where: 1) in Eq. (10) we assume that discrepancies between annotators are captured through their (latent) group distribution; 2) in Eq. (12), we make the (trivial) assumption that groups are associated to annotators rather than items. Both assumptions are consistent with viewing the annotator behavior as determined by the groups, as stated in Section 3.1.

The likelihood expression from Eq. (2) is derived from Eq. (12) by: 1) relying on the sample data to capture $P(X, A)$, i.e., via Empirical Risk Minimization; 2) marginalizing out the latent categorical variable Z ; and 3) using a posterior distribution to estimate $P(Z | A)$ based on the available data.

Concerning such posterior distribution, let S_a be a set of observed input/output pairs for a single annotator a , i.e., a sample from $P(X, Y | a)$. Then we start by observing that:

$$P(Z | X, Y) = \frac{P(X, Y, Z)}{P(X, Y)} \quad (13)$$

$$= \frac{P(Y | X, Z)P(X, Z)}{P(X, Y)} \quad (14)$$

$$= \frac{P(Y | X, Z)P(X)P(Z)}{P(X, Y)} \quad (15)$$

$$= \frac{1}{Z} P(Y | X, Z)P(Z) \quad (16)$$

where in Eq. (14) we use the fact that X and Z are assumed to be independent, again as stated in Section 3.1, while in Eq. (16) we simply model via a partition function all terms that do not depend on the group. The estimate from Eq. (4) is then derived by relying on the sample S_a for X and Y and using a prior for $P(Z)$.

B Comments on the Training Algorithm

The training algorithm we use in our implementation follows closely Algorithm 1, but a few points deserve further discussion. First, as the vast majority of modern differentiable Machine Learning

approaches, we rely on (mini-batch) stochastic gradient descent, rather than classical gradient descent. Specifically, the algorithm is implemented in PyTorch and we use the AdamW optimizer. We use the default PyTorch weight initializer for θ .

Second, after the training process is over we update the $\hat{P}(Z)$ prior to reflect the group prevalence observed in the training data. This is done as:

$$\hat{P}(z) \leftarrow \beta' \hat{P}_{\theta, S}(z) + (1 - \beta') \frac{1}{n} \quad (17)$$

where $\hat{P}_{\theta, S}(z)$ is the group prevalence, computed as in Section 3.2.2, and $\beta' \in [0, 1]$ is a tempering parameter. For $\beta' = 1$, the inferred group prevalence on the training data becomes the new prior; for $\beta' = 0$, the prior stays uniform.

Third, the algorithm features an optional tempering mechanism to mitigate overconfidence in the inferred group distribution for an annotator. Specifically, we determine the group distribution via a slightly more complex version of Eq. (4):

$$\sqrt[\beta'']{\frac{1}{Z} \prod_{\substack{j=1 \\ a_j=a_i}}^m \hat{P}_{\theta, S}(y_j | x_j, z) \hat{P}(z)} \quad (18)$$

where $\beta'' \geq 1$ is a second tempering parameter, which smooths the group distribution when it takes a value larger than 1. In particular, for $\beta'' = m$ the accumulation of evidence from multiple samples is largely ignored, while for $\beta'' = 1$ we recover Eq. (4). All reported experiments are performed with $\beta'' = 1$.

C Validation

Fig. 6 shows the detailed results of annotator group distribution for all configurations.

C.1 Modeling with Imbalanced Groups

In real world setting, the distribution of annotators across groups can be imbalanced. For this reason, we repeat the validation set of experiments, where in each configuration the total number of annotators

	$\Delta_{-1}L$	$\Delta_{-2}L$	F1	Accuracy
Balanced	.13	.17	.89 $\pm$.18	.93 $\pm$.16
Imbalanced	.06	.13	.80 $\pm$.17	.95 $\pm$.09

Table 3: Comparison of validation results with Balanced vs Imbalanced groups: *calibration gain* (ΔL); most relevant items (F1 score); and grouping Accuracy.

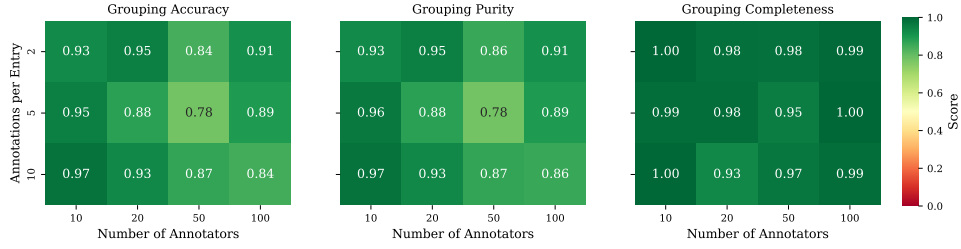


Figure 6: Grouping performance of FLOCKE across all configurations. The averaged score is showed, since for each configuration multiple sets of groups are tested.

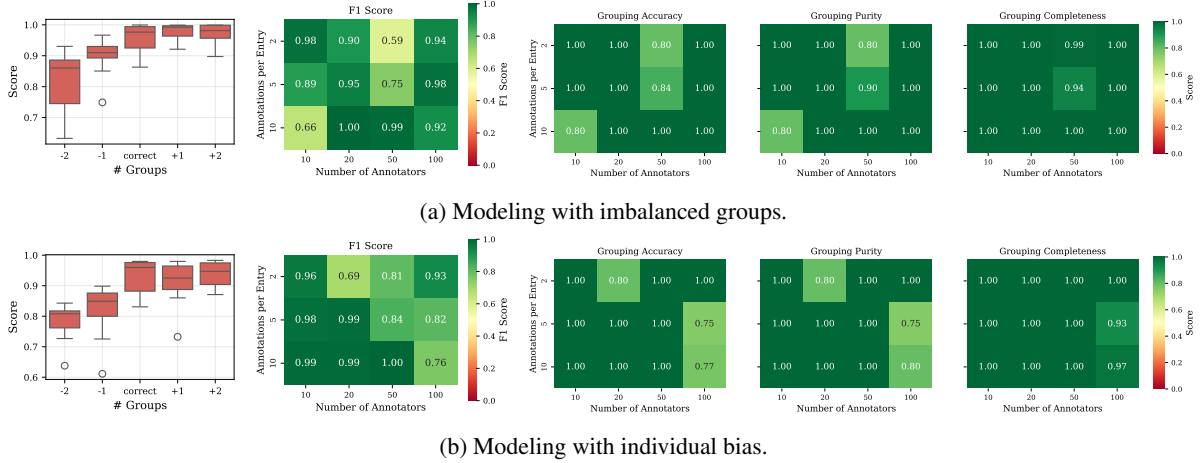


Figure 7: Detailed results of robustness checks, with (a) imbalanced groups and (b) individual bias.

is the same as in the previous validation, but group size ratios are randomly assigned so that the ratio of each group z is $r_z \in [0.1, 0.9]$. The results presented in Table 3 validate the possibility to learn groups under imbalances. In particular on average $\Delta_{-2}L = 0.13$, proving how the number of groups can be detected with imbalanced groups, despite a less accentuated difference. Similarly, a high grouping accuracy is observed (0.95), while F1 for most relevant items has a considerable drop. Detailed results are available in Fig. 7.

C.2 Modeling With Individual Bias

In order to assess the impact of individual bias, we introduce annotator-specific variations. These are assigned to each annotator independently, by selecting a random bias among *functionalities*. Such bias causes the group-level label to be changed with random probability $p \in [0.1, 0.4]$. Constraining individual-level bias prevents the emergence of spurious data groupings, aligning with our objective to model group behaviors rather than individual variations. A comparison of results is presented in Table 4. For all metrics, we observe scores close to the case without individual bias. Such observations

	$\Delta_{-1}L$	$\Delta_{-2}L$	F1	Accuracy
w/o bias	.13	.17	.89 $\pm$.18	.93 $\pm$.16
w/ bias	.10	.15	.82 $\pm$.16	.95 $\pm$.09

Table 4: Comparison of validation results with and without individual bias. Results are expressed in terms of *calibration gain* (ΔL); most relevant items (F1 score); and grouping Accuracy.

suggest that FLOCKE can learn group behaviors from annotations reflecting also individual biases. Detailed results are shown in Fig. 7.

C.3 Different Language Models

We validate FLOCKE using multiple language models. The results (Table 5) validate the method’s capabilities across different embedding spaces, despite some variations in specific outcomes.

	$\Delta_{-1}L$	$\Delta_{-2}L$	F1	Accuracy
MultiBERT	.13	.17	.89 $\pm$.18	.93 $\pm$.16
RoBERTa-base	.07	.13	.74 $\pm$.18	.72 $\pm$.24
RoBERTa-large	.07	.11	.73 $\pm$.18	.74 $\pm$.23
SBERT	.11	.16	.88 $\pm$.16	.92 $\pm$.11

Table 5: Validation results with different embeddings.