DeMeVa at LeWiDi-2025: Modeling Perspectives with In-Context Learning and Label Distribution Learning

Daniil Ignatev, Nan Li, Hugh Mee Wong, Anh Dang, Shane Kaszefski Yaschuk
Utrecht University, Utrecht, The Netherlands
{d.ignatev, n.li, h.m.wong, t.t.a.dang, s.p.kaszefskiyaschuk}@uu.nl

Abstract

This system paper presents the DeMeVa team's approaches to the third edition of the *Learning with Disagreements* shared task (LeWiDi 2025; Leonardelli et al., 2025). We explore two directions: in-context learning (ICL) with large language models, where we compare example sampling strategies; and label distribution learning (LDL) methods with RoBERTa (Liu et al., 2019b), where we evaluate several fine-tuning methods. Our contributions are twofold: (1) we show that ICL can effectively predict annotator-specific annotations (*perspectivist* annotations), and that aggregating these predictions into soft labels yields competitive performance; and (2) we argue that LDL methods are promising for soft label predictions and merit further exploration by the perspectivist community.

1 Introduction

In natural language processing (NLP), annotations are often treated as a gold standard, implying a single, unambiguous truth. However, for tasks that involve, among other things, cultural norms or subjectivity, human judgments can vary substantially, often reflecting diverse annotator backgrounds or personal perspectives (Plank, 2022; Cabitza et al., 2023). Customary approaches that aggregate these diverging annotations with techniques like majority voting disregard the potential validity of pluralistic interpretations, which may lead to the loss of valuable information about both the data instances and the people who annotated them. The *Learning with Disagreements* (LeWiDi) shared task shifts the focus to learning from unaggregated crowd labels, whether through learning from soft labels or through aligning models with specific annotators' viewpoints (i.e., *perspectivist* training).

The DeMeVa team ranks 2nd overall on the leaderboard of the LeWiDi 3rd Edition shared task (LeWiDi 2025; Leonardelli et al., 2025). In this system paper, we describe the contributions of the DeMeVa team and discuss both our highest-scoring method and the other approaches that did not make it onto the leaderboard. We hope that our interpretation of these results will offer insights into learning with disagreement in NLP.

We obtained our score on the leaderboard by employing in-context learning (ICL) for perspectivist modeling. ICL refers to the ability of pre-trained large language models (LLMs) to perform NLP tasks without task-specific training; in ICL, these models are instead conditioned on input-output examples ("demonstrations") provided in the prompt (Brown et al., 2020). Recent studies have demonstrated ICL's success on a wide range of tasks (see e.g. Dong et al., 2024). However, they have also shown that ICL is sensitive to the choice, order, and format of demonstrations. We explore how and to what extent ICL can be leveraged to steer LLMs toward the annotation patterns of individual annotators in natural language understanding.

In parallel with perspectivist ICL, our team also pursued alternative directions aimed at modeling label distributions. In this context, we drew on existing research from both NLP and other communities. Specifically, we refer to studies in *label distribution learning* (LDL), a research vein that focuses on modeling probability distributions over full label spaces and which has its roots in the broader machine learning community. We note that some of the insights from LDL have not yet fully found their way into NLP-specific research. In our experiments, we build on such works by using two LDL-specific fine-tuning methods, neither of which has been widely applied in NLP: *ordinal label distribution learning* (Wen et al., 2023) and predicting population-level label distributions via clustering (Liu et al., 2019a).

The structure of this paper is as follows. In Section 2, we briefly reintroduce the datasets and subtasks of the LeWiDi shared task. Next, we describe our ICL approaches in Section 3 and our LDL-

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 171–181

Dataset	Task	#E (train/dev/test)	#Ann/E	#Ann
CSC (Jang and Frassinelli, 2024)	Sarcasm detection	5628/704/704	4+	840
MP (Casola et al., 2024)	Irony detection	12017/3005/3756	5+	506
Par (as yet unpublished)	Paraphrase detection	400/50/50	4	4
VariErrNLI (Weber-Genzel et al., 2024)	NLI	388/50/50	4	4

Table 1: Overview of datasets used in LeWiDi 2025. **E** denotes entries, **Ann** denotes annotators.

related fine-tuning strategies in Section 4. Finally, we make our concluding remarks in Section 5.

2 Datasets, tasks, and evaluation metrics

In this section, we discuss the datasets and evaluation metrics of the LeWiDi 2025 shared task.

2.1 Datasets

The LeWiDi 2025 shared task includes 4 datasets covering various aspects of natural language understanding (see Table 1 for an overview).

CSC The *Conversational Sarcasm Corpus* (CSC; Jang and Frassinelli, 2024) is a richly annotated sarcasm dataset containing 7,040 context-response pairs. For each of these pairs, the authors provided self-ratings on a 6-point Likert scale, and third-party annotators (360 in total, with 6 per author in Part 1 and 4 per response in Part 2) rated the level of sarcasm in the responses on the same scale.

MP The *MultiPICo Dataset* (MP; Casola et al., 2024) is a multilingual, socio-demographically grounded dataset of irony on social media, comprising 18,778 post-reply pairs from Reddit and Twitter across 9 languages and 25 linguistic varieties. Each received a mean of 5.02 binary irony labels from a pool of 506 crowd annotators balanced by gender and nationality.

Par The *Paraphrase Detection Dataset* (Par; as of yet unpublished) contains 500 sentence pairs from the *Quora Question Pairs* dataset, each annotated by 4 expert annotators on a Likert scale ranging from -5 to +5 based on paraphrase quality. Annotators were asked to provide short explanations justifying their scores as well.

VariErrNLI The *VariErrNLI* Dataset (Weber-Genzel et al., 2024) is designed to disentangle genuine human label variation from annotation errors in natural language inference (NLI). It features a two-round annotation protocol applied to 500 multi-genre NLI (MNLI; Williams et al., 2018; Nie et al., 2020) items, resulting in 1,933 label-explanation pairs in the first round and 7,732 validity judgments in the second round. The dataset serves both as a benchmark for Automatic Error Detection methods and a resource to improve dataset trustworthiness. It also includes explanations for each annotation.

2.2 Tasks and evaluation metrics

LeWiDi 2025 introduces two tasks for the two established main approaches to unaggregated data: 1) Task A—*soft label modeling*, where systems generate probability distributions over all classes for each item; and 2) Task B—*perspectivist modeling*, where systems predict individual annotators' labels for specific items. At the same time, within each of these two tasks, the evaluation metrics vary depending on the structure of the concrete dataset they are paired with: e.g., Par and CSC, which both include Likert-scale values, require a different metric suite compared to datasets with unranked labels.

For Task A, the MP and VariErrNLI datasets make use of the Manhattan distance as the evaluation metric. The Manhattan distance measures the sum of absolute differences between the predicted and the target distributions. For VariErrNLI, this is extended to a *Multi-label Average Manhattan Distance* (MAMD), averaging the Manhattan distances across multiple labels. Performance on the Par and CSC datasets is assessed with the Wasserstein distance, which measures the minimum cost to transform one distribution into another.

Regarding Task B, MP and VariErrNLI are paired with the *error rate* (ER) and *multi-label error rate* (MER), respectively. ER measures the proportion of incorrectly matched values between predicted and target label vectors, while MER averages the error rates across multiple labels. For Par and CSC, the *average normalized absolute distance* (ANAD) is used, which normalizes the average absolute difference between Likert scale values based on the range of the scale. In all cases, a lower score indicates better performance, with a score of 0 indicating a perfect match.

3 In-context learning

Recent work has explored in-context learning for steering language models toward diverse human label distributions, primarily focusing on persona-based prediction for tasks like toxicity and hate speech detection (Sorensen et al., 2025; Radlinski et al., 2022; Ramos et al., 2024). In that vein, many studies focus solely on the effect of steering models with persona descriptions (Hu and Collier, 2024; Kambhatla et al., 2025; Sun et al., 2025); in the meantime, prompts that also incorporate annotations have been shown to elicit better predictions (Meister et al., 2025). While these inquiries are mostly based on more widely used datasets, LeWiDi 2025 presents new challenges on tasks that have received limited attention so far in the domain of perspectivist NLP such as paraphrase evaluation and sarcasm detection.

We explore different ICL strategies on these novel datasets to advance perspective-aware modeling, leveraging state-of-the-art generative models: OpenAI's GPT-4o (Achiam et al., 2023), Claude Haiku 3.5 (Anthropic, 2024), and Llama 3.1 70B-Instruct (Grattafiori et al., 2024). However, we do not explore persona-based steering as the LeWiDi 2025 datasets contain relatively few sociodemographic variables, making sociodemographic prompting infeasible.

3.1 System pipeline

To accomplish both tasks of LeWiDi 2025, we propose a two-step pipeline (Figure 1). First, we use ICL to prompt LLMs to predict individual annotators' labels based on their previous responses (Task B). We then use these predictions to calculate the final soft label (Task A).

The two key components of ICL are *demonstration selection* and *prompt engineering*. Our main focus is on finding the most appropriate example sampling method (demonstration selection). As for the prompt engineering component, we use a simple template adapted from Dutta et al. (2025) that is applicable to all datasets in the shared task (see Figure 2 for the prompt template and Appendix A for a filled example for the CSC dataset). The template is designed to be flexible enough to accommodate different tasks and input formats while also being straightforward enough for the LLM to leverage. For every experiment, we set the temperature to `0.0` to enforce greedy decoding and yield the most probable sequence with minimal randomness.

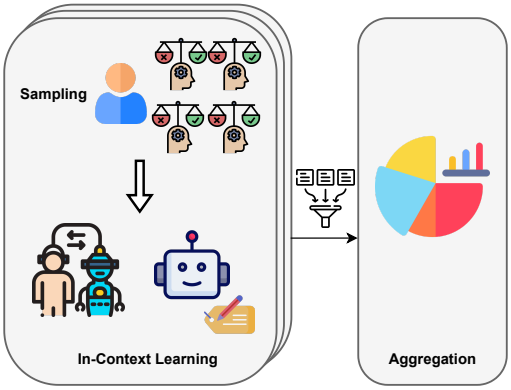

Figure 1: Our two-step pipeline to solve both tasks, based on ICL. In the first step (Task B), we sample examples from an annotator's past annotations and prompt the LLM to model annotator-specific behavior and predict labels for test inputs. In the second step, we aggregate these predictions into soft labels (Task A).

Prompt Template

```
[INST] You are an expert in guessing my
response against a {TASK_NAME} task.

Your task is to analyze and predict my
response to {INPUT_FORMAT} between <<<
and >>>, and label it with {RESPONSE
_FORMAT} {LABEL_EXPLANATION}.

Below are some of my previous responses.
You should learn my response behavior
from them and then make the prediction.

{EXAMPLES}
[/INST]

>>>
{INPUT}
>>>
```

Figure 2: Our ICL prompt template. The template supports varied tasks and input formats without sacrificing clarity.

3.2 Example selection strategies

ICL is sensitive to how demonstrations are sampled and supplied to the model. We therefore compare two strategies for example selection: *similarity-based* and *stratified label-based* sampling. Additionally, we examine whether explanations available in the Par and VariErrNLI datasets can improve model personalization when included in prompts. This test builds on the work started by Ye and Durrett (2022) and Jiang et al. (2023), who

stress the ambiguous role of explanations in NLI labeling.

The standard approach of retrieving semantically similar examples faces challenges with respect to perspectivist learning. BERT-based cosine similarity primarily ensures lexical and topical proximity (Kaster et al., 2021), but perspectivist tasks may require more nuanced selection. First, as Jiang and Marneffe (2022) show, annotators in NLU tasks rely on specific linguistic heuristics rather than topical similarity; hence, similarity with respect to these heuristics would offer a better selection criterion. Second, annotator-specific subsets can be arbitrarily small, which means they may lack enough similar examples for meaningful retrieval. Our two sampling strategies are as follows.

Similarity-based sampling For a test input q (the current query) and annotator a, let $\mathcal{D}_a$ denote the set of training examples annotated by a. Let $\mathbf{h}(x) \in \mathbb{R}^d$ be the sentence embedding of x produced by Sentence-Transformers (Reimers and Gurevych, 2020). We measure the relevance using the cosine similarity: $\mathrm{s}(q, x) = \cos\big(\mathbf{h}(q), \mathbf{h}(x)\big)$. We select k demonstrations starting with $S = \varnothing$ and at each step, we add to this set the element

$$x^* = \underset{x \in \mathcal{D}_a \setminus S}{\arg\max}\ \lambda\ \mathrm{s}(q, x) - (1-\lambda) \max_{x' \in S} \mathrm{s}(x, x').$$

and update $S \leftarrow S \cup \{x^*\}$ until $|S| = k$. We set $\lambda = 0.7$ to reduce redundancy among selected shots via the Maximal Marginal Relevance (MMR) method.

Stratified label-based sampling For each annotator a, let $\mathcal{D}_a$ denote the training set, $\mathcal{Y}_a$ the full set of their annotations, and $y_a(x) \in \mathcal{Y}_a$ the label assigned by this annotator for data sample x. We first drop labels that occur less than two times to ensure stratification. Let $L = \max\{|\mathcal{Y}_a|, k\}$. If $|\mathcal{D}_a| \leq L$ or only one label remains, we sample up to k examples uniformly from $\mathcal{D}_a$. Otherwise, we construct a stratified subsample $S' \subset \mathcal{D}_a$ that approximately preserves the empirical label proportions over $y_a(x)$. We do this using scikit-learn (Pedregosa et al., 2011), and we then draw k examples uniformly without replacement from S'.

We hypothesize that label-based sampling yields more representative examples by exposing models to diverse annotation patterns, which can be particularly effective for nuanced label scales (such as those found in CSC and Par) compared to binary tasks. This approach increases the likelihood that relevant linguistic heuristics appear in demonstrations, helping models learn annotator-specific decision patterns. We set the number of demonstrations to $k = 10$.

3.3 Model performance

We report the experiment results in Table 2. While the performance differences between ICL approaches are relatively subtle, they mostly yield substantial improvements over the baseline methods across the datasets and tasks.[1]

Similarity-based sampling performs best on MP, whereas label-based sampling tends to improve (lower) Task A distances on the other datasets without reducing the error rate. For MP, both error rate and distance are lower when using similarity-based sampling. This is to be expected: with binary labels, stratified label-based sampling is practically equivalent to random sampling. On the other three datasets, label-based sampling often results in improvements on Task A, while the error rate often changes insignificantly or even increases compared to stratified label-based sampling. Our explanation for this is that the metrics for Task A show more sensitivity toward numeric values of predictions, and label-based sampling offers more control of said numeric values since the model limits its outputs to within the provided label range. At the same time, since the error rate is not significantly influenced, our assumption that the sampled examples are more representative of this method does not appear to hold.

For Par and VariErrNLI, the results show that the inclusion of explanations further enhances performance; remarkably, this trend is more pronounced in Task A metrics compared to Task B metrics, as the error rate remains roughly in the same range for both settings. However, the calibration effect of label-based sampling is more notable (especially for GPT-4o), showing that it is amplified by reasoning examples. The fact that explanations improve performance in this regard may complement the results of Ni et al. (2025), who find that CoT-prompting helps steer RLHF models toward human perspectives. While explanations only contain one reasoning step, they still can be regarded as being analogous to more complex reasoning examples.

[1] MP stands out as an exception to that. We explain the poor performance of Llama and Haiku on MP by the fact that they do not adequately support several of the languages present in MP.

	Task A				Task B			
	CSC	MP	Par	VariErrNLI	CSC	MP	Par	VariErrNLI
baseline_random	1.549	0.689	3.35	1.0	0.355	0.5	0.38	0.5
baseline_most_frequent	1.169	0.518	3.23	0.59	0.238	0.316	0.36	0.34
GPT-4o +*sim*	0.84	0.466	1.17	0.46	0.175	0.294	0.13	0.26
GPT-4o +*strat*	**0.792**	**0.469**	1.25	0.44	**0.172**	**0.3**	0.14	0.25
Haiku-3.5 +*sim*	1.005	0.657	1.58	0.43	0.205	0.375	0.15	0.26
Haiku-3.5 +*strat*	0.95	0.684	1.47	0.42	0.201	0.392	0.16	0.27
Llama-3.1-70B-Inst +*sim*	1.192	0.691	1.41	0.44	0.226	0.392	0.14	0.24
Llama-3.1-70B-Inst +*strat*	1.157	0.706	1.38	0.36	0.227	0.399	0.15	0.22
+ *Explanation*:								
GPT-4o +*sim*	–	–	1.17	0.43	–	–	0.12	0.24
GPT-4o +*strat*	–	–	**1.12**	**0.38**	–	–	**0.13**	**0.23**
Haiku-3.5 +*sim*	–	–	1.36	0.44	–	–	0.13	0.24
Haiku-3.5 +*strat*	–	–	1.35	0.45	–	–	0.15	0.25
Llama-3.1-70B-Inst +*sim*	–	–	1.35	0.46	–	–	0.14	0.25
Llama-3.1-70B-Inst +*strat*	–	–	1.39	0.44	–	–	0.14	0.25

Table 2: Results of ICL Strategies on LeWiDi 2025. +*sim* denotes similarity-based example sampling, and +*strat* denotes stratified label-based sampling. We additionally experiment with including annotator explanations in the Par and VariErrNLI datasets (+*Explanation*). The results submitted to the leaderboard are shown in bold.

3.4 Discussion

To further validate the results of ICL, we examine its predictions on each development set in more detail. While the development sets of Par and VariErrNLI both comprise only 50 examples, those of MP and CSC consist of hundreds of items each, making it more challenging to inspect them thoroughly. We therefore sample a smaller subset of items from both datasets: for CSC, we randomly select 50 items, whereas for MP, we extract 50 random items for each included language (totaling 450 items). Although this strategy makes the analysis more feasible, it effectively prevents us from comparing per-annotator label distributions on CSC, as the down-sampled sets only include a few examples per worker. Nevertheless, we can still identify the strengths and weaknesses of our ICL methods based on specific data items.

One notable tendency is that models often predict unanimous agreement on instances that appear straightforward on the surface, but which are actually annotated differently. We illustrate this with an example from `MP-dev-1597` in Figure 3. While the reply is directly licensed by the first utterance, that utterance does not give an immediate and obvious reason for an ironic reply. However, more than half of the annotators labeled this example as ironic. Similar cases can also be identified in the three other datasets, as annotators often demonstrate vastly different annotation behaviors. These examples possibly show that complete pluralistic alignment of language models may be impossible to fully achieve (at least in the linguistic domain), as the model adhering to common sense in all examples appears to be more important in this context when compared to adhering to the plurality of views.

MP–dev–1597 (snippet)

[Post]: We once used coins such as Annas, paise, even half annas! and one could survive a day! The rupee used to be made of silver which would be a day's salary back then.

[Reply]: How old are you?

Figure 3: A sample from the MP development set. The majority of the annotators marked this example as ironic.

At the same time, we note that the tested models are generally successful in mimicking specific annotators' labeling strategies. This is best illustrated in the VariErrNLI and Par datasets, since the annotators' motivations are directly available for analysis. For example, in the Par dataset, annotator `Ann3` uses label `0` considerably more often

than their peers, consistently labeling most non-contradicting examples with 0 rather than negative values even when they are non-relevant. This reasoning is also reflected in Ann3's explanations. The models, particularly when combined with label-based sampling, tend to imitate this peculiarity while also never predicting 0 for annotators Ann1 and Ann2 (who rarely use it). Likewise, the predictions also reflect less subtle differences, like the annotators' inclination toward positive or negative scale values (for Par) and entailment or neutral labels (for VariErrNLI). For example, Ann3's preference for positive values is also discernible in the predicted labels. In this respect, it can be argued that ICL can be successful when used for perspectivist modeling of individual perspectives.

4 Fine-tuning approaches

In this section, we discuss the various fine-tuning approaches we have explored for Task A. While the overall performance of these approaches was ranked lower on the leaderboard than the in-context learning methods from Section 3—in part because we merely tackled a subset of Task A—we believe that these fine-tuning methods are still a valuable contribution to understanding how we can learn from disagreements. All fine-tuning experiments were done using the base RoBERTa (Liu et al., 2019b) model.

4.1 Approach 1: Cumulative distances for Likert scales

In the machine learning and computer vision communities, Geng and Ji (2013) introduced *label distribution learning* (LDL) as an alternative to *single-label* and *multi-label learning* (MLL). While MLL allows data instances to be assigned to multiple classes, LDL aims to solve the ambiguity problem (i.e., instances potentially belonging to several classes) by predicting *how much* each label describes an instance. In other words, just like the soft evaluation approaches developed in perspectivist NLP communities, LDL predicts a probability distribution over the set of available labels. Wen et al. (2023) remark that LDL algorithms generally fail to accurately predict distributions for tasks where the labels are inherently ordered, such as age estimation. They propose the *ordinal label distribution learning* (OLDL) paradigm and introduce evaluation metrics which take the ordinality of labels into account.

The Par and CSC datasets both contain annotations based on a Likert scale. These scales are ordered: higher ranks represent a higher degree of the measured concept. In the first part of our fine-tuning efforts, we experiment with using two evaluation metrics proposed by Wen et al. (2023) as loss functions when fine-tuning RoBERTa: *cumulative Jensen–Shannon divergence* and *cumulative absolute distance*. During experimentation, we freeze all but the last six layers.

Let CDF_P and CDF_Q be the *cumulative distribution functions* of distributions P and Q, respectively. We define the two loss functions as follows.

Cumulative Jensen–Shannon The *cumulative Jensen–Shannon* (CJS) divergence between P and Q is defined as:

$$\text{CJS}(P, Q) = \sum_{n=1}^{C} D_{js}(\text{CDF}_P(n)||\text{CDF}_Q(n)), \quad (1)$$

where $D_{js}(X||Y)$ denotes the Jensen–Shannon divergence between distributions X and Y.

Cumulative Absolute Distance The *cumulative absolute distance* (CAD) is defined as:

$$\text{CAD}(P, Q) = \sum_{n=1}^{C} |\text{CDF}_P(n) - \text{CDF}_Q(n)|. \quad (2)$$

We make the following observation: the evaluation metric used for both Par and CSC in Task A is the Wasserstein distance (WSD). Intuitively, the WSD reflects how much mass has to be moved, and how far, to transform one distribution into another. In the discrete 1-dimensional scenario, as is the case for Likert labels, the Wasserstein distance reduces to:

$$W_1(P, Q) = \sum_{n=1}^{C} |\text{CDF}_P(n) - \text{CDF}_Q(n)|, \quad (3)$$

which is the same as CAD (Equation 2). Indeed, Wen et al. (2023) proposed CAD as an adaptation of the Mallows distance, which is also known as the Wasserstein-2 distance.

Results We report our results in Table 3. We concluded that straightforwardly using one of the given formulas as a loss function for fine-tuning RoBERTa would not be sufficient. The reason for this is that although CAD is equal to the Wasserstein distance in 1D, minimizing CAD loss during

	CSC	Par
CJS	0.831 ± 0.01	1.677 ± 0.10
CJS+MAE	0.813 ± 0.00	1.694 ± 0.03
CAD	0.800 ± 0.01	1.590 ± 0.12
CAD+MAE	0.797 ± 0.01	1.558 ± 0.10

Table 3: Fine-tuning results (Wasserstein distance) for the CAD and CJS loss functions on the test sets. All results are averaged across three random seeds. Here, `CJS+MAE` is the average of predictions from CJS and MAE; `CAD+MAE` is defined in a similar fashion.

training might not guarantee good generalization: the cumulative nature of CAD/W1 could allow small prediction errors to be diffused across subsequent labels. As a result, we hypothesized that it might not strongly penalize localized prediction errors if the overall CDF stays close, potentially leading to blurry or smeared distributions. For this reason, we also experimented with combining CJS/CAD with the mean absolute error (MAE), encouraging the mode of the predicted distribution to align better with the "ground truth" distribution while still respecting the ordinal structure of the data. However, Table 3 suggests that this does not make a difference. For the CSC dataset in particular, we find that CAD/CAD+MAE can yield scores that are competitive with in-context learning (0.792 in Table 2).

4.2 Approach 2: Population-level label distributions

Liu et al. (2019a) introduce a strategy for learning label distributions designed to significantly reduce the total number of human labels required for each data item. They suggest that even if humans can interpret a data item in many ways, their annotations tend to reduce these interpretations to a limited number of distinct "ground truth" label distributions. Therefore, the annotations for any given item are seen as a sample drawn from one of these distinct underlying distributions. They found that this technique works well for datasets with 5-10 annotations per data item. Given that the Par dataset only has four annotations per sentence pair, we used this approach on this dataset alone. Liu et al. (2019a) also hypothesized that semantically similar items tend to have similar label distributions. For this reason, they proposed to (1) cluster the data into semantically similar groups using unsupervised learning, (2) aggregate the annotations of the clusters to create a single soft label for each cluster, and (3) use supervised learning to learn to predict the unified label distributions.

When dealing with the Par dataset, we assume that some sentence pairs are inherently more difficult to annotate than others. The annotations for these pairs may be more spread out and sparse as a result, while those for other samples may be more unified. We adopt the clustering and two-stage training methodology proposed by Liu et al. (2019a). However, instead of using a single soft label distribution for all items in a cluster, we trained the classifiers on the original soft labels and then included clustering information to push the predicted soft labels to fall within a certain range.

Model Specification For the clustering, we opted for k-means clustering with a maximum of 5 clusters. We clustered the sentence pairs into groups with similar soft label distributions and then used their cluster numbers to guide the training process. We then fine-tuned RoBERTa to predict the soft labels. To leverage the resulting clusters, we trained multitask classifiers with 2 prediction heads.

Soft Label Head The soft label head is a simple feedforward layer outputting logits over 11 annotation scores from -5 to 5. In this case, we used cross-entropy loss as the loss function.

Cluster Classification Head To classify the clusters, we used a separate feedforward layer for predicting the logits for n discrete cluster IDs. The head is trained to predict the corresponding cluster assignment of each example. For the loss function, we tried several options, namely KL divergence, Wasserstein distance, and all loss functions described in Section 4.1.

The overall training loss is the sum of the soft label loss and the weighted cluster classification loss:

$$L_{\text{total}} = L_{\text{soft}} + \alpha \cdot L_{\text{cluster}}. \quad (4)$$

In this formula, L_{total} represents the total training loss, L_{soft} is the loss for soft label prediction, and L_{cluster} is the loss for cluster prediction. α is a tunable parameter that varies the overall influence of L_{cluster}.

Results Our best score with this approach is a Wasserstein distance of 1.66 for the Par dataset. We achieved this by classifying the dataset into 3 clusters. While the performance is above the

baseline by a notable margin, this method still underperformed compared to the other fine-tuning method described in Section 4.1 and the in-context learning method described in Section 3.

4.3 Discussion

For the loss functions, it comes as no surprise that CAD yields better results than CJS on the test set, given that the evaluation metric is CAD/WSD. Table 3 suggests that a standard fine-tuning setup with this loss function might be enough to yield competitive scores on the CSC dataset.

Note that the Par dataset in particular had a relatively small number of annotators. Given that only four annotators were annotating on an 11-point Likert scale, sparse distributions are inevitable. We find that our methods are not able to handle this sparsity well enough to yield scores comparable to those for CSC. On a related note, we would like to make one additional observation: when working with sparse annotations, it is highly important to consider how the models are evaluated. When annotations are sparse, the "ground truth" distributions may only be a noisy, undersampled proxy of the underlying human opinion distribution. As well, relying on raw empirical frequencies can exaggerate annotation noise, and evaluating against them with strict distance metrics such as the Wasserstein distance may unfairly penalize models that produce smoother (and arguably more plausible) distributions. However, as it was not possible to apply smoothing to the unseen test set, we found that models optimized for smoother distributions will generally perform poorly according to the LeWiDi scoring mechanism. All results reported in this section were obtained without additional smoothing.

As with many other domains, it appears that the NLP community can take inspiration from the computer vision and machine learning communities (and vice versa). Indeed, the perspectivist approaches in NLP appear to have emerged independently from label distribution learning in CV/ML (and also with different objectives; note, for example, the fact that CAD and 1D WSD are the same), yet both grapple with similar challenges. We argue that perspectivist NLP could benefit from the probabilistic and distributional modeling techniques developed in these other communities.

5 Conclusion

In this paper, we introduced the two main approaches taken by the DeMeVa team for the LeWiDi 2025 shared task. Our comparison of ICL approaches on perspectivist modeling, while not yielding fully conclusive results, suggested that sampling examples based on labels can help generative models calibrate their predictions—especially for numeric outputs like Likert scale values. Models calibrated in this way can trace and mimic annotators' behavior down to more specific, granular details. However, their reliance on common sense (possibly induced by RLHF) may hinder their ability to recognize plurality when it is not overtly expressed.

The second contribution of this work is a call for the perspectivist NLP community to look outward. In particular, we can learn from how machine learning communities have addressed uncertainty and label distribution learning. While perspectivist NLP rightly centers the diversity of annotator perspectives, it can benefit from established techniques such as probabilistic modeling and smoothing methods that account for annotation noise and limited sample sizes. We have merely scratched the surface here by borrowing simple loss functions and a clustering method from LDL, but we believe that engaging with other fields can be beneficial to the perspectivist community as a whole.

Ethical Considerations

In this work, we make use of personalized annotations, which, *inter alia*, include sociodemographic variables related to the annotators. However, anonymization by their respective original authors ensures that this data cannot be used in a manner that is harmful to individuals.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This work is funded by the Dutch Research Council (NWO) through the AiNed Fellowship Grant NGF.1607.22.002, *Dealing with Meaning Variation in NLP*.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Alt-

man, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, and 260 others. 2023. Gpt-4 technical report.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

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.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128.

Sujan Dutta, Deepak Pandita, Tharindu Cyril Weerasooriya, Marcos Zampieri, Christopher M Homan, and Ashiqur R KhudaBukhsh. 2025. Annotator reliability through in-context learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 14230–14237.

Xin Geng and Rongzi Ji. 2013. Label distribution learning. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 377–383.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Hyewon Jang and Diego Frassinelli. 2024. Generalizable sarcasm detection is just around the corner, of course! 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 4238–4249, Mexico City, Mexico. 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.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine de Marneffe. 2023. Ecologically valid explanations for label variation in NLI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10622–10633, Singapore. Association for Computational Linguistics.

Gauri Kambhatla, Sanjana Gautam, Angela Zhang, Alex Liu, Ravi Srinivasan, Junyi Jessy Li, and Matthew Lease. 2025. Beyond sociodemographic prompting: Using supervision to align llms with human response distributions. *arXiv preprint arXiv:2507.00439*.

Marvin Kaster, Wei Zhao, and Steffen Eger. 2021. Global explainability of bert-based evaluation metrics by disentangling along linguistic factors. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8912–8925.

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 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.

Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher Homan. 2019a. Learning to predict population-level label distributions. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1111–1120, New York, NY, USA. Association for Computing Machinery.

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

Nicole Meister, Carlos Guestrin, and Tatsunori B 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.

Jingwei Ni, Yu Fan, Vilém Zouhar, Donya Rooein, Alexander Hoyle, Mrinmaya Sachan, Markus Leippold, Dirk Hovy, and Elliott Ash. 2025. Can reasoning help large language models capture human annotator disagreement? *Preprint*, arXiv:2506.19467.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? *arXiv preprint arXiv:2010.03532*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

Filip Radlinski, Krisztian Balog, Fernando Diaz, Lucas Dixon, and Ben Wedin. 2022. On natural language user profiles for transparent and scrutable recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2863–2874, New York, NY, USA. Association for Computing Machinery.

Jerome Ramos, Hossein A. Rahmani, Xi Wang, Xiao Fu, and Aldo Lipani. 2024. Transparent and scrutable recommendations using natural language user profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13971–13984, Bangkok, Thailand. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. 2025. Value profiles for encoding human variation. *arXiv preprint arXiv:2503.15484*.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 845–854, Albuquerque, New Mexico. Association for Computational Linguistics.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Changsong Wen, Xin Zhang, Xingxu Yao, and Jufeng Yang. 2023. Ordinal label distribution learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23424–23434.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392.

A Example of an ICL prompt

CSC–test–2143–Ann743

```
[INST] You are an expert in guessing my
response against a sarcasm detection task.


Your task is to analyze and predict my
response to a pair of context and
response between <<< and >>>, and label
it with an integer from 1 to 6 where 1
means not sarcastic at all and 6 means
completely sarcastic.

Below are some of my previous responses.
You should learn my response behavior
from them and then make the prediction.

Example 0:
[Context]: Steve is a fan of Bulgarian
folk music. Every week, he finds a
different song and plays it on his phone
and says, "I finally found one you'll
like! This one is really good. Come on!"
[Response]: Bulgarian folk music is for
old people Steve, didn't you say you
wanted to be young and cool?
[Label]: 2

Example 1:
[Context]: You are watching TV with Steve.
 Whenever you set the volume to an odd
number, Steve takes the remote control
away from you and sets the volume to an
even number.
[Response]: My mistake, I never useally
do that.
[Label]: 2

...

Example 9:
[Context]: Steve and you are hanging out
tonight. He shows up wearing a red tank
top, green shorts, and yellow sneakers.
[Response]: Did you go to a yard sale or
something?
[Label]: 5
[/INST]

>>>
[Context]: You walk into the room and
Steve is wearing his shoes on his hands.
When you see him, he says "look at me! I'
m Mr. Shoehand!"
[Response]: Are you 5 or 50?
[Label]:
>>>
```

Figure 4: ICL prompt for entry `CSC-test-2143` and Annotator `Ann743` (excerpt). The in-context examples are selected from `Ann743`'s annotations in the train set, following the stratified label-based sampling method.