# NLP-ResTeam at LeWiDi-2025:Performance Shifts in Perspective Aware Models based on Evaluation Metrics

**Olufunke O. Sarumi[1], Charles Welch[2], Daniel Braun[1]**
[1]University of Marburg, [2]McMaster University
{sarumio,daniel.braun}@uni-marburg.de[1], cwelch@mcmaster.ca[2]

## Abstract

Recent works in Natural Language Processing have focused on developing methods to model annotator perspectives within subjective datasets, aiming to capture opinion diversity. This has led to the development of various approaches that learn from disaggregated labels, leading to the question of what factors most influence the performance of these models. While dataset characteristics are a critical factor, the choice of evaluation metric is equally crucial, especially given the fluid and evolving concept of perspectivism. A model considered state-of-the-art under one evaluation scheme may not maintain its top-tier status when assessed with a different set of metrics, highlighting a potential challenge between model performance and the evaluation framework. This paper presents a performance analysis of annotator modeling approaches using the evaluation metrics of the 2025 Learning With Disagreement (LeWiDi) shared task and additional metrics. We evaluate five annotator-aware models under the same configurations. Our findings demonstrate a significant metric-induced shift in model rankings. Across four datasets, no single annotator modeling approach consistently outperformed others using a single metric, revealing that the "best" model is highly dependent on the chosen evaluation metric. This study systematically shows that evaluation metrics are not agnostic in the context of perspectivist model assessment.

## 1 Introduction

The primary aim of perspectivism in (NLP) is to preserve and leverage the diverse, subjective decisions of individual annotators, both in the modeling process and in the subsequent evaluation of those models (Frenda et al., 2024; Cabitza et al., 2023). Given the variety of annotator representation methods, a key challenge lies in how to effectively incorporate annotator-specific information during model training to capture these unique perspectives (Mostafazadeh Davani et al., 2022). The efficacy of such annotator modeling techniques is influenced by several critical factors. A foundational element is the annotation paradigm used to create the dataset (Rottger et al., 2022). Furthermore, the performance is heavily dependent on the dataset's statistical properties, including the number of training instances required to reliably model an annotator, the volume of annotations per annotator, the degree of inter-annotator agreement (IAA), and the number of annotations per instance. Sarumi et al. (2024) showed that the number of contributions from an annotator and the IAA are particularly crucial statistics to consider.

While existing approaches capture annotator diversity to varying extents, their evaluation has predominantly relied on conventional metrics like the F1-score (Uma et al., 2021; Plepi et al., 2022; Sullivan et al., 2023; Welch et al., 2022; Sarumi et al., 2025a) and in some cases Cross Entropy, especially for soft label prediction (Leonardelli et al., 2023). It has been argued, however, that such metrics are insufficient as they often collapse multiple valid perspectives into a single ground truth, failing to truly reflect the goals of a perspectivist evaluation (Rizzi et al., 2024). As part of our submission to LeWiDi 2025, we present a comparative study of different annotator modeling approaches. We analyze how their performance shifts when assessed using a range of evaluation metrics, including those provided by the organizers. Our aim is to advance a more nuanced view within the perspectivist framework. We hypothesize that the performance of a given modeling approach is not absolute but is contingent upon the evaluation metric used. A model that is best performing under one metric may not perform as well under another, especially when applied to datasets with different underlying statistical properties and task natures. To investigate this, we implemented five distinct modeling approaches



*Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025)*, pages 219–227

and evaluated them on the perspectivist subtask (B) using additional evaluation metrics.

## 2 Background and Summary

One of the primary challenges of the 2025 edition of the LeWiDi shared task is the two concurrent tasks designed to model and evaluate variations in annotations (Leonardelli et al., 2025). Task A, the soft label approach, focused on predicting the probability distribution of labels for each instance and Task B, the perspectivist approach, focused on predicting the individual label assigned by each annotator. The organizers introduced four new datasets and adopted a tailored evaluation framework for each, rather than relying on a single unifying metric.

The Conversational Sarcasm Corpus (CSC) (Jang and Frassinelli, 2024), consists of context–response pairs rated for sarcasm on a Likert scale from 1 to 6, with soft label evaluation based on Wasserstein distance and perspectivist evaluation based on Mean Absolute Distance (MAD). The MultiPico (MP) dataset (Casola et al., 2024) is a crowdsourced multilingual irony detection resource containing post–reply pairs from Twitter and Reddit, annotated with binary labels across 11 languages. The datasets also contained annotator metadata such as gender, age, nationality, and student or employment status. Evaluation for the soft label task used Manhattan distance, while the perspectivist task used the error rate. The Paraphrase Detection (PAR) dataset (MaiNLP Lab, 2025) contains question pairs collected from Quora and annotated on a Likert scale from –5 to +5, with each annotator providing a brief explanation for their score, as in CSC, evaluation for the soft task used Wasserstein distance and for the perspectivist task used MAD. Finally, the VariErrNLI dataset (Weber-Genzel et al., 2024) was designed for error detection by distinguishing between annotation mistakes and legitimate human label variation in natural language inference; it includes both labels and annotator explanations and was evaluated using the same metrics as the MP dataset. In this study, we used the official training and validation splits provided by the organizers, and our final models performed inference on the unlabeled test sets. The Dataset statistics are presented in Table 1.

## 3 System Overview

Our system architecture for the LeWiDi task is illustrated in Figure 1. Following dataset preprocessing, which involves the extraction and organization of the dataset along with annotator metadata, we designed an embedding pipeline that begins with pre-computations from a transformer model. For the MP dataset, we obtained high dimentional embeddings from XLM-RoBERTa model[1] because of the multilingual properties of the dataset. For other datasets, we employed the all-MiniLM-L12-v2 model,[2] from the Sentence-Transformers library. In our setup, after obtaining the embeddings for each sentence pair, the model's vocabulary was dynamically extended with two special tokens. The first token represents enrichment features, computed by calculating cosine similarity, Manhattan distance, and Euclidean distance, as well as element-wise multiplication and difference, to capture multiple similarity features between corresponding sentence pairs. The second token represents annotator features, following the strategies we developed for annotator modeling. For every annotator ID, we create three annotator tokens: a user ID token which uses the user-id of each annotator, a user passport token derived from annotator metadata, and a composite token linking the annotator to its label patterns. The user passport token incorporates all available information about the annotator. In addition to these three tokens, we explored their combinations with the composite token, specifically, composite with user ID and composite with user passport resulting in five annotator modeling approaches. Previously, these approaches were used for a single-sentence setup (Sarumi et al., 2024). We performed feature fusion, combining the different annotator strategies with the enrichment features (Sarumi et al., 2025b), which served as a constant base for the fusion. The resulting vector representation serves as input to our model, which includes two residual blocks to mitigate gradient vanishing, followed by a three-layer Multi-Layer Perceptron (MLP) and a multi-head self-attention mechanism designed to capture different aspects of the combined features. The model then branches into two types of prediction heads: a soft head for predicting the probability distribution of a label, and hard head, dedicated to predicting the specific label for an individual annotator. The soft head is trained with the Kullback–Leibler Divergence loss (KLDivLoss), while the hard head is trained with

[1] Multilingual XLM-RoBERTa model, Hugging Face Transformers library.

[2] all-MiniLM-L12-v2 model Sentence-Transformers library.

cross-entropy loss (CrossEntropyLoss). This architecture allows the model to simultaneously and jointly learn the label distributions and annotator-specific predictions.

## 4 Experimental Setup

Our system used the datasets provided by the organizers. Table 1 presents the statistics for the training and development splits of each dataset. Building on existing work, we implemented slightly modified variants of some annotator modeling techniques, as described earlier, and introduced a new approach, the User Passport Model. This model leverages extended annotator demographic profiles, making use of rich metadata.

All five annotator modeling approaches were trained on each dataset using a unified framework: consistent annotator representations, feature enrichment strategies, and training procedures were applied across datasets. We obtained precomputed sentence embeddings from SBERT all-MiniLM-L12-v2 for all datasets, with the exception of the MP dataset, for which XLM-RoBERTa embeddings were used. These embeddings were concatenated with the enrichment features and annotator representations to form the combined input representation.

The downstream model employed a multi-layer perceptron (MLP) backbone, extended with a multi-head self-attention mechanism (two heads), which we implemented from scratch. The first head ("soft head") was designed to predict label distributions and the second head ("hard head"), aligned with the perspectivist approach, was designed to predict the individual annotator labels. The two objectives were jointly optimized with a combined loss function, enabling the model to learn both soft and hard targets concurrently.

### 4.1 Methods

Here we describe the various annotator modelling approaches we implemented, drawing on existing literature as well as the new methods we introduced, namely the User Passport and Composite User Passport modelling techniques.

**User-ID Token** The User ID Token approach uses a single, unique special token for each annotator, using its ID as provided. This token serves as a lightweight identifier. The model learns a specific embedding for each of these tokens, which helps it understand that a particular annotation is tied to a particular user (Plepi et al., 2022).

**User-Passport Token** The User Passport is a unique special token that represents an individual annotator based on their demographic metadata, encoded as a trainable embedding. We dynamically process the annotator metadata file, which contains all available demographic information for each annotator. During training, the model implicitly encodes the demographic traits associated with each passport token, effectively creating a *passport* that captures the annotator's profile. This passport token is appended to the input text, enabling the model to make predictions while being aware of the specific annotator's profile.

**Composite Token** The Composite approach uses a special token whose embedding is computed as the average embedding of all instances in which an annotator assigned a specific label. The model learns an embedding for each composite token, capturing the annotator's characteristic judgment style and linking them directly to their specific type of annotation. (Plepi et al., 2022; Sarumi et al., 2024)

**Composite+User-ID Token** The Composite User ID approach combines the strengths of the previous two methods by appending both the unique User ID token and the Composite User Token to the input text. This provides the model with richer context, enabling it to capture both the annotator's individual identity and their characteristic judgment style for a given label. This dual-token strategy strengthens the link between annotator identity and annotator behaviour.

**Composite+User-Passport Token** The Composite User Passport Token combines the User Passport and the Composite Token by appending both the relevant composite token for the given annotator and the corresponding User Passport token to the input text. This creates a robust representation of the annotator, capturing both their demographic profile and their characteristic judgment style.

### 4.2 Evaluation Metrics

Following the definitions of the two tasks A and B, focused on predicting the probability distribution of a value (soft labels) and the individual hard labels of annotators, respectively, the performance of our

| | #A | #I | N | A/I | CL | K-$\alpha$ |
|---|---|---|---|---|---|---|
| CSC | 872 | 6,332 | 33± 14 | 4.54± 0.01 | 212± 76.73 | 0.34 |
| MP | 506 | 15,022 | 150±0.76 | 5.04±0.01 | 293±431.81 | 0.26 |
| PAR | 4 | 450 | 450± 0.00 | 4.00± 0.00 | 108± 45.49 | 0.09 |
| VariErr NLI | 4 | 434 | 419± 4.53 | 3.86± 0.04 | 177± 111.58 | -0.06 |

Table 1: Dataset statistics including the number of annotators (A), the number of total instances (I), the average number of annotations per annotator (N), average annotations per instance (A/I), the average context length (CL), the agreement as measured by Krippendorff's alpha. (*The statistics are based only on train and dev splits*).

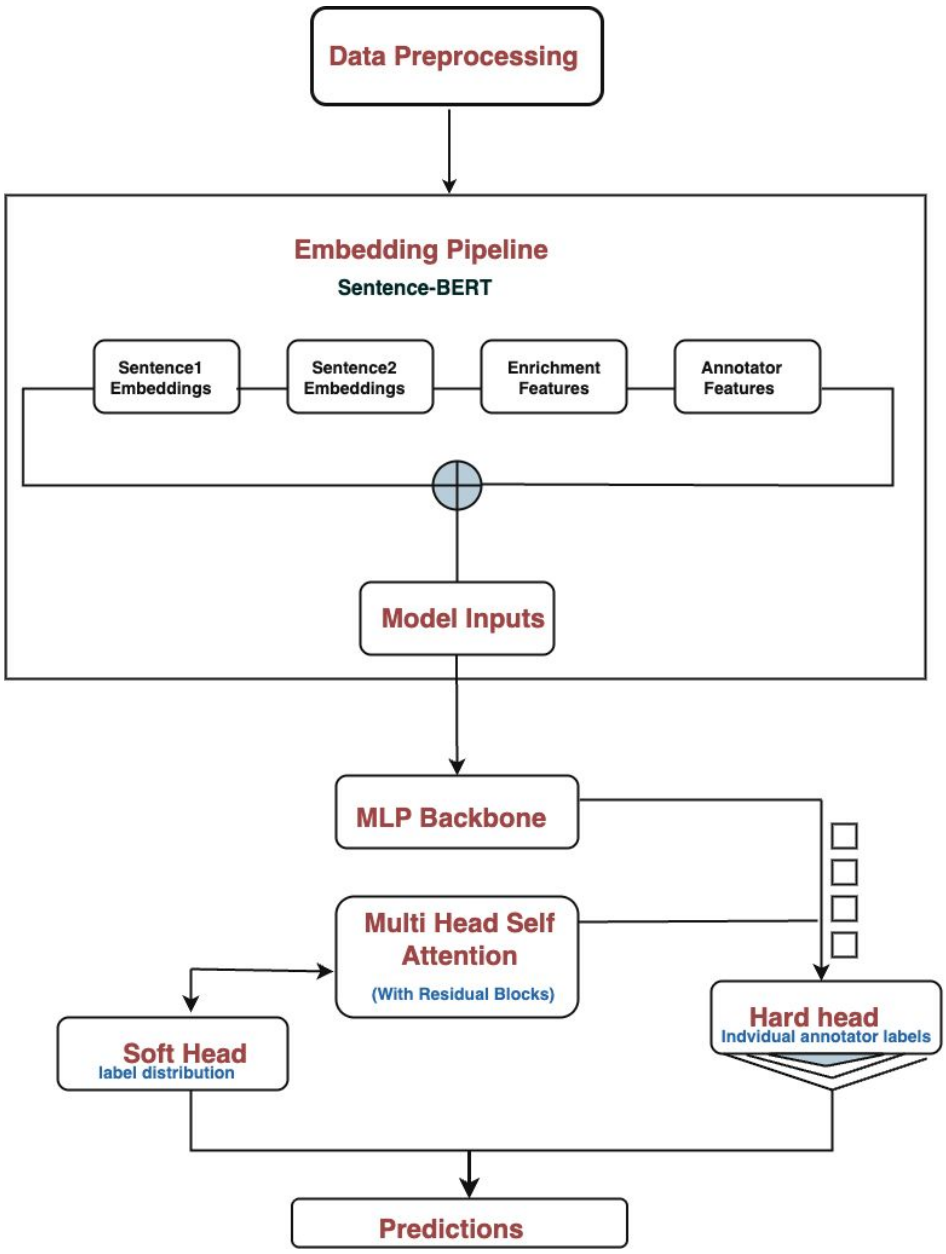


Figure 1: System architecture

system was primarily evaluated using the official metrics specified by the organizers. However, we also used additional evaluation metrics, not because they are inherently more suitable for the tasks, but to investigate whether the annotator model that performs best under one metric remains the best when evaluated with a different metric considering how dynamic it is for models to learn from disagreement. This allowed us to assess the sensitivity of model performance to evaluation criteria across different annotator modeling strategies. For the perspectivist task (Task B), we further analyzed model perfor-

mance using individual F1 scores and ROC-AUC scores. The CSC and Paraphrase datasets were evaluated using the official soft evaluation metric: Average Wasserstein Distance $(AWD)$. As seen in equation $(i)$.

$$AWD = \frac{1}{N}\sum_{i=1}^{N}\min_{\gamma\in\Gamma(p_i,t_i)}\sum_{h=1}^{n}\sum_{k=1}^{n}\gamma_{h,k}|h-k| \tag{1}$$

For the perspectivist evaluation of the same datasets, the Mean Absolute Distance $(MAD)$ between the actual labels and the predictions were measured. $(ii)$

$$MAD = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{a}\sum_{k=1}^{a}\frac{|t_{i,k}-p_{i,k}|}{s}\cdot 100 \tag{2}$$

The MultiPico and VariErr-NLi datasets, were evaluated with the average Manhattan distance $(AverageMD)$ in the Soft evaluation, see the equation $(iii)$, while the hard evaluation was based on Error rates computation as in equation $(iv)$ with slight modification as multi-label average MD and multi-label error rate for the VariErr-NLi dataset.

$$AverageMD = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{n}|p_{i,k}-t_{i,k}| \tag{3}$$

and

$$AverageER = \frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{1}{a}\sum_{k=1}^{a}|t_{i,k}-p_{i,k}|\right) \tag{4}$$

### 4.3 Training

The training was performed using the AdamW optimizer, with a fixed learning rate of $1\times10^{-3}$. A cosine annealing learning rate scheduler was applied with $T_{\max}=10$. We trained our models for 10 epochs, with early stopping based on the minimum validation soft metric and maximum hard metric as the case may be, using a patience of 5. The batch size was set to 16, and training used a single NVIDIA A100 40GB GPU. The loss functions combined KL divergence and Jensen-Shannon divergence for the soft label head, and cross-entropy loss for the hard label heads.

## 5 Results

For the evaluation of the LeWiDi shared task, two categories of metrics were used: soft label metrics and perspectivist metrics. In the soft label evaluation, the probability distribution (soft label) predicted by the system was compared against the distribution derived from human annotations. A lower distance between the predicted and human soft labels indicated better performance, with a perfect prediction yielding a distance of zero. For the CSC and PAR datasets, Wasserstein distance was used, while for the MP and VariErrNLI datasets, Manhattan distance was applied. In the perspectivist evaluation, the focus was predicting individual annotators labels. Performance was measured using Mean Absolute Distance (MAD) between predicted and actual annotator labels. Although participants could submit multiple runs, our late entry into the competition allowed only one submission before the evaluation phase closed. Based on the evaluation scores posted on the Leadersboard, our scores for the soft and the perspectivist tasks are shown in Table 6 where we compared our system's performance to the top-performing models on the leaderboard, including teams Opt-ICL (Leonardelli et al., 2025; Sanghani et al., 2025). These results placed us between 9th and 10th on the leaderboard based on average score. Our submission was based on our composite model, which, with the addition of more hidden layers, improved results for most datasets except PAR. Post-evaluation results from our improved models, computed using the Codabench platform, are presented in Table 2. Results based on the dev splits, which were not processed through Codabench, are reported in Table 3. Additional evaluations using traditional metrics such as F1-score and ROC-AUC are reported in Tables 4 and 5 respectively.

## 6 Discussion

The performance of annotator modeling techniques is not universal but is highly dependent on the characteristics of the dataset and the focus of the evaluation metric. We observe key differences in how models learn and perform on datasets with varying numbers of annotators, annotation strategies and subjective levels.

On the MP dataset, characterized by a large pool of annotators (>500), the highest number of instances, and the longest average context length, cf. Table 1 the Composite + User Passport model

| | Task A (Soft) | | | | Task B (Hard) | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **CSC** | **MP** | **PAR** | **VariErrNLI** | **CSC** | **MP** | **PAR** | **VariErrNLI** |
| User-ID Token | **1.171** | 0.519 | 3.320 | **0.59** | **0.241** | **0.322** | 0.350 | 0.350 |
| User Passport Token | **1.171** | 0.510 | **3.280** | **0.590** | 0.247 | **0.322** | 0.340 | 0.350 |
| Composite Token | 1.185 | **0.508** | 3.300 | **0.590** | 0.249 | 0.323 | 0.310 | 0.350 |
| Composite + User-ID | 1.193 | 0.533 | 3.300 | 0.600 | 0.248 | 0.336 | 0.340 | 0.350 |
| Composite + User Passport | 1.175 | 0.538 | **3.280** | 0.610 | 0.246 | 0.353 | **0.290** | 0.350 |

Table 2: Results for different annotator modeling approaches (**Post Evaluation computed on Codabench**). The specific evaluation metrics vary by task and dataset. **Task A (Soft)** metrics are Wasserstein Distance (CSC, PAR), Soft-Manhattan Distance (MP), and Soft-Multi-Label-Manhattan Distance (VariErrNLI). **Task B (Hard)** metrics are Mean Absolute Distance (CSC, PAR), Hard-Error rate (MP), and Hard-MultiLabel-Error rate (VariErrNLI). For all metrics, lower values are better. Best results are shown in **bold**.

| | Task A (Soft) | | | | Task B (Hard) | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **CSC** | **MP** | **PAR** | **VariErrNLI** | **CSC** | **MP** | **PAR** | **VariErrNLI** |
| User-ID Token | 1.278 | **0.513** | 2.812 | 0.885 | 0.228 | 0.323 | 3.620 | 0.705 |
| User Passport Token | 1.288 | 0.537 | **2.786** | 0.901 | 0.232 | 0.330 | 3.620 | 0.705 |
| Composite Token | 1.212 | 0.529 | 2.891 | **0.878** | 0.229 | 0.324 | **3.460** | **0.695** |
| Composite + User-ID | **1.177** | 0.538 | 2.999 | 0.889 | 0.227 | 0.324 | 3.660 | 0.705 |
| Composite + User Passport | 1.253 | 0.524 | 2.846 | 0.881 | **0.226** | **0.318** | 3.620 | 0.705 |

Table 3: Results for different annotator modeling approaches (**Post Evaluation (ours)**). Dataset abbreviations are: CSC, MP, PAR, and VariErrNLI. The specific evaluation metrics vary by task and dataset. **Task A (Soft)** metrics are Wasserstein Distance (CSC, PAR), Soft-Manhattan Distance (MP), and Soft-Multi-Label-Manhattan Distance (VariErrNLI). **Task B (Hard)** metrics are Mean Absolute Distance (CSC, PAR), Hard-Error rate (MP), and Hard-MultiLabel-Error rate (VariErrNLI). For all metrics, lower values are better. Best results are shown in **bold**.

| Method | CSC | MP | PAR | VariErrNLI |
|---|---|---|---|---|
| User-ID Token | 23.1 | 32.5 | 08.8 | 70.5 |
| User Passport Token | 23.4 | 34.6 | 14.5 | 70.5 |
| Composite Token | 23.2 | 38.0 | **16.5** | 69.5 |
| Composite + User-ID | **23.8** | 36.8 | 11.1 | 70.5 |
| Composite + User Passport | 23.4 | **39.1** | 11.3 | 70.5 |

Table 4: Full dataset result F1 scores on the **individual annotator** labels for each annotator representation method and dataset for the Task B

| Method | CSC | MP | VariErrNLI |
|---|---|---|---|
| User-ID Token | **67.2** | 52.7 | 88.5 |
| User Passport Token | 66.4 | 60.4 | **90.1** |
| Composite Token | 64.9 | 60.2 | 87.8 |
| Composite + User-ID | 65.2 | 60.6 | 88.9 |
| Composite + User Passport | 65.4 | **61.8** | 88.1 |

Table 5: Full dataset result ROC scores on the **individual annotator** labels for each annotator representation method and dataset for the Task B

consistently performed best across all evaluation strategies, including minimising the error rate, truth prediction measured by the F1-Score, and its classification ability measured by the ROC-AUC score, however, this was not observed on other datasets. A key characteristic of this annotator technique is its use of all demographic information available from the corpus metadata, which contributes to its robustness. The MP dataset has more demographic information than the other datasets.

In contrast, on the CSC dataset, Composite + User Passport performed best when error rate was being measured, further strengthening the ability of the model to minimise error, especially on large datasets; however, the CSC dataset has less demographic information than the MP dataset. We see the impact of this without their composite token in the ROC scores for CSC and VariErrNLI where User-ID token performs best for the CSC and User Passport performs best for VariErrNLI.

| | CSC | MP | PAR | VariErrNLI |
|---|---|---|---|---|
| **Soft Task** | | | | |
| Baseline (Random) | 1.543 | 0.687 | 3.350 | 0.676 |
| Ours | **1.393** | **0.551** | **3.136** | 1.000 |
| Top Submission | 0.746 | 0.422 | 0.983 | 0.233 |
| **Perspectivist Task** | | | | |
| Baseline (Random) | 0.352 | 0.499 | 0.367 | 0.497 |
| Ours | **0.291** | **0.326** | 0.418 | **0.345** |
| Top Submission | 0.156 | 0.289 | 0.080 | 0.124 |

Table 6: Leaderboard Evaluation Results. Best overall results are underlined.

| | Error Rate | MAD | F1 | ROC |
|---|---|---|---|---|
| User ID Token | | | ✓ | ✓ |
| User Passport Token | | | ✓ | ✓ |
| Composite Token | ✓ | ✓ | ✓ | |
| Composite User ID | | | ✓✓ | |
| Composite Passport | ✓ | ✓ | ✓✓ | ✓ |

| Legend | |
|---|---|
| CSC | ✓ |
| MP | ✓ |
| PAR | ✓ |
| VAR | ✓ |

Table 7: Performance shift analysis of Anotator models across different evaluation metrics for (Task B-Perspectivist approach)

The VariErrNLI dataset is highly subjective, with an agreement score of -0.06 and a very small number of annotators, with each annotator annotating more than 95% of the total instances. The User Passport model performs well while measured with ROC score, which suggests the model is particularly strong at capturing distinct classification features of the data, which did not translate to larger datasets. Across the datasets, all except VariErrNLI struggled with the F1 score evaluation, plateauing at 70.5, except for the composite model, with reduced performance and a slight reduction in error rate. This shows that different models capture different aspects of data. Some better account for individual labels in highly subjective corpora, which may preserve minority labels, while others output high scores in large corpora sometimes aggregating towards majority labels. Therefore, in modeling perspectives, there is a need for careful consideration of what has been measured vis-à-vis the minority and majority classes. An optimal model will ultimately harness the strength of different evaluation strategies.

## 7 Conclusion

Previous works have established that certain statistics, particularly the number of annotations per annotator and the IAA, are critical to the performance of annotator modeling approaches. Apparently, these factors reflect underlying dataset characteristics. Although prior findings were often based on evaluations using individual macro F1 scores, our observations as shown in Tables 1 and 7, confirm perspectivism even in evaluation and dataset characteristics. All datasets in our study exhibit low Krippendorff's alpha scores, indicating high disagreement among annotators with VariErrNLI dataset with the highest disagreement score of negative alpha value.

In conclusion, the choice of evaluation metric significantly influences which annotator modeling approach emerges as the best-performing model, with focus on the Task B Perspectivist Evaluation. Across CSC, MP, PAR, and VAR, no single approach consistently ranked highest across all metrics. Composite+User Passport ranked best consistently on the MP dataset but with lower scores

when compared across corpora. These results confirm that model rankings are not metric-agnostic; a model optimised for one evaluation metric may not retain its advantage when assessed with another, underscoring the need for further work that assesses and harnesses the strength of perspectivist systems while leveraging integrated evaluation approaches.

## Limitations

A limitation of our system was the absence of task-specific fine-tuning with a pre-trained language model. We hypothesize that this approach could significantly improve the results. The models we implemented were also slight variants of existing architectures, specifically adapted for this shared task. A full implementation of these models, without the modifications we made for the competition, could also lead to further performance gains. These two represent areas for future work and potential improvements in addition to exploring an integrated perspectivist evaluation system. Our code is publicly available on GitHub[3].

## References


Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 6860–6868. AAAI Press.

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 62th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*.

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.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, 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 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. 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 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.

MaiNLP Lab. 2025. The Paraphrase Detection (Par) Dataset. Unpublished manuscript, Ludwig-Maximilians-Universität Munich. Managed by the MaiNLP Lab, Ludwig-Maximilians-Universität Munich.

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.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402, Abu Dhabi, United Arab Emirates. 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 Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94, Torino, Italia. ELRA and ICCL.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Aadi Sanghani, Sarvin Azadi, Virendra Jethra, and Charles Welch. 2025. McMaster at LeWiDi-2025: Demographic-Aware RoBERTa. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.


[3] GitHub

Olufunke O. Sarumi, Béla Neuendorf, Joan Plepi, Lucie Flek, Jörg Schlötterer, and Charles Welch. 2024. Corpus considerations for annotator modeling and scaling. 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 1029–1040, Mexico City, Mexico. Association for Computational Linguistics.

Olufunke O. Sarumi, Charles Welch, Daniel Braun, and Jörg Schlötterer. 2025a. The impact of annotator personas on llm behavior across the perspectivism spectrum. *Preprint*, arXiv:2508.17164.

Olufunke O. Sarumi, Charles Welch, Lucie Flek, and Jörg Schlötterer. 2025b. Funzac at CoMeDi shared task: Modeling annotator disagreement from word-in-context perspectives. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 90–96, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Michael Sullivan, Mohammed Yasin, and Cassandra L. Jacobs. 2023. University at buffalo at SemEval-2023 task 11: MASDA–modelling annotator sensibilities through DisAggregation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 978–985, Toronto, Canada. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. 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.

Charles Welch, Chenxi Gu, Jonathan K. Kummerfeld, Veronica Perez-Rosas, and Rada Mihalcea. 2022. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1742–1752, Dublin, Ireland. Association for Computational Linguistics.