LeWiDi-2025 at NLPerspectives: Third Edition of the Learning with Disagreements Shared Task

Elisa Leonardelli[1], Silvia Casola[2], Siyao Peng[2], Giulia Rizzi[3], Valerio Basile[4], Elisabetta Fersini[4], Diego Frassinelli [2], Hyewon Jang[5], Maja Pavlovic[6] Barbara Plank[2], Massimo Poesio[6,7]
[1]Fondazione Bruno Kessler, [2]LMU Munich & MCML,
[3]Università Milano Bicocca, [4]Università di Torino,
[5]University of Gothenburg, [6]Queen Mary University of London, [7]Utrecht University

Abstract

Many researchers have reached the conclusion that AI models should be trained to be aware of the possibility of variation and disagreement in human judgments, and evaluated as per their ability to recognize such variation. The LeWiDi series of shared tasks on Learning With Disagreements was established to promote this approach to training and evaluating AI models, by making suitable datasets more accessible and by developing evaluation methods. The third edition of the task builds on this goal by extending the LeWiDi benchmark to four datasets spanning paraphrase identification, irony detection, sarcasm detection, and natural language inference, with labeling schemes that include not only categorical judgments as in previous editions, but ordinal judgments as well. Another novelty is that we adopt two complementary paradigms to evaluate disagreement-aware systems: the soft-label approach, in which models predict population-level distributions of judgments, and the perspectivist approach, in which models predict the interpretations of individual annotators. Crucially, we moved beyond standard metrics such as cross-entropy, and tested new evaluation metrics for the two paradigms. The task attracted diverse participation, and the results provide insights into the strengths and limitations of methods to modeling variation. Together, these contributions strengthen LeWiDi as a framework and provide new resources, benchmarks, and findings to support the development of disagreement-aware technologies.

1 Introduction

The assumption that natural language (NL) expressions have a unique and clearly identifiable interpretation has been recognized in AI as just a convenient idealization for over twenty years (Poesio and Artstein, 2005; Versley, 2008; Recasens et al., 2011; Passonneau et al., 2012; Plank et al., 2014b; Aroyo and Welty, 2015; Martínez Alonso et al., 2016; Dumitrache et al., 2019; Pavlick and Kwiatkowski, 2019; Jiang and de Marneffe, 2022). More recently, the increasing focus in NLP on tasks depending on subjective judgments (Kenyon-Dean et al., 2018; Simpson et al., 2019; Cercas Curry et al., 2021; Leonardelli et al., 2021; Akhtar et al., 2021; Almanea and Poesio, 2022; Casola et al., 2024) led to the realization that in many NLP tasks the traditional approach to dealing with disagreement of 'reconciling' different subjective interpretations is not tenable. Many AI researchers concluded therefore that rather than eliminating disagreements from annotated corpora, we should preserve them (e.g. Poesio and Artstein, 2005; Aroyo and Welty, 2015; Kenyon-Dean et al., 2018; Pavlick and Kwiatkowski, 2019; Uma et al., 2021b; Davani et al., 2022; Abercrombie et al., 2022; Plank, 2022). As a result, a number of corpora with these characteristics now exist, and more are created every year (Plank et al., 2014a; White et al., 2018; Dumitrache et al., 2019; Poesio et al., 2019; Nie et al., 2020; Cercas Curry et al., 2021; Leonardelli et al., 2021; Akhtar et al., 2021; Almanea and Poesio, 2022; Sachdeva et al., 2022; Casola et al., 2024; Jang and Frassinelli, 2024; Weber-Genzel et al., 2024). Much recent research has therefore investigated whether corpora of this type are also useful resources for training NLP models, and if so, what is the best way for exploiting disagreements (Sheng et al., 2008; Beigman Klebanov and Beigman, 2009; Rodrigues and Pereira, 2018; Uma et al., 2020; Fornaciari et al., 2021; Uma et al., 2021b; Davani et al., 2022; Casola et al., 2023). This research in turn led to questions about how such models can be evaluated (Basile et al., 2021; Uma et al., 2021b; Gordon et al., 2021; Fornaciari et al., 2022; Giulianelli et al., 2023; Lo et al., 2025). A succinct overview of the literature on how the problem affects data, modeling and evaluation in NLP is given in Plank (2022), and an extensive survey can be found in Uma et al. (2021b).

Such research also led to the establishment of the Learning With Disagreements (LeWiDi) shared tasks. The first edition, organized at SemEval 2021 Task 12 (Uma et al., 2021a), introduced the idea of providing a unified testing framework for modeling disagreement and evaluating systems on such data. The benchmark combined six widely used corpora spanning semantic and inference tasks as well as image classification tasks. While the resource attracted considerable attention (the benchmark was downloaded by more than 100 teams worldwide) participation in the evaluation was limited, possibly due to the difficulty of the provided baselines or the need for expertise in both NLP and computer vision. In addition, the benchmark only covered a single subjective task, i.e., humour detection, (Simpson et al., 2019), and a single language (English).

A second edition followed at SemEval 2023 (Leonardelli et al., 2023), designed to address these limitations and to better reflect the growing interest in subjective NLP tasks. In contrast to the first edition, all datasets were textual and the focus shifted entirely to inherently subjective phenomena such as misogyny, hate-speech and offensiveness detection, where training with aggregated labels makes much less sense. Moreover, Arabic was added as a second language. Finally, evaluation combined the soft-label approach also used in the first edition, based on cross-entropy, with the more traditional F1 metric. The reformulated task attracted broad interest in the community: more than 130 groups registered, with 30 submitting predictions and 13 contributing system papers.

The third edition of the LeWiDi shared task, described in this manuscript and co-located with the NLPerspectives Workshop at EMNLP 2025, builds on these experiences while further broadening the scope of the task. Like the earlier editions, its central goal is to provide a common evaluation framework for systems trained on disagreement-rich data. However, LeWiDi-2025 introduces several innovations. New tasks include natural language inference (NLI), irony detection, conversational sarcasm detection, and paraphrase detection. On the evaluation side, we move entirely to soft metrics, which are organized into two complementary tasks: (i) soft-label evaluation, refining methods from LeWiDi 2 with several distance-based metrics (e.g., Manhattan distance, Rizzi et al. 2024); and (ii) perspectivist evaluation, where systems must model the labeling behavior of individual annotators, again with newly developed metrics tailored to this setting. In addition, two of the datasets adopt Likert-scale annotation, posing further challenges for evaluation. LeWiDi 3 engaged a smaller but dedicated group of participants relative to the previous edition. A total of 53 individuals registered on the competition platform, with 15 teams providing submissions, which resulted in 9 system papers.

2 The LeWiDi 3 Benchmark

The four selected datasets are summarized in Table 1, illustrated with examples in Table 2, and described in detail in the following sections.

All datasets were released in a harmonized *json* format, identical to that of the previous LeWiDi edition, to ensure consistent access across datasets and shared tasks editions. Each item contains the same fields,[1] while the field *other info* is dataset-specific and includes additional subfields particular to each dataset. Annotator age and gender is available for all four datasets, with some datasets providing further attributes. This metadata was distributed separately in an additional json file. All datasets are publicly available.[2]

2.1 The Conversational Sarcasm Corpus (CSC)

The CSC dataset (Jang and Frassinelli, 2024) is a dataset of sarcasm in English, which contains around 7,000 context–response pairs. Each pair is rated on a 1 (not at all) – 6 (completely) scale both by the speakers who generated the responses and by multiple external observers (4 - 6 per speaker). The contexts consist of situation descriptions involving an imagined interlocutor, and the responses stem from the responses given by online participants. The generators of the responses as well as evaluators rated the level of sarcasm of the responses.

2.2 The MultiPICo dataset (MP)

The MP dataset (Casola et al., 2024) is a multilingual perspectivist corpus consisting of short exchanges from Twitter and Reddit. Each entry in the corpus represents a post-reply pair. Crowdsourced workers had to determine whether the reply was ironic given the post (binary label). The corpus includes 11 languages: Arabic, Dutch, English,

[1] *item_id*, *text*, *task*, *number of annotations*, *number of annotators*, *disaggregated annotations*, *annotator IDs*, *language*, *hard label*, *soft labels*, *split*, and *other info*.

[2] https://le-wi-di.github.io/

Dataset	Task	Labels	Lang(s)	N. Items (N. Annotations)	N. Ann. per item	Pool Annotators	Textual type	Annotators' Metadata	Other info
CSC	Sarcasm detection	Likert scale [1 to 6]	En	7,036 (31,984)	Variable: 4 to 6	872	context+ response	gender, age	context + speaker
MP	Irony detection	[0,1]	Ar,De,En, Es,Fr,Hi, It,Nl,Pt	18,778 (94,342)	Variable: 2 to 21	506	post+ reply	gender, age, ethnicity, [...+6]	source, level, language variety
Par	Paraphrase detection	Likert scale [-5 to 5]	En	500 (2,000)	4	4	question1 + question2	gender, age, nationality, education	explanations
VEN	Natural Language Inference	[contradiction (C), entailment (E), neutral (N)]	En	500 (1,933)	Variable: 1 to 6	4	context + statement	gender, age, nationality, education	explanations

Table 1: Key statistics about the datasets used in the 3rd LeWiDi shared task.

French, German, Hindi, Italian, Portuguese, and Spanish. It also contains sociodemographic information about the annotators, including gender, age, nationality, race, and student or employment status. While the statistics may vary slightly across languages, each post-reply pair is typically annotated by an average of 5 workers.

2.3 The VariErr NLI dataset (VEN)

VariErr NLI (Weber-Genzel et al., 2024) was designed for automatic error detection, distinguishing between annotation errors and legitimate human label variations in NLI tasks. The dataset was created using a two-round annotation process: initially, four annotators provided labels and explanations for each NLI item; subsequently, they assessed the validity of each label-explanation pair. It comprises 1,933 explanations for 500 re-annotated items from the Multi-Genre Natural Language Inference (MNLI) corpus for Round 1 and 7,732 validity judgments for Round 2. The LeWiDi 2025 Shared Task focuses on Round 1 (and therefore we refer to it just as *VEN*), where annotators could assign one or more labels from Entailment, Neutral, Contradiction to each Premise ("context") - Hypothesis ("statement") pair and provide corresponding explanations.

2.4 The Paraphrase Detection dataset (Par)

The Par dataset focuses on paraphrase detection. It is structurally similar to *VEN*, but unlike *VEN*, the labels here are scalar and each annotator provides only a single score per item. It consists of 500 question pairs sampled from the Quora Question Pairs (QQP) dataset, each annotated independently by the same four annotators. Annotations are given on a Likert scale from -5 to 5, indicating the perceived degree of paraphrastic relation between the questions, and are accompanied by short textual explanations. As this dataset had not been released previously, it was new to the participants of LeWiDi-2025.

3 Task definition

The main goal of the shared task is to provide a unified testing framework for learning from disagreements and evaluating models on such datasets. Given the heterogeneous nature of the datasets, participants were free to design dataset-specific approaches; however, they were encouraged to adopt a unified crowd learning methodology or framework across all datasets, rather than optimizing a separate best-performing model for each dataset.

3.1 Task A and Task B

LeWiDi-2025 defines two complementary tasks.

Task A: Soft-label prediction. Participants are required to predict a *probability distribution* over the possible labels for each item. Evaluation is based on the predicted distribution and the gold soft label distribution. This task continues the line of soft-label modeling from previous editions, but is now applied across expanded datasets, including those with Likert-scale judgments.

Task B: Perspectivist prediction. Participants must predict the *individual label choices of annotators*, i.e., model how a specific annotator would label a given instance. Evaluation measures the agreement between predicted and actual annotator-level responses. This task emphasizes capturing annotator bias and perspective.

Participants may choose to submit to one or both tasks, and across any subset of the provided datasets.

Dataset (detection of)	Example	Annotations (Task B) *AnnotatorId:Label*	Soft labels (Task A) *Label:Probability*
CSC (Sarcasm)	context: "You walk into the room and Steve is there and Steve says “hi!”" response: "hi"	A812:1, A813:3, A814:1, A815:2	[1:0.5, 2:0.25, 3:0.25, 4:0, 5:0, 6:0]
MP (Irony)	post: "@USER Oh dear" reply: "@USER It's ok, wine has fixed everything"	A26:1, A64:1, A70:1	[0:0, 1:1]
Par (Paraphrase)	Q1: "Have you seen an alien craft?" Q2: "Have you ever seen an alien?"	A1:-1, A2:-3, A3:5, A4:4	[-5:0, -4:0, -3:0.25, -2:0, -1:0.25, 0:0, 1:0, 2:0, 3:0, 4:0.25, 5:0.25]
VEN (NLI)	context: "yeah i can believe that" statement: "I agree with what you said."	A1:E, A2:N, A3:N, A4:E	[C:[0:1, 1:0] E:[0:0.5, 1:0.5] N:[0:0.5, 1:0.5]]

Table 2: Examples from the four datasets included in LeWiDi-2025. For each item, the annotators' IDs and their corresponding annotations are shown, along with the derived soft-label distributions. Task B required predicting an individual annotator's label given their ID, while Task A required predicting the full soft-label distribution for the item.

Codabench served as the official competition platform, where participants registered to access the data and to submit their results.[3]

3.2 Phases

The competition consisted of three phases:

Practice phase: Participants received training and development data (with full metadata) to design and test their models. They could submit their results (on the development data) to *Codabench* and compare results on a public leaderboard.

Evaluation phase: Participants submitted predictions on unseen test data (without labels). Rankings were computed for each dataset and across datasets, with missing submissions replaced by the organizer's baseline score.

Post-campaign phase: To support long-term research, the test data and gold labels were later released publicly and remain available through our website[3].

3.3 Baselines

We provided two simple baselines: (i) a *random* baseline, where each distribution (Task A) or prediction (Task B) was assigned a random prediction, and (ii) a *most frequent* baseline, where all items were assigned the most frequent distribution within the training set (Task A) or label. These baselines were intentionally kept minimal so as not to discourage participation, unlike in the first edition of the shared task.

[3] https://www.codabench.org/competitions/7192/

4 Evaluation metrics

Two complementary paradigms for disagreement evaluation were employed in LeWiDi-2025: soft-label and perspectivist evaluation.

4.1 Soft-label Evaluation

In soft-label evaluation, annotator judgments are represented as probability distributions (soft labels), and system predictions are evaluated against these human-derived soft labels by measuring the distance between the two distributions. Previous editions of LeWiDi employed cross-entropy as the distance metric. However, Rizzi et al. (2024) demonstrated that cross-entropy exhibits several counterintuitive properties, whereas the Manhattan and Euclidean distances provide a more suitable alternative in the context of binary classification. At the same time, they highlighted the limitations of the analyzed metrics in providing fair comparisons for multiclass classification tasks.

Based on previous findings, here we address the broader settings introduced in this edition of the shared task, i.e., multiclass and multilabel classification, as well as labels on a Likert scale. In LeWiDi-2025, both the Manhattan distance and the Wasserstein (Earth Mover's) distance are adopted as the primary soft evaluation metrics. Specifically, the Average Manhattan Distance is applied to the *MP* and *VEN*[4] datasets, while the Average Wasserstein Distance is used for the ordinal-scale datasets

[4] Considering the nature of the dataset itself, a multilabel adaptation of the Average Manhattan distance has been proposed. Additional details are reported in Appendix A.

(i.e. *Par* and *CSC*).

In particular, for what concerns the Average Wasserstein Distance (AWD), the cost of transporting probability mass from one bin to another is defined as the absolute difference between their positions, forming a symmetric, non-negative ground distance matrix with zeros on the diagonal.

4.2 Perspectivist Evaluation

The perspectivist evaluation focuses on assessing a system's ability to model the individual label choices of annotators. For datasets with nominal categories (*MP*, *VEN*), performance is measured using error rate; for datasets with ordinal categories (*Par*, *CSC*), a normalized absolute distance is used.

In particular, the average error rate (AER) (Equation 1), which measures the degree of error between corresponding pairs of target and predicted value vectors is computed as follows:[5]

$$AER = \frac{1}{N}\sum_{i=1}^{N} ER(i) \tag{1}$$
$$= \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{a - \sum_{k=1}^{a}|t_{i,k} - p_{i,k}|}{a}\right) \tag{2}$$

Where the Error Rate (ER) for a single sample i with target label vector $\vec{t}_i = [t_1, t_2, ... t_a]$, and predicted label vector $\vec{p}_i = [p_1, p_2, ... p_a]$ is defined as:

$$ER(i) = 1 - \frac{a - \sum_{k=1}^{a}|t_{i,k} - p_{i,k}|}{a} \tag{3}$$

Here, a denotes the length of the vectors (i.e., the number of annotators), and N is the total number of samples.

The Average Normalized Absolute Distance (ANAD) across all samples is defined as:

$$ANAD = \frac{1}{N}\sum_{i=1}^{N} NAD(i) \tag{4}$$
$$= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{a}\sum_{k=1}^{a}\frac{\left|t_{i,k} - p_{i,k}\right|}{s} \times 100 \tag{5}$$

Where the Normalized Absolute Distance (NAD) for a single sample i with target label vector $t_i = [t_1, t_2, ..., t_a]$, and predicted label vector $p_i = [p_1, p_2, ..., p_a]$ is:

$$NAD(i) = \frac{1}{a}\sum_{k=1}^{a}\frac{|t_{i,k} - p_{i,k}|}{s} \times 100 \tag{6}$$

[5]A multilabel adaptation of the average error rate has been adopted for *VEN*; see Appendix A for further details.

with a denoting the number of annotators, and s the scaling factor given by the range of the Likert scale.

5 Participating systems

The third edition of the LeWiDi shared task attracted a smaller but more focused community compared to the previous edition. In total, 53 people subscribed to the competition *Codabench*, and 15 teams submitted predictions. Among them, 6 teams participated across all datasets and both tasks; 2 teams submitted for three datasets and both tasks (excluding *VEN*); and 5 teams focused on a single dataset with submissions only for Task A. In terms of system papers, 9 were submitted: 6 from teams who participated in multiple tasks and datasets, and 2 from teams who worked on a single dataset and Task A. Task A was overall more popular, as the majority of teams who submitted exclusively for one dataset contributed only to Task A, while 11 teams engaged also with Task B.

5.1 Systems overview

This section provides an overview of the participating systems, focusing on the 9 participating teams that submitted system papers, describing their architectures, methodologies, and key features relevant to the evaluation tasks.

`Opt-ICL` (Sorensen and Choi, 2025) combines in-context learning (ICL) with fine-tuning in a two-stage approach. They first apply post-training, by exposing an LLM to over 40 datasets rich in human disagreement (Sorensen et al., 2025), and then, for each dataset, conduct supervised fine-tuning, using in-context demonstrations from all the individual annotators along with annotator demographics. At inference, the model performs per-rater prediction by constructing a prompt with as many training examples from that annotator as possible, followed by the input to be labeled. They derive soft label distributions from perspectivist predictions.

`DeMeVa` (Ignatev et al., 2025) employs LLMs with ICL, modeling perspectivism through annotators' past behavior. They focus on criteria for selecting demonstrative examples for LLMs (10 per annotator), comparing semantic and label-based strategies, with the latter performing better for multi-label datasets. They derive soft label distributions from perspectivist predictions.

`twinhter` (Nguyen and Van Thin, 2025) built a BERT-based model that integrates annotator perspectives by creating a new (text, annotator) pair. They create a separate training instance for each annotator's view and combine it with their background information when available, enabling the model to capture individual interpretations of the same input.

`McMaster` (Sanghani et al., 2025) implemented a demographic-aware RoBERTa model that incorporates information such as age, gender, nationality, and evaluated it across all four datasets. The authors find that nationality and ethnicity in particular show the largest gains in performance, while also noting the limitations of relying on such features.

`BoN Appetite Team` (Ruiz et al., 2025) investigated three test-time scaling methods, a way to improve LLMs performances: two benchmark algorithms (Model Averaging and Majority Voting), and a Best-of-N (BoN) sampling method. Their results show that the benchmark methods (Averaging and Voting) reliably boost performance, while BoN sampling does not transfer well from mathematical domains.

`PromotionGo` (Huang et al., 2025) submitted only to the *MP*-Task A with an XLM-R–based system, ranking first. They deployed three main strategies to develop a competitive system: data augmentation, including lexical swaps, prompt-based reformulation, and large-scale back-translation into nine languages; optimization for alignment to the evaluation metric (Manhattan Distance) by using L1 loss as a loss function; ensemble learning, by training multiple models on shuffled data splits and averaging predictions to improve robustness.

`Uncertain Mis(Takes)` (Staliūnaitė and Vlachos, 2025) addressed only the *VEN*-Task A, ranking first. They aim to quantify ambiguity in NLI instances, relying on the hypothesis that if a given instance is ambiguous, then the explanations for different labels will not entail one another. For each item, they generate 128 LLM explanations. With a fine-tuned entailment model they cluster them and quantify their Semantic Entropy (SE). The explanation clusters' SE scores are combined with text embeddings for soft label distribution prediction.

`NLP-ResTEAM` (Sarumi et al., 2025) proposed a multi-task architecture. Special 'tokens' are added to the input, including several tokens aiming at modeling the annotators based on their ID, their demographic features, their annotation behavior, or combinations of those. The system produces two outputs from a textual input and an annotator's information: one is a soft-label, the other a prediction of that specific annotator's (hard) label.

`LPI-RIT` (Sawkar et al., 2025) builds upon the DisCo (Distribution from Context) architecture (Weerasooriya et al., 2023), a neural model that jointly predicts item-level, annotator-level, and per-annotator label distributions. They tackled both soft-label and perspectivist tasks simultaneously. They also introduced several extensions to DisCo, such as integrating annotator metadata through pretrained sentence encoders, and modified loss functions to better align with evaluation metrics.

6 Results and discussion

This section presents the official results of the shared task and discusses key trends across systems and datasets. We also examine the role of evaluation metrics and summarize insights from ablation studies conducted by participating teams.

6.1 Results and statistics

Table 3 and 4 report the overall leaderboard for Task A and Task B respectively. If a team did not submit predictions for a particular dataset or task, we used the random baseline results to compute the overall ranks and average positions. Ranks were calculated with statistical ties taken into account. Specifically, we used the Wilcoxon signed-rank test at the instance level to identify clusters of tied systems. Predictions that were not significantly different ($p = 0.05$) from the top-performing system in a given cluster were considered ties. A new cluster was formed when a system's performance was found to be statistically different from that of the best-performing system in the previous cluster. Leadboards for each specific dataset are reported in Appendix B.

6.2 General discussion

As in the previous edition of the shared task, we observed a great variety in design choices, but some trends emerge.

System choices Some teams (`OCP-ICL`, `DeMeVa`, `BoN Appetit Team`) used large language models relying on in-context learning (ICL) or test-time scaling methods. Others built on transformer models (RoBERTa, BERT, or XLM-R) and trained on the shared task data with annotator-aware extensions (`McMaster`, `twinhter`, `NLP-ResTeam`), or with data augmentation and ensembles but without

Rank	(av.pos)	Team	CSC WS	CSC (rank)	MP MD	MP (rank)	Par WS	Par (rank)	VEN MMD	VEN (rank)
			SOFT EVALUATION							
1	*(1.5)*	Opt-ICL	**0.746**	*(1)*	**0.422**	*(1)*	**1.200**	*(1)*	0.449	*(3)*
2	*(2.75)*	DeMeVa	**0.792**	*(1)*	0.469	*(6)*	**1.120**	*(1)*	0.382	*(3)*
3	*(3)*	twinhter	0.835	*(5)*	0.447	*(5)*	**0.983**	*(1)*	**0.233**	*(1)*
4	*(4.25)*	McMaster	0.803	*(3)*	0.439	*(3)*	1.605	*(4)*	0.638	*(7)*
5	*(4.75)*	BoN Appetite Team	0.928	*(6)*	0.466	*(6)*	1.797	*(4)*	0.356	*(3)*
6	*(5.5)*	aadisanghani*	0.803	*(3)*	0.439	*(3)*	3.051	*(7)*	BSL	*(9)*
7	*(7)*	PromotionGo	BSL	*(11)*	**0.428**	*(1)*	BSL	*(7)*	BSL	*(9)*
8	*(7.25)*	*Most frequent baseline*	1.170	*(7)*	0.518	*(8)*	3.231	*(7)*	0.595	*(7)*
9	*(7.5)*	Uncertain Mis(Takes)	BSL	*(11)*	BSL	*(11)*	BSL	*(7)*	**0.308**	*(1)*
10	*(8.5)*	NLP-ResTeam	1.393	*(9)*	0.551	*(9)*	3.136	*(7)*	1.000	*(9)*
10	*(8.5)*	LPI-RIT	1.451	*(9)*	0.540	*(9)*	3.715	*(7)*	BSL	*(9)*
12	*(8.75)*	cklwanfifa*	BSL	*(11)*	BSL	*(11)*	BSL	*(7)*	0.469	*(6)*
12	*(8.75)*	harikrishnan_gs*	1.295	*(8)*	BSL	*(11)*	BSL	*(7)*	BSL	*(9)*
12	*(8.75)*	tdang*	BSL	*(11)*	BSL	*(11)*	1.665	*(4)*	BSL	*(9)*
15	*(9.5)*	*Random baseline (BSL)*	1.543	*(11)*	0.687	*(11)*	3.350	*(7)*	0.676	*(9)*

Table 3: Overall Task A (soft evaluation) results as an average of a system's rank across datasets. * indicates that no system description was available for the team.

Rank	(av.pos)	Team	CSC MAD	CSC (rank)	MP ER	MP (rank)	Par MAD	Par (rank)	VEN MER	VEN (rank)
			PERSPECTIVIST EVALUATION							
1	*(1.5)*	Opt-ICL	**0.156**	*(1)*	**0.289**	*(1)*	0.119	*(2)*	0.270	*(2)*
2	*(2)*	DeMeVa	0.172	*(2)*	0.300	*(2)*	0.134	*(2)*	0.228	*(2)*
3	*(3.25)*	twinhter	0.228	*(5)*	0.319	*(6)*	**0.080**	*(1)*	**0.124**	*(1)*
4	*(3.75)*	McMaster	0.213	*(3)*	0.311	*(2)*	0.199	*(4)*	0.343	*(6)*
5	*(4.75)*	*Most frequent baseline*	0.239	*(5)*	0.316	*(2)*	0.362	*(6)*	0.345	*(6)*
6	*(5)*	aadisanghani*	0.213	*(3)*	0.311	*(2)*	0.491	*(6)*	BSL	*(9)*
6	*(5)*	BoN Appetite Team	0.231	*(5)*	0.414	*(9)*	0.228	*(4)*	0.272	*(2)*
8	*(6.5)*	NLP-ResTeam	0.291	*(8)*	0.326	*(6)*	0.418	*(6)*	0.345	*(6)*
9	*(7)*	cklwanfifa *	BSL	*(10)*	BSL	*(10)*	BSL	*(6)*	0.271	*(2)*
10	*(7.5)*	LPI-RIT	0.331	*(9)*	0.324	*(6)*	0.437	*(6)*	BSL	*(9)*
11	*(8.75)*	*Random baseline (BSL)*	0.352	*(10)*	0.499	*(10)*	0.367	*(6)*	0.497	*(9)*

Table 4: Overall Task B (perspectivist evaluation) results as an average of a system's rank across datasets. * indicates that no system description was available for the team.

explicit annotator features (`PromotionGo`). Finally, hybrid systems included `LPI-RIT`, which combined sentence-transformer embeddings with the DisCo architecture, and `Uncertain (Mis)Takes`, which modeled disagreement via semantic entropy over LLMs' generated explanations.

Towards Unified Approaches A clear difference from the previous edition (where teams tailored systems to each dataset) is that all participants who submitted for more than one dataset pursued general-purpose pipelines, aiming to capture patterns of disagreement across datasets with a unified approach. The majority instantiates a separate model for each dataset but follows the same pipeline, while others use a single model uniformly for all datasets.

Overall Rankings and Local Exceptions As a consequence of the shift away from dataset-specific solutions toward general-purpose pipelines, a clearer view of which approaches generalize better was enabled. In fact, differently from the previous edition, some systems ranked consistently among the best across all datasets and tasks. LLM-based systems with ICL secured the top positions in the overall leaderboard, with `OCP-ICL` and `DeMeVa` ranking first and second. However, fine-tuned transformer models, such as `twinhter` and `McMaster` were competitive and `twinhter` outperformed LLMs on smaller datasets *Par* and *VEN*. Moreover, the specific leaderboards revealed notable exceptions: teams that focused on tailored solutions for a single dataset, `PromotionGo` on Par and `Uncertain (Mis)Takes` on *VEN*, achieved first place locally.

Annotator information The majority of teams (six) used annotator information extensively, de-voting effort to find the optimal way for encoding annotator information. Two types of information were available: annotators' previous behavior and demographics. Some systems used annotator examples in in-context prompts to learn annotator views with LLMs (`Opt-ICL`, `DeMeVa`) or implicitly by training on each pair annotation-item or by passing annotator ID (`twinhter`, `NLP-ResTeam`, `LPI-RIT`). Demographics information usage was tested by `Opt-ICL`, `McMaster`, `twinther` and `NLP-ResTeam`. Notably, all of the best-performing systems incorporated some form of annotator information. Further details on the impact of annotator information are in Section 6.5.

Data Augmentation Strategies `Opt-ICL` post-trained LLMs using over 40 additional datasets. `NLP-ResTEAM` synthesized examples via paraphrasing and back-translation. `PromotionGo` applied extensive lexical (swap and reformulation) and translation-based augmentation. Further details on the impact of data augmentation are given in Section 6.5.

Task A vs Task B Leaderboard rankings for the two complementary tasks were largely similar. Not all systems attempted Task B, but of those that did, several derived the soft labels for Task A from the perspectivist labels for Task B. All three top-performing systems adopted this strategy, indicating that understanding annotator behavior contributes to overall prediction quality. Other systems adopted a multi-task strategy, using one output head for the soft label, the other for the perspectivist information.

6.3 Individual datasets results

CSC Two major observations stand out regarding *CSC*. The first relates to the role of demographic information. Most participating teams have used annotator information in their systems, regardless of their ranking. However, the winning team (`Opt-ICL`) reports through an ablation study that using demographic information did not significantly improve their results. This might be because the demographic information provided in *CSC* consists only of gender and age, with missing data, reported by the `twinther` team. Another observation is related to the importance of fine-tuning. While the most successful teams have used a combination of in-context learning while leveraging annotators information, two of these teams (`DeMeVa` and `McMaster`) report that fine-tuning RoBERTa has yielded comparable results to in-context learning with larger models. The winning team (`Opt-ICL`) also reports that dataset-specific fine-tuning was a crucial contributor to the results.

MP With respect to the other dataset included in the shared task, *MP* presented and additional challenge due to its multilinguality. This challenge was approached by leveraging pre-trained multilingual backbones (the majority of the teams) and/or by fine-tuning on the multilingual data. While the dataset is very metadata-rich, the top-2 best performing models for both tasks either did not incorporate annotators' sociodemographic data or only noticed a slight improvement when doing so. Fine-tuning was used for most systems. Submissions to Task A showed in general better results (with only two teams performing worse than the most frequent BSL), while only the winning team performed significantly better in Task B; we hypothesize this could be due to the large number of annotators in the dataset.

VEN & Par *VEN* and *Par* are two datasets with similar designs: (1) the same four annotators annotated all instances in the corpora, (2) all annotators are required to provide explanations to supplement their annotated labels. Due to these design similarities, we observe that the Perspectivist rankings of *Par* and *VEN* are extremely similar, with `twinhter` ranking first and `Opt-ICL` and `DeMeVa` in the tied second place. All three systems incorporated explanations into the context and demonstrated that models (both BERT-based ones and LLMs) can leverage this richer textual input to better understand labeling rationales and thus enhance performance. `DeMeVa` observed that including explanations in prompts helps better understand individual annotators' preferences, e.g., Ann3 for positive labels in *Par*. Additionally, `Uncertain (Mis)Takes` participated only and won first place in the *VEN* Task A using LLM-generated explanations and semantic entropy scores. Overall, explanations proved to be a valuable resource, either as explicit input features or as generated reasoning traces, and consistently contributed to stronger performance on datasets in both soft-label and perspectivist evaluations.

6.4 The new evaluation metrics: an assessment

The introduction of new evaluation metrics aimed to overcome the limitations of cross-entropy and to provide more reliable measures of model performance across diverse settings, including binary, multilabel, and ordinal-scale datasets based on the Likert scale. In practice, the Manhattan and Wasserstein distances offered intuitive and robust evaluations of soft label predictions, while the Error Rate and Average Normalized Absolute Distance enabled perspectivist assessments that better reflected annotator behavior and label structure.

For the multilabel scenario, evaluation relies on the Mean Absolute Manhattan Distance (MAMD) and the Mean Error Rate (MER).[6] These metrics have been designed to consider each label dimension independently, while simultaneously capturing the overall structure of label co-occurrence within an instance. By design, partially correct predictions incur a lower penalty than completely incorrect predictions. This allows the evaluation to reflect both the distribution of individual labels across annotators and their joint occurrence within the same instance, providing a nuanced measure of system performance in multilabel settings.

For datasets with ordinal labels (i.e., Likert-type scales), the Average Normalized Absolute Distance (ANAD) and the Average Wasserstein Distance (AWD) explicitly incorporate the ordinal nature of the labels. Unlike simple accuracy-based measures, these metrics penalize predictions proportionally to their deviation from the true label. In this way, systems are penalized less when producing outputs that are closer to the correct ordinal value, even if not exact, thereby providing a more faithful evaluation of performance on ordinal data.

Across all metrics, the lower bound remains consistent, with a score of 0 indicating a perfect match. A limitation, however, is that the upper bound is in some cases dataset-dependent (e.g., for the Wasserstein distance), which prevents direct comparisons across datasets.

6.5 Post-Submission Experiments and Ablation studies

Beyond their official submissions, all teams conducted supplementary analyses to gain a deeper understanding of their systems. These ablation studies and evaluations of alternative strategies enriched the competition with valuable insights and underscored the participants' commitment. The results demonstrated that the effectiveness of different approaches varied across datasets, reflecting both the specific characteristics of the data and the influence of the evaluation metrics employed.

[6]Further details are reported in Appendix A.

One major focus investigated was the role of annotator information. For LLM-based systems such as `OCP-ICL` and `DeMeVa`, provide in-context rater examples at inference time proved decisive: `OCP-ICL` showed that such examples drove large gains across datasets while demographics had negligible impact, and `DeMeVa` demonstrated that stratified selection of annotator examples improved consistency over random or similarity-based sampling. In contrast, for fine-tuned transformer-based models, annotator metadata and embeddings were more influential. `McMaster` found that demographic embeddings, particularly nationality and ethnicity, improved their RoBERTa system; `twinhter` observed stronger benefits from annotator metadata on small-annotator datasets; `LPI-RIT` reported that simple annotator ID tokens stabilized predictions; and `NLP-ResTeam` showed that label-style composite embeddings often outperformed demographics, though the best choice varied depending on the evaluation metric.

Ablation studies across papers revealed mixed effects of augmentation across teams. `OCP-ICL` found that post-training on over 40 dataset improved results only for *MP*, while for the other datasets was indifferent. `NLP-ResTeam` concluded that augmentation helped for small datasets (*Par* and *VEN*), while `PromotionGo` found that combining augmentation strategies worked best.

7 Conclusions

We are delighted that the third edition of the LeWiDi shared task continued to attract the attention of the community researching disagreement and variation in NLP. Again, we found that the participating teams engaged actively with the tasks, tackling interesting issues such as how best to use annotator information and the relation between soft-label modelling and perspectivist modelling.

Our hope is that the shared task and the datasets we released will stimulate further research in this area, by the participant groups and others. We believe that further thinking is still needed on issues such as the most appropriate form of evaluation for tasks in which human subjects express ordinal judgments, or the usefulness of modelling individual annotators or groups of annotators. To promote this, the *Codabench* page will remain open to submissions after the deadline so that researchers can continue test their models on the datasets.

Limitations

While this edition broadened the range of datasets, the scope remained restricted to text, leaving open the question of how disagreement-aware methods would perform in other modalities such as vision, speech, or multimodal tasks. Another open issue is that all annotators present in the test sets were also seen during training and development. As a result, the shared task did not directly evaluate systems' ability to generalize to unseen annotators, an ability that is likely to be critical in real-world applications.

Acknowledgments

The work of E. Leonardelli has been partially supported by the European Union's CERV fund under grant agreement No.101143249 (HATEDEMICS). The work of V. Basile was funded by the 'Multilingual Perspective-Aware NLU' project in partnership with Amazon Science. M. Pavlovic was supported by a PhD studentship from DeepMind. B. Plank acknowledges funding by the ERC Consolidator Grant DIALECT 101043235. M. Poesio was partially supported by NWO through the AINed Fellowship Grant NGF.1607.22.002 'Dealing with Meaning Variation in NLP'.

References

Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma, editors. 2022. *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. European Language Resources Association, Marseille, France.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2021. Whose opinions matter? Perspective-aware models to identify opinions of hate speech victims in abusive language detection. *CoRR*, abs/2106.15896.

Dina Almanea and Massimo Poesio. 2022. ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291, Marseille, France. European Language Resources Association.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Beata Beigman Klebanov and Eyal Beigman. 2009. From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, et al. 2024. Multipico: Multilingual perspectivist irony corpus. In *Proceedings of the conference-association for computational linguistics. Meeting*, volume 1, pages 16008–16021. Association for Computational Linguistics (ACL).

Silvia Casola, Soda Marem Lo, Valerio Basile, Simona Frenda, Alessandra Teresa Cignarella, Viviana Patti, and Cristina Bosco. 2023. Confidence-based ensembling of perspective-aware models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3496–3507, Singapore. Association for Computational Linguistics.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. 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 ACL*, 10:92–110.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, volume 1, pages 2164–2170. Association for Computational Linguistics.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Tommaso Fornaciari, Alexandra Uma, Massimo Poesio, and Dirk Hovy. 2022. Hard and soft evaluation of NLP models with BOOtSTrap SAmpling - BooStSa. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 127–134, Dublin, Ireland. Association for Computational Linguistics.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–14. ACM.

Ziyi Huang, Nishanthi Rupika Abeynayake, and Xia Cui. 2025. Promotiongo at lewidi-2025: Multilingual irony detection using l1 loss, data augmentation and ensemble learning. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.

Daniil Ignatev, Nan Li, Hugh Mee Wong, Anh Dang, and Shane Kaszefski-Yaschuk. 2025. Demeva at lewidi-2025: Modeling perspectives with in-context learning and label distribution learning. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. 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.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, volume 1, pages 1886–1895. 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.

Elisa Leonardelli, Alexandra Uma, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, and Massimo Poesio. 2023. Semeval-2023 task 11: Learning with disagreements (lewidi). *arXiv preprint arXiv:2304.14803*.

Soda Marem Lo 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*, Suzhou, China. Association for Computational Linguistics.

Héctor Martínez Alonso, Anders Johannsen, and Barbara Plank. 2016. Supersense tagging with inter-annotator disagreement. In *Proceedings of the 10th Linguistic Annotation Workshop*, pages 43–48. Association for Computational Linguistics.

Nguyen Huu Dang Nguyen and Dang Van Thin. 2025. twinhter at lewidi-2025: Integrating annotator perspectives into bert for learning with disagreements. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):219–252.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

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.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014a. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014b. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83. Association for Computational Linguistics.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1778–1789. Association for Computational Linguistics.

Marta Recasens, Ed Hovy, and M. Antònia Martí. 2011. Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

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 *3rd Workshop on Perspectivist Approaches to NLP, NLPerspectives 2024*, pages 84–94. European Language Resources Association (ELRA).

Filipe Rodrigues and Francisco C. Pereira. 2018. Deep learning from crowds. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 1611–1618.

Tomas Ruiz, Siyao Peng, Barbara Plank, and Carsten Schwemmer. 2025. Bon appetit team at lewidi-2025: Best-of-n test-time scaling can not stomach annotation disagreements (yet). In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Aadi Sanghani, Azadi Sarvin, 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.

Olufunke O. Sarumi, Charles Welch, and Daniel Braun. 2025. Nlp-resteam at lewidi-2025:performance shifts in perspective aware models based on evaluation metrics. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.

Mandira Sawkar, Samay U. Shetty, Deepak Pandita, Tharindu Cyril Weerasooriya, and Christopher Homan. 2025. Lpi-rit at lewidi-2025: Improving distributional predictions via metadata and loss reweighting with disco. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 614–622.

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting humorousness and metaphor novelty with Gaussian process preference learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5716–5728. Association for Computational Linguistics.

Taylor Sorensen and Yejin Choi. 2025. Opt-icl at lewidi-2025: Maximizing in-context signal from rater examples via meta-learning. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.

Taylor Sorensen, Benjamin Newman, Jared Moore, Chan Park, Jillian Fisher, Niloofar Mireshghallah, Liwei Jiang, and Yejin Choi. 2025. Spectrum tuning: Post-training for distributional coverage and in-context steerability. Preprint.

Ieva Staliūnaitė and Andreas Vlachos. 2025. Uncertain (mis)takes at lewidi-2025: Modeling human label variation with semantic entropy. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. 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.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft-loss functions. In *Proceedings of the 8th AAAI Conference on Human Computation and Crowdsourcing*, pages 173–177.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021b. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Yannick Versley. 2008. Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation*, 6(3):333–353.

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.

Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with disco. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4679–4695.

Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2018. Lexicosyntactic inference in neural models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4717–4724, Brussels, Belgium. Association for Computational Linguistics.

Appendix

A Evaluation Metrics for the Multilabel setting

In this section we outline how the adopted metrics were adapted to handle multilabel classification.

A.1 Multilabel Average Manhattan Distance (MAMD)

To account for the multilabel setting, the Average Manhattan Distance (AMD) was adapted into the Multilabel Average Manhattan Distance (MAMD) reported in equation 8. For each sample, the average Manhattan distance across all label-specific distributions is computed. The final score is then obtained as the average of such values over all samples.

$$AMD(i) = \frac{1}{L}\sum_{j=1}^{L}\sum_{k=1}^{n}\left|p_{i,j,k} - t_{i,j,k}\right| \tag{7}$$

$$MAMD = \frac{1}{N}\sum_{i=1}^{N} AMD(i) \tag{8}$$

With:

- N is the total number of samples,
- L is the number of labels (e.g., Entailment, Neutral, Contradiction for the VEN dataset),
- n is the length of each distribution,
- $t_{i,j,k}$ is the k-th value of the j-th target distribution for sample i,
- $p_{i,j,k}$ is the corresponding predicted value.

A.2 Multilabel Error Rate (MER)

The metric adopted for the perspectivist evaluation is the Multilabel Error Rate (MER), which quantifies the average dissimilarity between predicted and target label vectors across multiple samples. The Multilabel Error Rate (MER) is computed as the average of the average Error Rate values across all samples as shown in Equation 9:

$$\begin{aligned} MER &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{L}\sum_{j=1}^{L} ER(i)\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{L}\sum_{j=1}^{L} 1 - \frac{a - \sum_{k=1}^{a}|t_{i,j,k} - p_{i,j,k}|}{a}\right) \end{aligned} \tag{9}$$

Here,

- N is the total number of samples.
- L is the number of possible labels (i.e., the number of label-specific vectors to evaluate per sample, such as Entailment, Neutral, Contradiction).
- a is the length of a target or predicted vector (i.e., the number of annotators contributing to each label vector).
- $t_{i,j,k}$ is the k-th element of the j-th target vector for sample i.
- $p_{i,j,k}$ is the k-th element of the j-th predicted vector for sample i.

B Datasets specific leaderboards

CSC					
TASK A			TASK B		
	TEAM	WS		TEAM	MAD
1	Opt-ICL	0.746	**1**	Opt-ICL	0.156
1	DeMeVa	0.792	**2**	DeMeVa	0.172
3	McMaster	0.803	**3**	McMaster	0.213
3	aadisanghani	0.803	**3**	aadisanghani	0.213
5	twinhter	0.835	**5**	twinhter	0.228
6	BoN Appetit Team	0.928	**5**	BoN Appetit Team	0.231
7	*Most frequent BSL*	1.170	**5**	*Most frequent BSL*	0.239
8	harikrishnan_gs	1.295	**8**	NLP-ResTeam	0.291
9	NLP-ResTeam	1.393	**9**	LPI-RIT	0.331
9	LPI-RIT	1.451	**10**	*Random label BSL*	0.352
11	*Random label BSL*	1.543			

Table 5: Results for the CSC dataset

MP					
TASK A			TASK B		
	Team	MD		Team	ER
1	Opt-ICL	0.422	**1**	Opt-ICL	0.289
1	PromotionGo	0.428	**2**	DeMeVa	0.300
3	McMaster	0.439	**2**	McMaster	0.311
3	aadisanghani	0.439	**2**	aadisanghani	0.311
5	twinhter	0.447	**2**	*Most frequent BSL*	0.316
6	BoN Appetit Team	0.466	**6**	twinhter	0.319
6	DeMeVa	0.469	**6**	LPI-RIT	0.324
8	*Most frequent BSL*	0.518	**6**	NLP-ResTeam	0.326
9	LPI-RIT	0.540	**9**	BoN Appetit Team	0.414
9	NLP-ResTeam	0.551	**10**	*Random label BSL*	0.499
11	*Random label BSL*	0.687			

Table 6: Results for the MP dataset

Par					
TASK A			TASK B		
	Team	WS		Team	MAD
1	twinhter	0.983	**1**	twinhter	0.080
1	DeMeVa	1.120	**2**	Opt-ICL	0.119
1	Opt-ICL	1.200	**2**	DeMeVa	0.134
4	McMaster	1.605	**4**	McMaster	0.199
4	tdang	1.665	**4**	BoN Appetit Team	0.228
4	BoN Appetit Team	1.797	**6**	*Most frequent BSL*	0.362
7	aadisanghani	3.051	**6**	*Random label BSL*	0.367
7	NLP-ResTeam	3.136	**8**	NLP-ResTeam	0.418
7	*Most frequent BSL*	3.231	**8**	LPI-RIT	0.437
7	*Random label BSL*	3.350	**8**	aadisanghani	0.491
7	LPI-RIT	3.715			

Table 7: Results for the Par dataset

VEN					
TASK A			TASK B		
	Team	MMD		Team	MER
1	twinhter	0.233	**1**	twinhter	0.124
1	Uncertain Mis(Takes)	0.308	**2**	DeMeVa	0.228
3	BoN Appetit Team	0.356	**2**	Opt-ICL	0.270
3	DeMeVa	0.382	**2**	cklwanfifa	0.271
3	Opt-ICL	0.449	**2**	BoN Appetit Team	0.272
6	cklwanfifa	0.469	**6**	McMaster	0.343
7	*Most frequent BSL*	0.595	**6**	NLP-ResTeam	0.345
7	McMaster	0.638	**6**	*Most frequent BSL*	0.345
9	*Random label BSL*	0.676	**9**	*Random label BSL*	0.497
10	NLP-ResTeam	1.000			

Table 8: Results for the VEN dataset