NLPerspectives 2025

The 4th Workshop on Perspectivist Approaches to NLP

Proceedings of the Workshop

November 8, 2025

The NLPerspectives organizers gratefully acknowledge the support from the following sponsors.

Silver

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-350-0

Introduction

The NLPerspectives workshop brings together researchers and practitioners to explore how Natural Language Processing (NLP) systems can better capture, represent, and engage with diverse perspectives. Building on the prior workshops of NLPerspectives, this year's 4th edition workshop centered on the challenges and opportunities of representing multiple, and sometimes conflicting, viewpoints in annotation, modeling, and evaluation. Contributions spanned theoretical frameworks, empirical studies, and methodological innovations, highlighting both crowd-truth approaches to data annotation and novel techniques for integrating plurality into NLP models. The program featured research papers, a position paper, and hosted a shared task called Learning with Disagreement. This proceedings volume documents the range of ideas and findings presented at NLPerspectives, including the shared task, offering insight into the state of the field and charting directions for future work at the intersection of language technology, social values, and human-centered AI.

Until recently, language resources supporting many tasks in Natural Language Processing (NLP) and other areas of Artificial Intelligence (AI) have been based on the assumption of a single 'ground truth' label sought via aggregation, adjudication, or statistical means. However, the field is increasingly focused on subjective and controversial tasks, such as quality estimation or abuse detection, in which multiple points of view may be equally valid; subjectivity, indeed, is considered one of the main causes of Human Variation.

The fourth edition of the workshop builds upon these ideas to explore interdisciplinary synergies throughout the programme, starting with the keynote speaker Dr. Jose Camacho-Collados. In particular, this event is aimed at widening the discussed methodology to include not only current and ongoing work on collecting and labelling non-aggregated datasets, but also approaches to mining, modelling and inclusion of diverse perspectives in data, evaluation, and applications of multi-perspective Machine Learning models. In addition, it involved techniques from social science and Human-Computer Interaction, such as participatory approaches and how they can be implemented at all stages of the supervised learning pipeline.

Working with language as data poses unique challenges. Words and their definitions evolve over time, and even within the same time period, interpretation of language is contextual, influenced by the cultural, geographic and linguistic environment of people. Bringing interdisciplinary approaches from Feminist theories, Critical Discourse Analysis, and indigenous epistemologies, among others, to Computational Linguistics can provide valuable methods for working with the subjectivity and uncertainty of language data.

Besides the community of data perspectivism and human label variation, we expect the workshop to attract researchers and industry practitioners, as happened in previous editions, interested in learning from disagreement, personalization, and participatory design. For the first time, the workshop has hosted a shared task, Learning with Disagreement (Le-Wi-Di), which explores new approaches to modelling and evaluation of perspectivist data.

The Shared Task on Learning with Disagreement (LeWiDi) aims to raise the visibility of the challenge of variation in interpretation, and to encourage the community to engage with the problem. The objective of the shared task is to provide a unified testing framework for learning from disagreements and evaluating with such datasets. Two editions of the shared task were organized as part of SEMEVAL: in 2021, focused on ambiguity in language and vision; and in 2023, focused on disagreement in subjective tasks. These shared tasks created benchmarks that have since been widely downloaded, and started an ongoing discussion on how to evaluate perspectivist models. This new edition, co-located within the workshop, differs from the previous ones in: the inclusion of two new classification tasks in which disagreements are prevalent, one on irony, and one on Natural Language Inference; the addition of two generative tasks, paraphrasing and summarization; testing new approaches to evaluation, inspired by recent research.

In its fourth edition, the workshop accepted 14 submissions—13 research papers and one stance paper, with one designated as non-archival. Additionally, 9 teams submitted successful entries to the shared task.

One of the primary motivations for the pursuit of perspectivism is expanding the number of voices represented in NLP datasets, modelling, and evaluation. The workshop therefore has diversity and inclusion as one of its central tenets.

As at previous events, this edition of NLPerspectives will host presenters, panellists, and keynote speakers representing diverse demographics, research backgrounds, and career stages. Our organizing and programme committee also includes a balance of geographical locations, genders, and career stages.

For access, the most important contributions and available resources are listed and reported in the manifesto page of perspectivist data (https://pdai.info/) and proceedings and available slides of previous editions are online on our website: https://nlperspectives.di.unito.it/.

To encourage people coming from various countries to join the discussion, the workshop is planned to be in a hybrid format. As in the previous editions, we will provide financial aid to scholars (especially students) who may not otherwise be able to attend.

Organizing Committee

Organizing Committee

Gavin Abercrombie, Heriot-Watt University
Valerio Basile, University of Turin
Davide Bernardi, Amazon Alexa
Shiran Dudy, Northeastern University
Simona Frenda, University of Turin
Sara Tonelli, Fondazione Bruno Kessler

Shared Task Organizers

Silvia Casola, LMU Munich, Germany
Elisabetta Fersini, University of Milan Bicocca, Italy
Elisa Leonardelli, Bruno Kessler Foundation, Italy
Maja Pavlovic, Queen Mary University, London, UK
Massimo Poesio, Queen Mary University, London, UK; University of Utrecht, Netherlands
Barbara Plank, LMU Munich, Germany
Alexandra Uma, Builder AI, UK
Siyao Peng, LMU Munich, Germany

Program Committee

Chairs

Gavin Abercrombie, Heriot Watt University
Valerio Basile, University of Turin
Shiran Dudy, Northeastern University
Simona Frenda, Heriot-Watt University
Sara Tonelli, FBK

Program Committee

Riza Batista - Navarro, Department of Computer Science, The University of Manchester
Davide Bernardi, Amazon
Agostina Calabrese, The University of Edinburgh
Silvia Casola, LMU Munich
Amanda Cercas Curry, Bocconi University
Teddy Ferdinan, Wrocław University of Science and Technology
Samuel Goree, Stonehill College
Marco Guerini, Fondazione Bruno Kessler
Nancie Gunson, Heriot-Watt University
Annette H a u t l i - J a n i s z, University of Passau
Cassandra L. Jacobs, University at Buffalo
Aiqi Jiang, Heriot Watt University
Kamil Kanclerz, Wrocław University of Science and Technology
Anna Koufakou, Florida Gulf Coast University
Sofie Labat, Ghent University
Marta Marchiori Manerba, Università di Pisa
Michele Mastromattei, University of Rome Tor Vergata
Massimo Poesio, Queen Mary University of London and University of Utrecht
Julia Romberg, GESIS – Leibniz Institute for the Social Sciences
Pratik Sachdeva, University of California, Berkeley
Manuela Sanguinetti, University of Cagliari, Department of Mathematics and Computer Science
Erhan Sezerer, Amazon Research
Tiago Timponi Torrent, Federal University of Juiz de Fora

Keynote Talk
Cultural Awareness in Multilingual Language Models - A Perspectivist Personal Perspective

Jose Camacho Collados
Cardiff University
2020-11-08 12:30:00 – Room: **Room 1**

Abstract: Language models have become ubiquitous in NLP and beyond. In particular, the new wave of large language models (LLMs) are increasingly used to communicate and solve practical problems in many languages and countries, and by an increasingly diverse set of users. However, even though there is no doubt that these models open up plenty of opportunities, there are important issues and research questions that arise when it comes to LLMs and their application in different languages and cultures. For instance, the language coverage in language models drastically decreases for less-resourced languages and as such, their performance. And not only the general performance is affected, but general-purpose LLMs may be implicitly biased to specific cultures and languages depending on their underlying training data.
In this talk, I will discuss how language models reflect on cultural diversity, including potential shortcomings and how language coverage and cultural awareness may be intrinsically intertwined. I will also share some lessons learned based on recent research in this area – in particular, I will focus on the development of BLEnD, a large effort to develop a cultural benchmark of everyday knowledge for dozens of languages and countries.

Bio: Jose Camacho-Collados is a UKRI Future Leaders Fellow and Professor at the School of Computer Science of Cardiff University, where he co-founded the Cardiff Natural Language Processing group (Cardiff NLP). Before joining Cardiff University, he completed his PhD in Sapienza University of Rome and was a Google AI PhD Fellow.
Jose has worked in multiple NLP areas with a particular focus on semantics, multilinguality and computational social science with an interdisciplinary perspective. In this area, he has been developing specialised and efficient NLP models for social media applications, such as TweetNLP and related efforts. His work has received several recognitions, including awards at top NLP conferences, or the 2023 AIJ Prominent Paper Award. He is also the co-author of the "Embeddings in Natural Language Processing" book.

Table of Contents

A Disaggregated Dataset on English Offensiveness Containing Spans

Pia Pachinger[1], **Janis Goldzycher**[2], **Anna M. Planitzer**[3], **Julia Neidhardt**[1], **Allan Hanbury**[1]
[1]Faculty of Informatics, TU Wien. Correspondence: pia.pachinger@tuwien.ac.at
[2]Department of Computational Linguistics, University of Zurich
[3]Political Commmunication Research Group, University of Vienna

Abstract

Toxicity labels at sub-document granularity and disaggregated labels lead to more nuanced and personalized toxicity classification and facilitate analysis. We re-annotate a subset of 1983 posts of the *Jigsaw Toxic Comment Classification Challenge* and provide disaggregated toxicity labels and spans that identify inappropriate language and targets of toxic statements.

Manual analysis shows that five annotations per instance effectively capture meaningful disagreement patterns and allow for finer distinctions between genuine disagreement and that arising from annotation error or inconsistency. Our main findings are: (1) Disagreement often stems from divergent interpretations of edge-case toxicity (2) Disagreement is especially high in cases of toxic statements involving non-human targets (3) Disagreement on whether a passage consists of inappropriate language occurs not only on inherently questionable terms, but also on words that may be inappropriate in specific contexts while remaining acceptable in others (4) Transformer-based models effectively learn from aggregated data that reduces false negative classifications by being more sensitive towards minority opinions for posts to be toxic. We publish the new annotations under the CC BY 4.0 license.

Content warning: This paper contains examples of offensive language to describe the data.

1 Introduction

The amount of toxic content on the internet is increasing and causes harm. Especially implicit offensiveness still often goes undetected (Zhang et al., 2022). Given the absence of widely accepted definitional distinctions between the terms *offensiveness* and *toxicity* (Pachinger et al., 2023), we use them interchangeably. Creating effective automated content moderation systems requires two key elements: nuanced understanding of online norm violations and incorporation of diverse opinions on what content warrants moderation.

The Need for Perspectivist Offensiveness Detection Perceptions of what content is harmful depend on individual, contextual, and geographical factors (Hershcovich et al. 2022; Sandri et al. 2023; Abercrombie et al. 2023 i.a.). Researchers report disagreements in the annotation of toxicity explained by previous annotations by the same annotator (Wich et al., 2020), sociodemographics (e.g. Kocoń et al. 2021; Aroyo et al. 2023) beliefs (Sap et al., 2022), and moral values and geocultural factors (Davani et al., 2023). The human perception of what constitutes harmful language is inherently reflected in the classifiers trained on human-labeled data to identify toxic speech. If perceptual differences in what makes a remark toxic are not taken into account when training these systems, what is considered toxic may be disproportionately influenced by the societal majority or, in the worst-case scenario, by an arbitrary group of annotators. Therefore, a one-size-fits-all approach to content moderation is unable to account for the diverse needs of different users (Cresci et al. 2022; Plank 2022; Jhaver et al. 2023 i.a.).

Perspectivism in machine learning (ML) represents a paradigm shift from consensus-based labeling to embracing annotator diversity. For each data instance, multiple labels are collected and, where possible, maintained as disaggregated annotations alongside annotator metadata throughout the ML pipeline (Cabitza et al., 2023). This paradigm aligns with descriptive annotation approaches, which embrace annotator subjectivity as a meaningful signal rather than noise. Descriptive annotation enables the modeling of diverse beliefs and perspectives, contrasting with prescriptive annotation that enforces uniform interpretation through strict guidelines (Röttger et al., 2022).

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 1–14
November 8, 2025 ©2025 Association for Computational Linguistics

The Need for Offensiveness Annotations Beyond Classes In (semi-) automated content moderation, explainability contributes to a greater understanding and trust of users (Molina and Sundar, 2022) and content moderators (Bunde, 2021). Experienced moderators work more efficiently when provided with structured explanations that pinpoint harmful content and articulate why it violates community standards (Calabrese et al., 2024). In text annotation, a span refers to a contiguous sequence of tokens within a document that is marked and labeled. Rather than annotating entire documents or sentences, spans allow annotators to identify and categorize specific portions of text. Annotators can pinpoint exactly which parts of a text exhibit the phenomenon of interest. Text classification models can learn from the specific linguistic features within marked spans, leading to more accurate predictions about similar text segments. Different spans within the same text can receive different labels, capturing the complexity of real-world documents where multiple phenomena may coexist, facilitating error analysis, model debugging, and model explainability (Lyu et al., 2024).

Main Contributions By re-annotating 1,983 comments from the Jigsaw Toxic Comment Classification dataset (cjadams et al., 2017), 1,561 of which received annotations from four or five annotators, we classify toxic utterances at the post level and we identify spans comprising the targets of toxic utterances, and spans comprising vulgar expressions while maintaining disaggregated labels from multiple annotators. We find that high disagreement among five annotators meaningfully signals cases where subjective elements influence the perception of toxic statements.

Our analysis demonstrates that disagreements arise from divergent interpretations of borderline toxic content. We find substantial disagreement on toxicity classifications involving non-human targets. Further, when evaluating whether a span consists of inappropriate language, disagreement occurs not only on inherently questionable terms, but also on words that may be inappropriate in specific contexts while remaining acceptable in others. Lastly, while annotators generally recognize targets of toxic language, repeated target mentions within comments can pose hurdles for human annotators and for ML approaches to extracting spans comprising targets.

Despite differences in definitions and annotation approaches, we find broad agreement between the Jigsaw annotations and our own. Additionally, experiments show that transformer-based models effectively learn from aggregated data, which reduces false negative classifications by being more sensitive towards minority opinions for posts to be toxic. We release the new annotations under CC BY 4.0 licensing with the underlying comment text being governed by Wikipedia's CC-SA-3.0 [1].

2 Related Work

Recent advances in toxicity annotation include the development of granular labeling frameworks, publication of disaggregated annotation datasets, and diversification of annotator demographics.

Disaggregated Toxicity Annotations In recent years, researchers have started to publish disaggregated toxicity annotations. For example, Kumar et al. 2021 release a labeled toxicity dataset that contains 107,620 texts and annotations by 17,280 annotators. Another example is the dataset published by Kennedy et al. 2020, which contains 50,000 texts and annotations by 11,000 Mechanical Turk workers. See Frenda et al. 2024 for more perspectivist datasets on online toxicity.

Toxicity Annotations Beyond Classes Additionally, there has been a surge in datasets related to offensive text detection with span and free-text annotations. Existing data with annotated spans include spans of the targets of offensive statements (Calabrese et al., 2022a; Zampieri et al., 2023; Pachinger et al., 2024), the spans contributing to the offensiveness label (Mathew et al., 2021; Pavlopoulos et al., 2021), spans comprising a violation of a moderation policy (Calabrese et al., 2022a), and the spans comprising vulgar language (Pachinger et al., 2024). More recently, free-text annotations related to toxicity labels were released (Sap et al., 2020; Zhang et al., 2022; Zhou et al., 2023). The spans and free text can be used to create inherently faithful explain-then-predict methods for offensive text detection (Kim et al., 2022; Zhang et al., 2022; Zhou et al., 2023). Furthermore, they can be used to create post-hoc explanations (Risch et al., 2020).

[1] https://huggingface.co/datasets/pia-p/disaggregated_offensiveness
https://github.com/pi-pa/disaggregated_offensiveness
https://web.ds-ifs.tuwien.ac.at/disaggregated_offensiveness

Annotator Populations in Toxicity Datasets Toxicity annotation studies on English toxic content typically rely on annotators from English-speaking countries, particularly the United States. Zhou et al. 2023 engage native English speakers for labeling offensive content, while Sap et al. 2020 recruit annotators exclusively from the U.S. and Canada. Calabrese et al. 2022b similarly restrict their pool to English-speaking countries. While Zhang et al. 2022 broaden their criteria to include anyone with English proficiency, they do not systematically ensure demographic diversity. This geographic concentration raises questions about the generalizability of toxicity judgments, particularly given that English is the dominant lingua franca spoken by a wide variety of people and perceptions of harmful content vary significantly across countries.

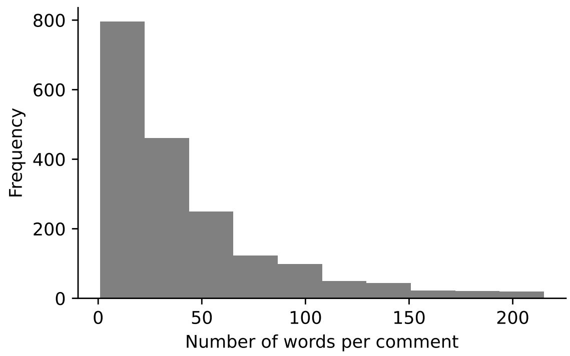


Figure 1: The distribution of the comment lengths of all comments shorter than the 95% percentile of the comments.

3 Data Source

We source the data from the *Toxic Comment Classification Challenge* from Jigsaw. It contains Wikipedia comments which have been labeled by human raters for toxic behavior. Annotated labels in the dataset are *toxic*, *severe toxic*, *obscene*, *threat*, *insult*, *identity hate*. The data is published under the CC0 License, with the underlying comment text being governed by Wikipedia's CC-SA-3.0. We source 1700 not toxic comments. Additionally, since we are interested in nuanced toxicity cases, we select comments labeled as *toxic* without any additional toxic categories and sample 1700 of these. We further select *insults* without any additional toxic categories, excluding comments with more severe or multiple toxic labels. From this pool of data, 1983 posts are annotated. Figure 1 shows the distribution of comment lengths in our dataset, excluding outliers.

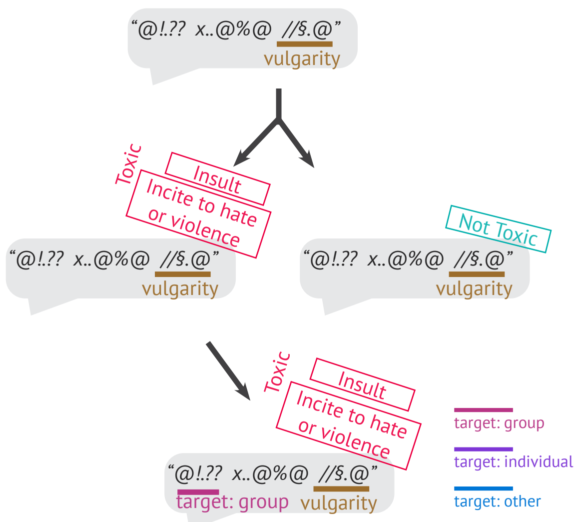


Figure 2: The annotation strategy for this dataset

4 Annotation Schema

We adopt the annotation schema used for the German AustroTox dataset (Pachinger et al., 2024), making the two datasets containing different data sources and cohorts of annotators compatible and allowing for multilingual analyses. Observe the annotation strategy in Figure 2. We classify each comment as insult, incite to hate or violence or not offensive. Since we do not source for incites to hate or violence in the Jigsaw dataset, the number of posts labeled as *Incite to hate or violence* in our dataset is limited. Therefore, we create an *Offensiveness / Toxicity* class by merging classes *Insult* and *Incite to hate or violence*. For non-offensive and offensive comments, we annotate vulgarities since both offensive and non-offensive posts can contain vulgarities. For offensive posts, we additionally annotate the targets of the offensive statement and the type of target. If the target is only mentioned via a pronoun, we annotate the pronoun as the target. Adopting a definition of *vulgarity* similar to that employed by Risch et al. (2021), We use the following definitions for classes and spans:

Insult An insult pursues the recognisable goal of disparaging the addressee or the object of reference.

Incite to Hate or Violence An incite to hate or violence against a person or a group of people. It is often hard to draw the line between insults and incites to hate, as insults always somewhat incite hate. For this annotation task, we define insults to be less severe than incites to hate or violence.

Offensive / Toxic An insult or an incitement to hate or violence.

Not Offensive / Not Toxic Not an insult nor an incite to hate or violence.

Vulgarity Rude, obscene, foul or boorish language that is inappropriate for civilized discourse.

Target Group The target of an offensive post is a group of persons or an individual insulted based on shared group characteristics.

Target Individual The target of an offensive post is a single person, not insulted based on shared group characteristics.

Target Other The target of an offensive post is not a person or a group of people.

We position our work within the descriptive annotation paradigm, recognizing that toxicity perception contains inherent subjective elements (Röttger et al., 2022). Determining whether someone intends to disparage a target, distinguishing between hate incitement and mere insults, and identifying when comments cross into inciting violence all involve subjective judgment calls. Similarly, the threshold for what constitutes inappropriate passages in civilized conversation varies across readers and contexts. Among our annotation tasks, identifying the target of offensive statements and categorizing the target type represents the most objectively answerable component. Our definition of vulgar language is deliberately expansive, extending beyond conventional sexual, scatological, and religious profanity to include any language inappropriate for civilized discourse. We recognize that determinations of vulgarity are both context-dependent and inherently subjective, as language deemed acceptable in casual forums may prove inappropriate in structured, goal-oriented discussions and edge-case acceptability varies by reader.

5 Annotation Campaign

We conduct the annotation with master's students in data science. Thirty percent of annotators are registered as female in this course, though this institutional designation may not reflect their actual gender identity. The majority of annotators are between 19 and 26 years old, with all demonstrating at least B2-level English proficiency. Most annotators originate from Eastern Europe.

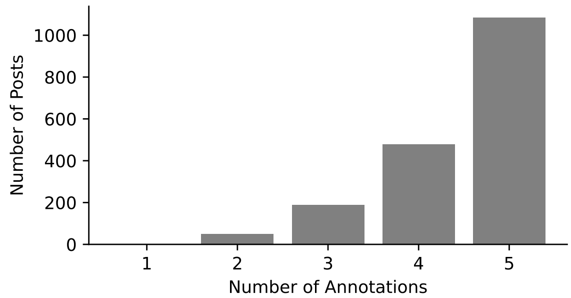


Figure 3: The number of annotators per post.

We acknowledge the importance of representing the perspectives of a population by including diversity in the annotator pool (Clemmensen and Kjærsgaard, 2022). Our annotator pool's homogeneous sociodemographics, constrained by resources and participants in the course, present a limitation in our data. On the other hand, the demographic composition in our dataset contrasts with typical NLP annotation practices, where annotators are predominantly recruited from English-speaking countries. However, English-language online discussions reach global audiences with diverse cultural backgrounds and perspectives on toxicity. We therefore argue that our annotator pool provides valuable demographic diversity to the current landscape of toxicity datasets. Furthermore, opinions on what constitutes toxicity are influenced by a multitude of factors beyond just sociodemographics. While we do not explicitly capture annotator characteristics for individual annotators, these factors are implicitly reflected in the disaggregated annotations (Geva et al., 2019; Wich et al., 2020). Consequently, we view our dataset as a valuable addition to the broader collection of resources, capturing user perspectives in various ways.

The annotation campaign was reviewed by the ethics committee of the first author's institution. Each annotator annotates about 200 comments, which takes approximately 1.5 to two hours. The dataset contains a higher proportion of offensive comments than the typical distribution in a user forum, but we only source comments with labels *toxic* or *insult* and exclude more severe labels. The annotators are explicitly informed that they have the option to cease annotation if they feel overwhelmed by the task without facing consequences, and about the publication of the data, and they receive comprehensive compensation through course

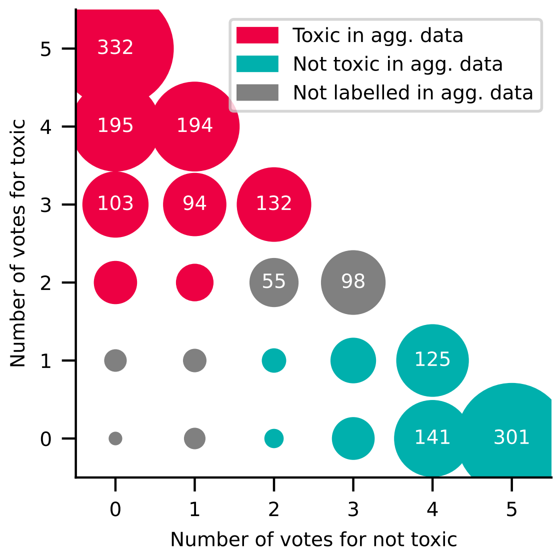


Figure 4: The disagreement in the toxicity annotations. The colors denote the label of the aggregated dataset.

credits for their efforts. Figure 3 shows the number of annotators per comment. The vast majority of posts is annotated by five annotators.

6 Disagreements in the Annotations

We calculate inter-annotator agreement using Krippendorff's Alpha across all annotation categories and conduct manual analysis to identify factors correlating with disagreement at both the post and span levels.

Disagreements in Offensiveness Annotations Figure 4 visualizes the distribution of annotator agreement on offensiveness labels on the post-level. Most posts show complete annotator agreement on whether they are toxic or non-toxic. We report a Krippendorff's alpha of 0.57 for binary offensiveness classification. While this falls below the $\alpha \geq 0.667$ threshold typically recommended for tentative conclusions in prescriptive annotation paradigms (Krippendorff, 2004), it aligns with values reported in comparable toxicity detection studies: Sap et al. (2020) report a Krippendorff's Alpha of 0.51, and Wulczyn et al. (2017) report an Alpha of 0.45.

To understand the origins of disagreements in posts where annotators highly disagree, we manually analyze 60 posts that are annotated as toxic by 3 annotators and as non-toxic by 2 annotators. Table 1 presents all factors related to disagreement that we identify through this analysis. We identify subjective elements in 46 of these posts, primarily involving grey-zone or nuanced toxicity that falls into borderline categories. Observe an example where toxicity is subjective and open to interpretation: "*Quiet, you. Whether you are a troll or not is irrelevant - your edits are trolling, are uncivil, and are ridiculous.*"

Additionally, 12 posts contain toxicity directed at non-human targets. We find that toxicity toward non-human entities is typically perceived as less severe and therefore more open to interpretation. The example from above illustrates such a case. In this example, the insult has two potential targets: the person being addressed or their edits, which are labeled as "ridiculous." Since the criticism targets the person's edits rather than the person themselves, readers may perceive the insult as less severe. Despite viewing our annotation guidelines as descriptive, in some cases, it is possible to definitively say that an utterance is an insult or incitement to hate or violence. We identify 7 such clear-cut cases among the 60 posts we analyze. 5 of the 60 high-disagreement posts we analyze contain quotes with toxic language, such as: "*Vandalism. How's about I stick ""W*nkers Haa HAAa"" in your block log?*". We obfuscate vulgarities and specific targets in this paper.

Subjective whether insult	41
Non-human target	12
Definitely insult	6
Subjective whether incite to h. or v.	5
Toxicity in quote	5
Calls target to leave conversation	5
Definitely not toxic	4
Lack of context	4
Particularly long post	4
Toxic against self	3
Spam (not in our toxicity definition)	2
Definitely incite to hate or violence	1

Table 1: Factors related to disagreements in the offensiveness classifications we identified in 60 comments

Disagreements in Vulgarity Annotations Figure 5 displays disagreement patterns in span annotations comprising vulgarities. We visualize only spans that at least one annotator marks as vulgar. Most spans perceived as vulgar receive annotations from only one or two annotators. Correspondingly, vulgar span annotations achieve a Krippendorff's Alpha of 0.05, indicating substantial disagreement among annotators.

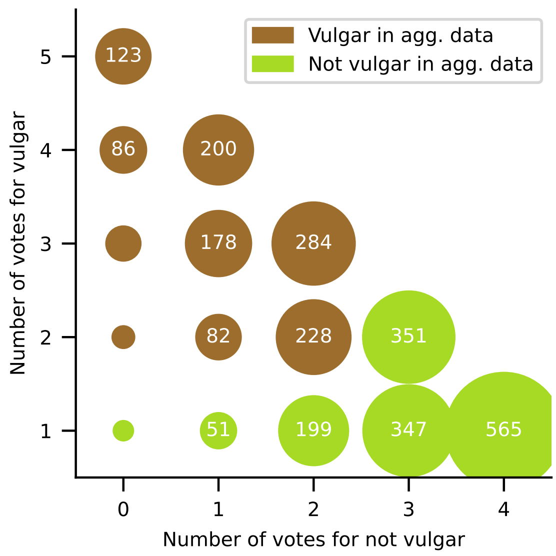


Figure 5: The disagreement on the spans comprising vulgarities. The colors denote the label of the aggregated dataset.

To understand this low agreement, we manually evaluate 50 spans that receive 3 votes for vulgarity while remaining unannotated by the other 2 annotators. Our analysis reveals that contextual factors critically influence vulgarity perception. Out of these 50 spans, 10 spans are unambiguously vulgar. Take, for example, the post "*Hey, I said it was ""a"" seat, not ""the"" seat, you dumb motherf#$ker!!*". We classify the span "*motherf#$ker*" as unambigously vulgar. Further, 17 of the 50 spans are subjectively vulgar (where the word's inherent nature is debatable). Recall that our definition of vulgar language is expansive, extending beyond conventional sexual, scatological, and religious profanity to include any language inappropriate for civilized discourse. Taking the previous example, consider the span "*dumb*". While this term carries multiple meanings, within this particular setting, it functions as a clear synonym for stupid. The acceptability of such language varies significantly among individuals. We hypothesize that the imagined forum context plays a critical role, where informal forums may accept different languages than goal-oriented discussion spaces.

Moreover, 15 spans demonstrate context-dependent vulgarity (where the same word becomes vulgar or benign depending on usage). Take, for example, the post "*He jailed 50 000 murders, thiefes, rapers, criminals, drug-sellers, prostitutes and many more in only 9 month what you couldn´t do in your 6000 years of history. Stupid losers*" The span "*Stupid*" falls under both categories. Whether it is appropriate for civilized discourse depends on the reader. But, additionally, the word is used in direct speech and against a group that might be vulnerable, which might make it appear more inappropriate to some than in other settings. Lastly, 6 spans are clearly non-vulgar, and one is incomprehensible. This distribution demonstrates that vulgarity annotation involves both lexical ambiguity, whether words are inherently vulgar, and contextual complexity, whether usage renders otherwise benign words inappropriate.

Disagreements in Target Annotations Figure 6 displays the distribution of annotator votes for spans constituting targets of toxic statements. The visualization includes all target types and only spans that at least one annotator identifies as a target. Most target spans receive annotations from only one annotator. We calculate disagreement based on annotators who label posts as toxic and obtain a Krippendorff's alpha of -0.05 for target annotations, indicating substantial disagreement similar to vulgarity span annotations. Identical spans labeled as different target types are treated as disagreements.

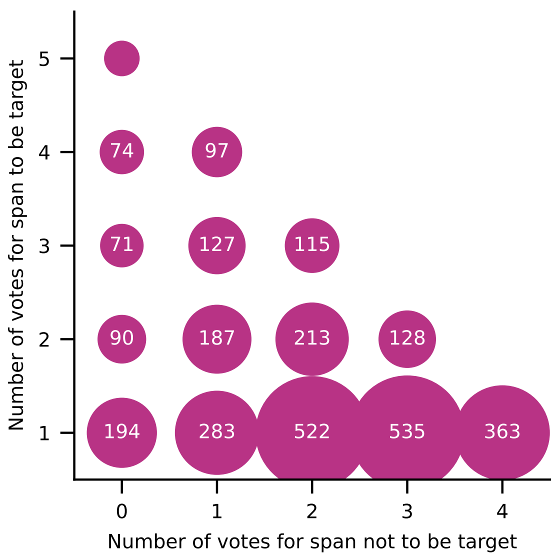


Figure 6: The disagreement on the spans comprising targets

We analyze 30 posts with 5 votes identifying a span as a target. All except one are correctly annotated. In 10 posts, the target appears multiple times, and in 2 cases, other potential targets appear in the data. We further analyze 30 spans with 3 votes for target classification and 2 votes against.

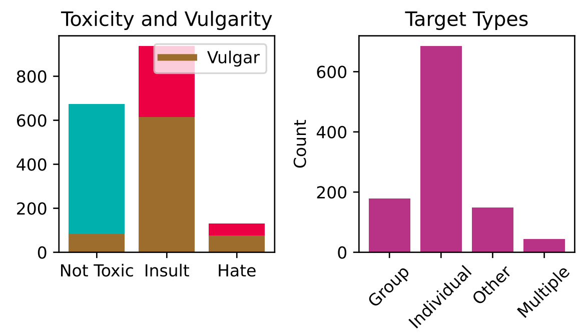


Figure 7: Post-level classes defined by the classes and spans labeled in the aggregated data.

All except one span are indeed targets of toxic remarks. In 16 cases, the target appears multiple times in the post, and in 7 cases, multiple potential targets are mentioned.

In summary, the frequent repetition of targets within posts and target mentions as pronouns creates annotation challenges. This repetition creates inconsistency; some annotators mark the target closest to the most toxic passage, while others mark its first appearance in the post. We advise authors to provide clear instructions for annotating targets of toxic statements, given their highly diverse manifestations.

7 Disagreement Between the Jigsaw Dataset and our Dataset

To enable meaningful comparisons with both the Jigsaw dataset annotations and the German AustroTox (Pachinger, 2024) dataset using a different data source but shared annotation framework, we aggregate the data using an approach that reduces false negative classifications by being more sensitive towards minority opinions for posts to be toxic.

Data Aggregation We adopt the same aggregation strategy as the AustroTox dataset for both classes and spans. This approach prioritizes avoiding false negatives for minority perspectives by creating broader decision boundaries. Specifically, we exclude examples with high disagreement that lean slightly toward non-toxic, while labeling examples with disagreement that lean slightly toward toxic as *toxic*. Figure 4 illustrates how different combinations of toxic versus non-toxic vote counts are labeled for post-level offensiveness classification. In especially, we label posts as non-toxic only when they receive at most one toxic vote and at least two non-toxic votes. We discard posts with less than two votes for one class, posts with 2 votes for a post to be toxic and 2 or 3 votes for a post to be not toxic. The remaining posts are labeled as offensive. The aggregated dataset results in a Krippendorff's alpha of 0.64. This value is higher than for the disaggregated data due to the fact that we discard instances with high disagreement and viewer perceptions of toxicity.

Figure 5 shows how we aggregate the spans comprising vulgarities. We label spans as vulgarities if they are annotated by at least two annotators and are not left out by more than two annotators. Spans in the comments comprising the different target types are annotated by majority voting of those who labeled the comment as offensive. If two spans receive a majority for a target span, both are annotated as the respective type of target. We combine the aggregated post-level classifications (*offensive* and *not offensive*) with span annotations to create fine-grained categories. This allows us to identify which types of toxicity are most prevalent in the dataset according to broad annotator consensus. Figure 7 shows the distribution of these categories and the frequency of different target types appearing in toxic utterances.

Disagreement Between the Jigsaw Dataset and our Dataset Figure 8 compares the label distributions between the Jigsaw dataset and our aggregated dataset. The majority of posts classified as offensive in our dataset correspond to the labels *toxic* or *insult* in the Jigsaw dataset. While the definition of toxicity differs between the two datasets, this broad alignment suggests general agreement among majority opinions regarding what consti-

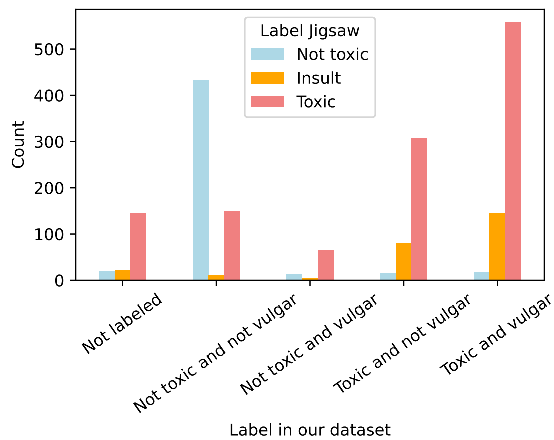


Figure 8: Labels in our vs. labels in the Jigsaw dataset.

tutes offensive content. In the area of toxic comment classification and hate speech detection, the labels *profane* and *vulgar* are often used similarly. Comments labeled as *not toxic and vulgar* in our dataset are categorized as *toxic* in the Jigsaw dataset, yet they do not receive the label *profane* in the Jigsaw dataset, suggesting that we use a broader definition for vulgarity than was used for the original data. The most significant divergence involves 155 posts that our dataset labels as *not toxic* and *not vulgar* but that Jigsaw categorizes as *toxic*.

We manually review 50 of the 155 posts that our dataset classified as non-toxic and non-vulgar, but that the Jigsaw dataset labeled as toxic. We find 14 comments to be genuinely non-toxic, while 12 fall into a subjective gray area where toxicity judgments could reasonably vary. An additional 8 comments appear toxic under broader definitions than ours, for instance, posts containing accusations of lying or spam that we would not classify as toxic, and one comment is toxic according to our definition. Further, 8 comments contain vulgar language, with 5 additional posts falling into a subjective category for vulgarity. Five comments use slang or specialized abbreviations that could lead to different interpretations across annotators. Lastly, 4 posts lack sufficient context for reliable assessment.

In summary, our analysis reveals three primary sources of disagreement between the datasets. Subjective interpretation challenges are the most prevalent issue, affecting 26 of the 50 comments. These include posts lacking sufficient context, containing specific language or slang, or falling into gray areas for toxicity or vulgarity assessment where annotators can reasonably disagree. Definitional differences explain 8 cases where comments appear toxic under Jigsaw's broader criteria but not ours. Potential labeling inconsistencies exist as well, with 8 comments appearing vulgar despite being labeled non-vulgar in our dataset, and 14 comments seeming non-toxic despite Jigsaw's toxic classification.

8 Classification Experiments

We conduct experiments on the aggregated data in order to show that the labels provide learnable signals. We conduct experiments on binary offensiveness classification, token classification of vulgar passages, and passages constituting the different types of targets. We fine-tune and evaluate encoder-based models, and we evaluate the few-shot performance of decoder-based models in a 10-fold cross-validation setting.

We fine-tune encoder-based models on all three tasks independently. This means that the target detection task inherently includes offensiveness classification, as we only annotate targets of offensive statements. We choose ELECTRA Large[2] (Clark et al., 2020) and Roberta Large[3] (Liu et al., 2019) for our experiments, as they exhibit good performance at the SemEval-2023 task 10: explainable detection of online sexism (Kirk et al., 2023). Additionally, we assess the in-context learning performance of the following large language models: GPT 3.5(*gpt-3.5-turbo-1106*) (Ouyang et al., 2022), GPT 4 (*gpt-4-1106-preview*) (et al., 2024), Llama 3[4] (AI@Meta, 2024) and Mistral [5] (Jiang et al., 2023). We use the same prompts as Pachinger et al. 2024. They contain an offensiveness definition, the post to be classified, and for the five-shot scenario, randomly sampled annotated example posts. Due to limited performance for Mistral and LLama3, we adjust the prompt, requiring them to respond with only *0* or *1*, and we define the token with the higher logit as the Llama3 and Mistral's prediction. We tokenize the spans generated by the generative models with the Roberta tokenizer. We compute the Micro F1 by adding up the values of the confusion matrix for the three target classes using Nakayama's (2018) framework.

Results Table 2 presents the evaluation results. Several important limitations should be noted, particularly for the in-context learning experiments. We did not fine-tune the decoder-based models or perform prompt optimization, meaning the decoder-based model results represent a lower bound of achievable performance rather than optimal outcomes. These results demonstrate that the models can achieve reasonable performance on several tasks, which serves our primary objective. The models perform better on our dataset than on AustroTox, which results in a Binary and F1 score of 0.76 for offensiveness classification and 0.71 for vulgarity token-classification, and a Micro F1 score of 0.24 for target classification (Pachinger et al.,

[2] https://huggingface.co/google/electra-large-discriminator

[3] https://huggingface.co/FacebookAI/roberta-large

[4] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[5] https://huggingface.co/mistralai/Mistral-7B-v0.1

			Offensive Post-level, 2 cls		Vulgarity Token-level, 2 cls		Target Token-level, 4 cls	
		Params	Binary	Macro	Binary	Macro	Micro	Macro
Electra	Large	335M	**.88** $\pm$ 04	.79 $\pm$ 15	.64 $\pm$ 24	**.87** $\pm$ 07	.08 $\pm$ 12	.35 $\pm$ 16
Roberta			**.90** $\pm$ 02	**.86** $\pm$ 03	**.77** $\pm$ 03	**.89** $\pm$ 02	**.27** $\pm$ 03	**.59** $\pm$ 04
Mistral	0-Shot	7.24B	.48 $\pm$ 05	.55 $\pm$ 04	-	-	-	-
	5-Shot		.77 $\pm$ 03	.73 $\pm$ 03	-	-	-	-
Llama3	0-Shot	8B	.78 $\pm$ 03	.75 $\pm$ 04	-	-	-	-
	5-Shot		.82 $\pm$ 02	.75 $\pm$ 03	-	-	-	-
GPT 3.5	0-Shot	-	**.89** $\pm$ 02	**.85** $\pm$ 02	.46 $\pm$ 04	.72 $\pm$ 02	.16 $\pm$ 02	.50 $\pm$ 02
	5-Shot		**.89** $\pm$ 02	**.85** $\pm$ 03	.47 $\pm$ 02	.73 $\pm$ 01	.18 $\pm$ 03	.52 $\pm$ 03
GPT 4	0-Shot	-	.87 $\pm$ 03	**.84** $\pm$ 03	.41 $\pm$ 06	.70 $\pm$ 03	.15 $\pm$ 02	.49 $\pm$ 02
	5-Shot		**.89** $\pm$ 02	**.86** $\pm$ 02	.43 $\pm$ 04	.71 $\pm$ 02	.18 $\pm$ 02	.52 $\pm$ 02

Table 2: Mean F_1 scores and standard deviations of ten-fold cross-validation on the different tasks. Cls stands for the number of classes for the respective task. The Micro F1 scores were computed, leaving out the negative class since the negative class is highly prevalent. Values in bold are statistically insignificantly different.

2024). We attribute this to the general prevalence of English in NLP and to the distinct data sources. Further, the fine-tuned smaller language models perform better in all tasks on our data. However, fine-tuning the decoder-based models would likely improve their performance significantly. These results suggest that fine-tuning yields better outcomes in this setting, particularly for detecting vulgar content.

In line with the results of the experiments on the AustroTox dataset, we find that especially the vulgar token detection task profits from fine-tuning. None of the models achieve good performance on target token detection. However, several factors explain these poor results. First, this is a challenging four-class classification task where the evaluation using the Micro-F1 score focuses only on target classes, excluding the predominant non-target class. This evaluation approach, combined with the sparse distribution of target spans in the data, inherently produces lower scores compared to the other two tasks. The high level of human annotator disagreement provides additional insight into these performance issues. Targets frequently appear multiple times within posts and are often referenced only through pronouns, creating ambiguity. Given that human annotators struggled with the task, the poor model performance becomes more understandable.

9 Conclusion

We re-annotate posts from the Jigsaw Toxic Comment Classification Challenge, providing disaggregated toxicity labels and spans that identify inappropriate language and targets. This sub-document granularity enables more nuanced and personalized toxicity classification. Manual analysis demonstrates that five annotations per instance effectively distinguish meaningful disagreement from annotation inconsistencies. We find high levels of disagreement on borderline toxicity cases, particularly for toxic statements targeting non-human entities. Additionally, when annotating spans comprising inappropriate language, disagreement occurs both on inherently questionable terms and on context-sensitive words that may be acceptable in some settings but inappropriate in others. Finally, experiments show that transformer-based models effectively learn from aggregated data that reduces false negative classifications by being more sensitive towards minority opinions for posts to be toxic.

Acknowledgements

At the time of conducting the main parts of this project, Pia and Anna were funded by the Vienna Science and Technology Fund (WWTF) [10.47379/ICT20015]. Janis is funded by the University of Zurich Research Priority Program project *Digital Religion(s)*. We thank the participants in the annotation campaign. Their dedication and effort have been invaluable to the success of this project. Lastly, the financial support by the Christian Doppler Research Association is gratefully acknowledged.

References

Gavin Abercrombie, Dirk Hovy, and Vinodkumar Prabhakaran. 2023. Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 96–103, Toronto, Canada. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342.

Enrico Bunde. 2021. Ai-assisted and explainable hate speech detection for social media moderators–a design science approach. *Proceedings of the 54th Hawaii International Conference on System Sciences*.

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

Agostina Calabrese, Leonardo Neves, Neil Shah, Maarten Bos, Björn Ross, Mirella Lapata, and Francesco Barbieri. 2024. Explainability and hate speech: Structured explanations make social media moderators faster. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 398–408.

Agostina Calabrese, Björn Ross, and Mirella Lapata. 2022a. Explainable abuse detection as intent classification and slot filling. *Transactions of the Association for Computational Linguistics*, 10:1440–1454.

Agostina Calabrese, Björn Ross, and Mirella Lapata. 2022b. Explainable abuse detection as intent classification and slot filling. *Transactions of the Association for Computational Linguistics*, 10:1440–1454.

cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic comment classification challenge. `https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge`. Kaggle.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Line H Clemmensen and Rune D Kjærsgaard. 2022. Data representativity for machine learning and ai systems. *arXiv preprint arXiv:2203.04706*.

Stefano Cresci, Amaury Trujillo, and Tiziano Fagni. 2022. Personalized interventions for online moderation. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, pages 248–251.

Aida Mostafazadeh Davani, M. C. D'iaz, Dylan K. Baker, and Vinodkumar Prabhakaran. 2023. Disentangling perceptions of offensiveness: Cultural and moral correlates. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.

OpenAI et al. 2024. Gpt-4 technical report.

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*, pages 1–28.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X Zhang. 2023. Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–33.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Jiyun Kim, Byounghan Lee, and Kyung-Ah Sohn. 2022. Why is it hate speech? masked rationale prediction for explainable hate speech detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6644–6655, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5):102643.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 299–318.

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

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, pages 1–70.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Maria D Molina and S Shyam Sundar. 2022. When ai moderates online content: effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, 27(4):zmac010.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Pia Pachinger. 2024. Austrotox: A dataset for target-based austrian german offensive language detection. *Comited to ACL.*

Pia Pachinger, Janis Goldzycher, Anna Planitzer, Wojciech Kusa, Allan Hanbury, and Julia Neidhardt. 2024. Austrotox: A dataset for target-based austrian german offensive language detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11990–12001.

Pia Pachinger, Allan Hanbury, Julia Neidhardt, and Anna Planitzer. 2023. Toward disambiguating the definitions of abusive, offensive, toxic, and uncivil comments. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 107–113, Dubrovnik, Croatia. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682.

Julian Risch, Robin Ruff, and Ralf Krestel. 2020. Offensive language detection explained. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 137–143.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the germeval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12.

Paul Röttger, 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.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020. Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 191–199, Online. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Marcos Zampieri, Skye Morgan, Kai North, Tharindu Ranasinghe, Austin Simmmons, Paridhi Khandelwal, Sara Rosenthal, and Preslav Nakov. 2023. Target-based offensive language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 762–770, Toronto, Canada. Association for Computational Linguistics.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2022. Rethinking offensive text detection as a multi-hop reasoning problem. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3888–3905, Dublin, Ireland. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. Cobra frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315.

Appendix

Figure 9 provides a detailed breakdown of factors contributing to disagreements in toxicity annotation, expanding on the analysis from Section 6 and offering a granular view of the data summarized in Table 1. Similarly, Figure 10 presents the detailed label distribution across post-level annotations and spans in the aggregated dataset (Section 7), providing additional granularity beyond Figure 7. Finally, Figure 11 displays the multitask system prompt used in our experiments (Section 8).

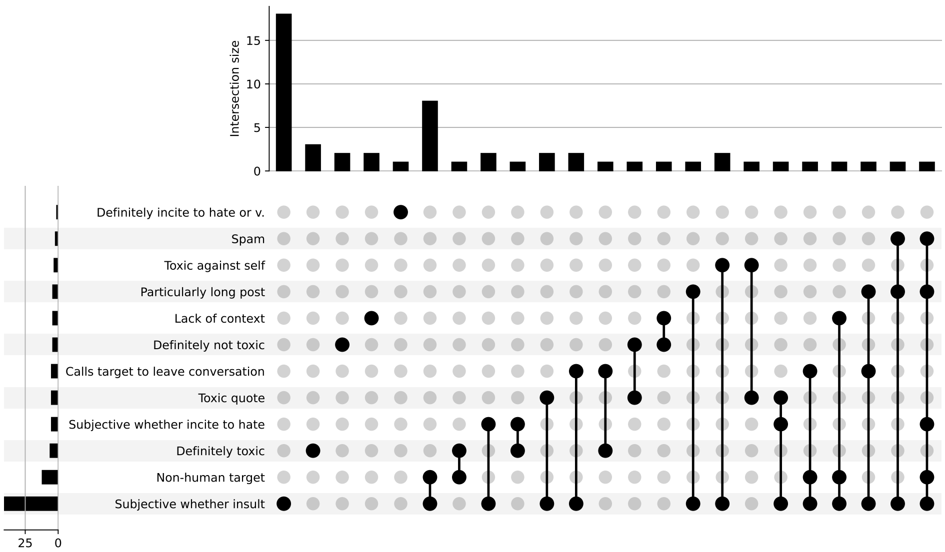


Figure 9: Factors related to disagreement between annotators in offensiveness labels in 60 posts with 3 annotators saying that the post is toxic and 2 annotators saying that it is not toxic.

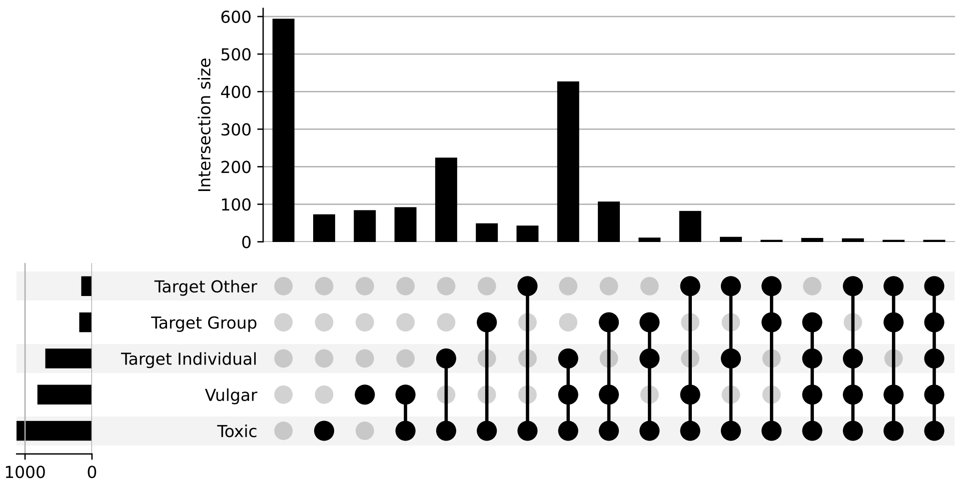


Figure 10: The fine grained class labels of the aggregated data and their co-appearance.

```
You receive a user comment. Your task is to answer
the following questions about the comment:

1. Is the comment toxic? (1 = toxic, 0 = non-toxic)
Definition: An offensive or toxic comment includes derogatory remarks towards
individuals, groups of people, or other entities. It may also incite hatred or
violence against individual persons or groups of people.

2. Who or what is the target of toxicity?
Mark at least one of the following targets of toxicity if the comment is toxic:
"Target_Group", "Target_Individual", or "Target_Other".

3. Mark vulgarities with "Vulgarity". Vulgarities can occur in toxic and
non-toxic comments.

Respond in JSON format with the following fields:
```json
{
    "Label": <0 or 1>,
    "Tags": [
        {
            "Tag": <"Target_Group", "Target_Individual", "Target_Other",
                    or "Vulgarity">,
            "Token": <Span of the target or the vulgarity>
        },
        ...
    ]
}
```
```

Figure 11: The multitask system prompt we use for the neural experiments.

CINEMETRIC: A Framework for Multi-Perspective Evaluation of Conversational Agents using Human-AI Collaboration

Vahid Sadiri Javadi **Zain Ul Abedin** **Lucie Flek**
Conversational AI and Social Analytics (CAISA) Lab, University of Bonn
{vahid.sadirij, zainabedin, lflek}@uni-bonn.de

Abstract

Despite advances in conversational systems, the evaluation of such systems remains a challenging problem. Current evaluation paradigms often rely on costly homogeneous human annotators or oversimplified automated metrics, leading to a critical gap in socially aligned conversational agents, where pluralistic values (i.e., acknowledging diverse human experiences) are essential to reflect the inherently subjective and contextual nature of dialogue quality. In this paper, we propose **CINEMETRIC**, a novel framework that operationalizes pluralistic alignment by leveraging the perspectivist capacities of large language models. Our approach introduces a mechanism where LLMs simulate a diverse set of evaluators, each with distinct personas constructed by matching real human annotators to movie characters based on both demographic profiles and annotation behaviors. These role-played characters independently assess subjective tasks, offering a scalable and human-aligned alternative to traditional evaluation. Empirical results show that our approach consistently outperforms baseline methods, including LLM as a Judge and as a Personalized Judge, across multiple LLMs, showing high and consistent agreement with human ground truth. CINEMETRIC improves accuracy by up to 20% and reduces mean absolute error in toxicity prediction, demonstrating its effectiveness in capturing human-like perspectives.

1 Introduction

What makes a conversation good? If we ask ten people, we might get ten different answers. As shown in Figure 1 (Human Evaluators), a response that one person finds relatively empathetic might strike another as less empathetic or even offensive. These differences highlight that the quality of the dialogue is inherently subjective and multifaceted (Foster et al., 2009). Yet the way we evaluate conversational systems today often assumes that there is an objective fact (by using automatic evaluation

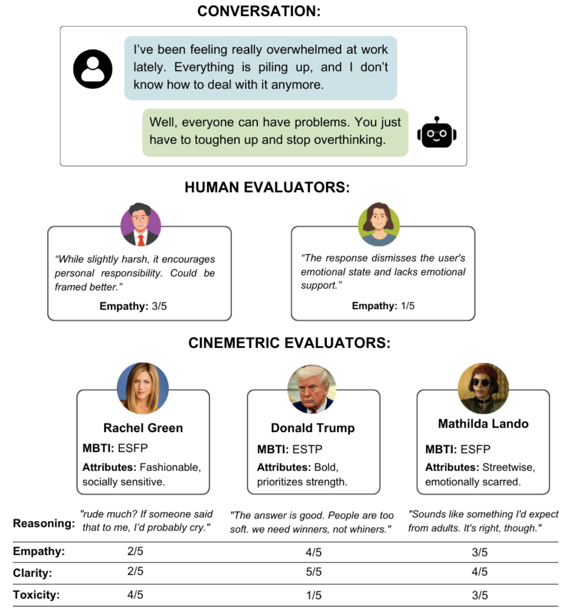


Figure 1: Comparison of Human vs. CINEMETRIC Role-Played Evaluators. A user expresses emotional distress, and the agent responds. Human evaluators and three distinct characters evaluate the agent's response. In the CINEMETRIC evaluation, each character reflects a unique personality profile, resulting in diverse ratings and subjective commentary across various dimensions.

metrics) or a single definitive measure of quality (by aggregating ratings from crowdworkers or domain experts) to rate qualities such as coherence or overall satisfaction (Siro, 2023). However, this "one-size-fits-all" human annotation approach has its own blind spots. It implicitly assumes a homogeneous pool of evaluators, averaging out individual differences. In reality, the background, values, cognitive styles, or personality of an annotator may significantly influence how they perceive the quality of a conversation (Prabhakaran et al., 2021; Gautam and Srinath, 2024). Thus, both purely automatic metrics and aggregated human ratings risk missing the plurality of perspectives inherent in dialogue quality, struggling to capture the nuanced dimensions in a scalable and robust way.

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 15–26

To address this gap, we introduce CINEMETRIC, a novel framework grounded in perspectivist principles for pluralistic alignment (Feng et al., 2024b; Castricato et al., 2024). This idea breaks away with from the singularity of current methods by embracing pluralism (Feng et al., 2024a). As illustrated in Figure 1, the core idea behind this framework is to simulate a panel of diverse evaluators (i.e., movie and public characters), each embodied as a distinct perspective-driven persona, capable of assessing a conversation through different lenses. These personas are defined by interpretable attributes such as gender, personality traits (e.g., MBTI types), thinking style (e.g., analytical vs. intuitive), and more. We then task the LLM to role-play these personas and evaluate conversational turns along multiple dimensions, including but not limited to toxicity, persuasiveness, clarity, and empathy. This work is guided by the following research questions:

RQ1. How can the perspectivist role-playing of diverse personas by large language models enhance the pluralistic alignment of conversational agents?

RQ2. How can Human-AI Collaboration be used to design an evaluation framework that captures the diversity of human values and preferences in LLM outputs?

RQ3. To what extent can perspective-driven evaluations, as instantiated by CINEMETRIC, approximate human judgments and enhance alignment with diverse human evaluative preferences?

In this paper, we propose several steps that help answer our RQs: *(i)* We investigate the conceptual foundations of perspectivism and the need for pluralistic alignment in conversational systems. (Section 2), *(ii)* We detail the design of the CINEMETRIC framework, outline our methodology for persona construction and evaluation design (Section 3). *(iii)* We describe our experimental setup, including how we simulate annotator perspectives through character personas, construct evaluation tasks, and define comparison baselines (Section 4). *(iv)* We present empirical results demonstrating the effectiveness of our method in capturing individual annotator perspectives compared to the baselines (Section 5).

2 Background

In this section, we discuss how perspectivism can be operationalized by the role-playing technique. We investigate the current approaches for evaluating conversational systems, and finally, we explore the pluralistic alignment in conversational agents.

2.1 Perspectivism and Role-Playing

Perspectivism, rooted in Nietzsche's philosophical tradition, refers to the idea that there is no singular objective viewpoint for many problems (Anderson, 1998; Cox, 1997), instead, understanding is shaped by diverse perspectives. Recent work in NLP has embraced this notion by treating annotator disagreements not as noise but as a valuable signal (Uma et al., 2021). For example, Basile (2020) advocates disaggregating annotation labels to preserve individual annotators' viewpoints instead of enforcing a single "ground truth," thereby capturing genuine differences in opinion. This perspectivist approach aims to avoid marginalizing minority opinions and to train models that recognize a spectrum of valid interpretations (Muscato et al., 2025).

One effective way to apply perspectivism in conversational systems is through role-playing or persona-based prompting of LLMs. Role-playing represents a core human ability to simulate different viewpoints and engage in perspective-taking(Jones, 1973). Prior work has shown that role prompts can improve the clarity and relevance of responses by aligning them with the implied perspective of the role (e.g., a "doctor" role yielding medically grounded explanations) (Tseng et al., 2024; Wang et al., 2024; Sun et al., 2024). At the same time, researchers caution that persona prompts can reinforce stereotypes if the model's training data contains biased representations of that role. (Park et al., 2025; Tseng et al., 2024; Tan and Lee, 2025). These studies highlight how role-playing with LLMs provides a versatile framework to inject perspectivism into conversational agents.

2.2 Evaluating Conversational Systems

Evaluating dialogue systems remains a difficult problem in NLP. Traditional evaluation methods for conversational systems have typically fallen into two categories: automated metrics and human evaluation. Automated metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and more recent neural embedding-based approaches like BERTScore (Zhang et al., 2019), which computes similarity scores using contextual embeddings, and contextually sensitive models such as ADEM (Lowe et al., 2018) and DynaEval (Lowe et al., 2018), which enhance correlation with human judgments by considering dialogue context and structure. However, these metrics often fail to capture the nuanced aspects of dialogue quality that matter to humans (Liu et al., 2016). Human evaluation, while more aligned with user experiences, faces challenges of cost, scalability, and annotator variance (Smith et al., 2022; Liu et al., 2022).

A recent trend is to leverage large language models themselves as automatic judges of dialogue (i.e., LLM-as-a-Judge) (Gu et al., 2024; Chan et al., 2023). Instead of a fixed metric formula, one can prompt an advanced LLM (e.g., GPT-4) with a conversation and ask it to provide a rating or feedback, possibly with an explanation. For example, Chiang and Lee (2023) showed that ratings given by ChatGPT-based evaluators correlated more strongly with human judgments than traditional metrics like BLEU or BERTScore. Furthermore, Dong et al. (2024) demonstrated that the standard LLM-as-a-Judge setting is not sufficiently reliable for personalization tasks, showing low agreement with human ground truth. They identified persona sparsity as a major cause of this unreliability. Thus, efforts to infuse evaluation with multiple perspectives from different backgrounds are a direct motivation for the CINEMETRIC framework.

2.3 Pluralistic Alignment in Conversational Agents

As conversational agents become more powerful and widespread, the goal of alignment, i.e., ensuring that an agent's behavior is consistent with human values and intentions, has taken center stage. Traditional alignment approaches, such as reinforcement learning from human feedback (RLHF) (Bai et al., 2022), typically optimize models to perform on average according to the preferences of a broad user base or a set of guidelines (e.g., being helpful, truthful, and harmless). Kirk et al. (2023) show that standard RLHF tends to collapse a model's behavior towards a central norm, reducing the richness of responses it can generate.

The concept of pluralistic alignment begins by questioning *"Whose values?"* current systems are aligned to (Bergman et al., 2024), and it has emerged as a response to the limitations of monolithic evaluation approaches (Conitzer et al., 2024). Gabriel (2020) argues that conversational agents should be designed to acknowledge and respect the diversity of human values rather than optimizing for a single objective function. This perspective aligns with Rawls (1971) concept of *"reasonable pluralism"*, which recognizes that a just society must accommodate diverse and sometimes conflicting conceptions of the good. Moreover, Feng et al. (2024b) have argued that alignment must be reconceived as a socially situated process, acknowledging the pluralism of society rather than pretending there is a single correct value system for a conversational agent. Therefore, researchers proposed diversity-aware alignment frameworks. For instance, pluralistic alignment as defined by Sorensen et al. (2024) is the capacity of conversational agents to handle a plurality of values or preferences, instead of being narrowly tuned to one.

Given the fact that evaluation is inherently multi-perspective, and that we can now harness LLMs to simulate those perspectives in a principled and reproducible way, CINEMETRIC offers a novel solution, namely, a framework that explicitly encodes pluralism into the evaluation pipeline.

3 Methodology: CINEMETRIC

The CINEMETRIC framework proceeds in three steps as shown in Figure 2: *(i)* **Perspective Source** (i.e., selecting representative human evaluators from multi-perspective datasets (Section 3.1)), *(ii)* **Perspective Making** (i.e., creating a set of "persona" movie characters and their corresponding perspectives for each sampled human evaluator (Section 3.2)), and *(iii)* **Perspective Taking** (i.e., leveraging the movie characters and their perspectives by LLMs to make predictions on held-out human evaluator annotations (Section 3.3)).

3.1 Perspective Source

The first step in our framework involves sourcing diverse human perspectives that serve as the grounding for the rest of the framework. To construct this perspective source, we draw from existing datasets that include both: (i) annotations made by individual human evaluators on subjective tasks such as toxicity classification or multiple-choice opinion questions, and (ii) demographic metadata about each evaluator (e.g., age, gender, location, race, political orientation, marital status, education level, etc.). From each dataset, we randomly sample a fixed number of evaluators.

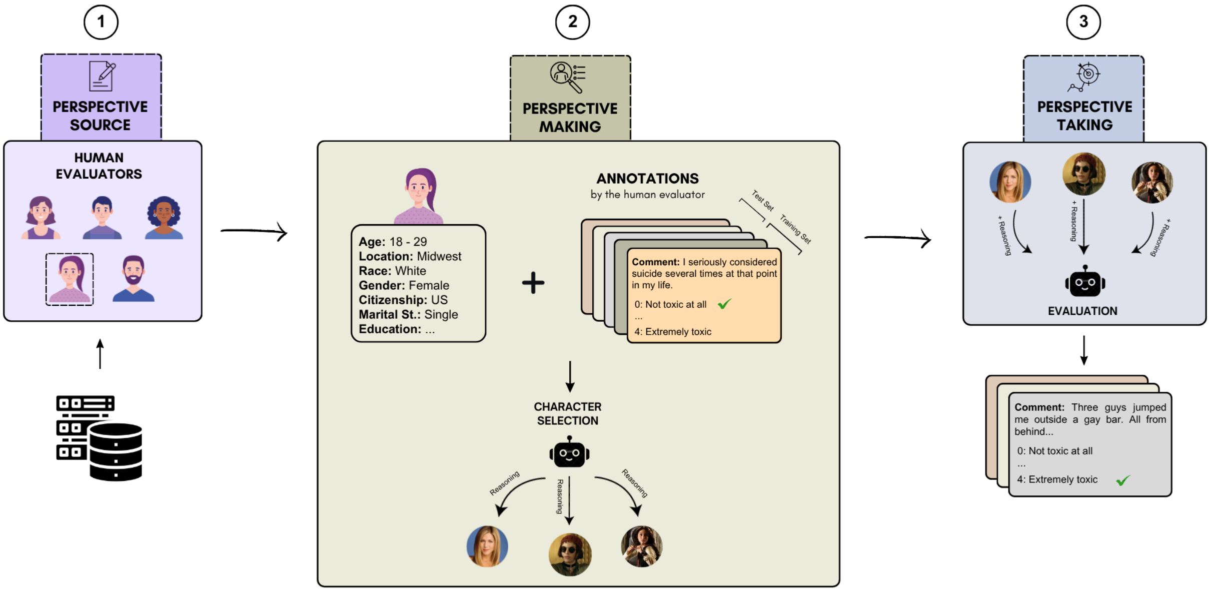


Figure 2: A high-level overview of **CINEMETRIC**, consisting of three steps: ① Perspective Source (See Section 3.1), ② Perspective Making (See Section 3.2), ③ Perspective Taking (See Section 3.3). See Appendix B for prompts for each step. To demonstrate how CINEMETRIC operates in practice, we provide a detailed example in Appendix C centered on a human annotator and the corresponding personas assigned to represent their perspective. We also include the reasoning behind the selection of each movie character, highlighting how their traits align with the evaluator's values and annotation.

3.2 Perspective Making

In this step, we transform each human evaluator into a small set of movie characters who share that evaluator's demographic profile and annotation tendencies.

Concretely, for each sampled human evaluator, we perform the following steps using a large language model: First, we compile the annotator's demographic metadata (e.g. age, gender, region, education, etc.) and some examples of their annotations or responses (training set). Then, we task the LLM to list five movie characters who are demographically and behaviorally similar to this person, given the demographic data as well as the examples of annotations. The LLM also provides a detailed reasoning for each character selection, explaining how the movie character's traits, background, and personality connect to the annotator's profile and annotation patterns, meaning that the characters together capture multiple facets of the human evaluator's perspective. Finally, as the LLM may hallucinate non-existent characters, we verify each suggested character by asking the LLM to check the existence of the movie character in the suggested movie. Any suggested character that the LLM cannot confirm is discarded. From the original five, we keep the first three characters that pass validation. In rare cases where fewer than three valid characters are found, we repeat the generation to produce additional candidates.

After this process, each real annotator is represented by three personas (movie characters), each described by a name and a short rationale for the match. These personas are intended to embody different, plausible user perspectives aligned with the original annotator's demographics and behavior.

3.3 Perspective Taking

Finally, we use the LLM to role-play each of the three movie characters to predict how the persona would respond to each held-out test query of an annotator (test set). Therefore, each of the three personas produces a predicted label or answer for the query. We compute the final prediction by majority vote among the three. If all three differ, we break ties by a fixed rule, choosing the prediction of the first-listed persona.

4 Experimental Setup

To comprehensively evaluate the effectiveness of our proposed approach, we conduct experiments across a diverse set of tasks and models, employing various techniques for comparison.

4.1 Evaluation Tasks

We focus our evaluation on two distinct subjective tasks that require reasoning over human perspectives. In particular, we utilize:

OpinionQA (Santurkar et al., 2023): a multiple-choice question-answering dataset based on U.S. public opinion surveys. OpinionQA contains responses from thousands of respondents, each annotated with 12 demographic features (e.g. age, gender, region, education, political ideology, race, etc.). Each respondent answered 50 questions on various topics, with an average of 3–4 answer choices per question. In our setting, the LLM selects the most likely answer based on a simulated annotator's perspective.

DP (Diversity of Perspectives) (Kumar et al., 2021): a large toxicity annotation dataset. In DP, 17,280 participants each assigned a score from 0 (not toxic) to 4 (very toxic) to 20 social-media comments (drawn from Twitter/Reddit/4chan) and each annotator provided demographic information and personal background. In total, the dataset contains 107,620 comment judgments linked to annotator metadata. The task captures how annotators with diverse backgrounds perceive offensive content differently. In our setting, the LLM assigns a score to the online comments.

In our experiments, for each dataset, we randomly select 100 annotators. Each annotator has a set of annotations (for DP) or answers (for OpinionQA) and demographic information. To simulate their perspectives using CINEMETRIC, we randomly sample 5 examples from each annotator's data for training, used to select the movie characters and construct the perspective-driven reasoning, and 10 examples for testing the evaluation performance. This results in a total of 1,000 test instances (100 annotators × 10 examples), each representing a comment or question to be evaluated by our framework.

4.2 Benchmarking Models

To evaluate the performance of our approach across a wide range of Large Language Models, we experiment with the following LLM families:

1. **DeepSeek**: DeepSeek-V3
2. **OpenAI**: GPT-4.1
3. **Google**: Gemini 2.5 Flash
4. **Mistral**: Mistral Medium 3

These models were selected to cover a wide spectrum of capabilities, sizes and families, enabling us to test CINEMETRIC's robustness across different LLMs. We use the same persona-generation and inference prompts across models. More details on the implementation can be found in Appendix A.

4.3 Methods Studied

We compared our proposed approach against several techniques. These approaches are chosen to cover a range of strategies.

LLM-as-a-Judge (Zheng et al., 2023): In this approach, the LLM is directly prompted to answer questions or assess toxicity without any personalization. This represents the default, non-perspectivist evaluation strategy.

LLM-as-a-Personalized-Judge (Dong et al., 2024): In this approach, the LLM receives demographic metadata of the target annotator to judge user preferences based on personas. This technique constructs personas but does not simulate perspective-taking. This represents a personalization baseline.

Ours: CINEMETRIC: This approach represents the core of our proposed method (described in Section 3), which uses training examples and demographics to match each annotator to movie characters. These characters are then role-played to predict labels on the test set, with majority voting.

Ours: CINEM. w/o Training Examples: To measure the effectiveness of our proposed approach, we remove behavioral data (i.e., the annotator's annotations) from the perspective making step, using only demographic metadata to select movie characters.

Ours: CINEM. w/o Character Names: An ablation, in which the LLM receives the annotator's metadata and behavioral examples, but does not use character names for perspective-taking (i.e., only for perspective-making). Instead, the LLM directly simulates the annotator.

These variants allow us to examine the impact of behavior-based persona construction (i.e., incorporating examples of evaluators' annotations) and the usefulness of using well-known movie characters for grounding perspectives.

4.4 Evaluation Metrics

For OpinionQA, we report accuracy in predicting the annotator's answer. For DP, we report both accuracy (exact match with the annotator's score) and mean absolute error (MAE) to account for near-miss predictions in ordinal toxicity judgments.

Method	DeepSeek	OpenAI	Google	Mistral
	DeepSeek V3	GPT 4.1	Gemini Flash 2.5	Mistral Medium
LLM as a Judge	37.71	45.26	43.56	43.83
LLM as a Personalized Judge	43.27	48.83	49.12	45.42
CINEM. w/o Training Examples	50.00	50.29	48.83	46.79
CINEM. w/o Character Names	52.92	51.16	51.46	**52.92**
CINEMETRIC	**57.31**	**52.33**	**53.53**	48.75

Table 1: The performance (accuracy) of different methods with various LLMs on the OpinionQA dataset.

Method	DeepSeek	OpenAI	Google	Mistral
	DeepSeek V3	GPT 4.1	Gemini Flash 2.5	Mistral Medium
LLM as a Judge	31.11 (1.183)	45.83 (0.9)	43.06 (0.967)	31.37 (1.07)
LLM as a Personalized Judge	37.22 (0.981)	45.00 (0.9)	41.34 (0.934)	27.33 (1.064)
CINEM. w/o Training Examples	37.50 (0.972)	46.11 (0.872)	43.89 (0.844)	31.11 (1.05)
CINEM. w/o Character Names	43.33 (0.847)	47.50 (0.867)	52.78 (0.683)	35.46 (0.904)
CINEMETRIC	**46.94** (0.747)	**49.61** (0.808)	**54.72** (0.653)	**38.27** (0.891)

Table 2: On DP, CINEMETRIC consistently outperforms other techniques. We present the performance (accuracy & MAE) of different methods with various LLMs on the DP dataset.

5 Results & Analysis

We evaluate the performance of CINEMETRIC and competing methods on two datasets across a diverse set of LLMs. As shown in Tables 1 and 2, CINEMETRIC consistently outperforms all baseline approaches, demonstrating its robustness and adaptability across different model families and evaluation formats.

5.1 Performance on OpinionQA

Table 1 presents detailed experimental results on OpinionQA. CINEMETRIC achieves the highest accuracy across all LLMs, surpassing both the LLM as a Judge and LLM as a Personalized Judge baselines. For instance, on DeepSeek V3, CINEMETRIC achieves 57.31% accuracy, which is a substantial improvement over the strongest baseline (Personalized Judge) by significant margins of about 15%. Similar gains are observed on GPT-4.1 (52.33% vs. 48.83%), and on a smaller model like Mistral Medium (48.75% vs. 45.42%). These results indicate that the combination of character-based simulation and perspective-driven alignment significantly enhances model performance in capturing annotator perspectives.

5.2 Performance on DP

Results on DP are presented in Table 2. Performance on the DP dataset reinforces our findings on OpinionQA. CINEMETRIC achieves the highest accuracy on every model, with notable improvements in both categorical prediction and mean absolute error (MAE). For example, on DeepSeek V3, CINEMETRIC reaches 46.94% accuracy with a MAE of 0.747, compared to 37.22% and 0.981 for the Personalized Judge baseline. On GPT-4.1, CINEMETRIC maintains its lead with 49.61% accuracy and a MAE of 0.808. Gemini shows particularly strong results, where CINEMETRIC achieves 52.78% accuracy, again with the lowest MAE (0.683), reflecting better ordinal sensitivity. Even on Mistral, the least capable model in our suite, CINEMETRIC improves performance to 38.27% accuracy with a MAE of 0.891, surpassing all alternative approaches.

5.3 Analysis of CINEMETRIC Aspects

Our baselines (i.e., CINEM. w/o Training Examples & CINEM. w/o Movie Characters) highlight the individual contributions of CINEMETRIC's components. Removing the behavioral training examples (CINEM. w/o Training Examples) consistently reduces accuracy and increases MAE across models, underscoring the value of using human-authored examples to align LLM behavior. When movie characters are excluded (CINEM. w/o Movie Characters), performance generally drops as well, though the magnitude of the decline varies by model. Notably, for Mistral on OpinionQA, the version without movie characters slightly outperforms the full model. This suggests that in resource-constrained models, reducing simulation complexity may be beneficial, possibly due to prompt length limitations or reduced capacity for role-play reasoning. Nevertheless, across all other settings, the full CINEMETRIC framework provides the best overall performance, reaffirming the utility of combining character-based simulation with perspective-driven alignment.

6 Conclusion

In this work, we introduced CINEMETRIC, a novel evaluation framework that operationalizes perspectivist alignment by simulating diverse evaluative standpoints through LLM role-play. By drawing on a rich set of character-based personas, our approach provides a scalable, pluralistic alternative to monolithic evaluation practices. Through comprehensive experiments on two diverse benchmarks and across four leading LLM families, we demonstrated that CINEMETRIC consistently outperforms existing evaluation strategies in both accuracy and MAE. Our results highlight the value of perspective-driven simulation in enhancing the human-likeness and value-diversity sensitivity of automated evaluations. In particular, CINEMETRIC achieves stronger agreement with human judgments than standard LLM-based or personalized-LLM evaluation baselines.

Limitations

Dataset limitations: In this study, we evaluated CINEMETRIC using only two benchmark tasks (OpinionQA and DP), which are diverse in format and domain, but do not exhaust the full range of scenarios in which perspectivist evaluation may be useful. Further evaluation on broader datasets, including open-domain conversations and underrepresented demographic viewpoints, will be explored in future work to strengthen the generalizability of our framework.

Analysis limitations: While our results show that CINEMETRIC exhibits higher agreement with human annotators compared to existing approaches, our current analysis focuses primarily on aggregate accuracy and mean absolute error. We do not yet conduct fine-grained error analyses on persona-specific disagreements or examine how specific attributes (e.g., gender, neurotype) contribute to evaluation variance. Additionally, our agreement metrics are indirect (e.g., accuracy on human-labeled responses), rather than derived from inter-rater correlation with actual human raters on a per-instance basis. A deeper investigation into persona-level contributions and alignment dynamics will help better characterize the interpretability and fairness of CINEMETRIC.

References

R Lanier Anderson. 1998. Truth and objectivity in perspectivism. *Synthese*, 115:1–32.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Valerio Basile. 2020. It's the end of the gold standard as we know it: Leveraging non-aggregated data for better evaluation and explanation of subjective tasks. In *International Conference of the Italian Association for Artificial Intelligence*, pages 441–453. Springer.

Stevie Bergman, Nahema Marchal, John Mellor, Shakir Mohamed, Iason Gabriel, and William Isaac. 2024. Stela: a community-centred approach to norm elicitation for ai alignment. *Scientific Reports*, 14(1):6616.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2024. Persona: A reproducible testbed for pluralistic alignment. *arXiv preprint arXiv:2407.17387*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. 2024. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*.

Christoph Cox. 1997. The" subject" of nietzsche's perspectivism. *Journal of the History of Philosophy*, 35(2):269–291.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can LLM be a personalized judge? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141, Miami, Florida, USA. Association for Computational Linguistics.

KJ Feng, Inyoung Cheong, Quan Ze Chen, and Amy X Zhang. 2024a. Policy prototyping for llms: Pluralistic alignment via interactive and collaborative policy-making. *arXiv preprint arXiv:2409.08622*.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024b. Modular pluralism: Pluralistic alignment via multi-llm collaboration. *arXiv preprint arXiv:2406.15951*.

Mary Ellen Foster, Manuel Giuliani, and Alois Knoll. 2009. Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 879–887.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437.

Sanjana Gautam and Mukund Srinath. 2024. Blind spots and biases: exploring the role of annotator cognitive biases in nlp. *arXiv preprint arXiv:2404.19071*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Delmos J Jones. 1973. Culture fatigue: The results of role-playing in anthropological research. *Anthropological Quarterly*, pages 30–37.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 299–318.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Xiangyang Liu, Tianxiang Sun, Junliang He, Jiawen Wu, Lingling Wu, Xinyu Zhang, Hao Jiang, Zhao Cao, Xuanjing Huang, and Xipeng Qiu. 2022. Towards efficient NLP: A standard evaluation and a strong baseline. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3288–3303, Seattle, United States. Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2018. Towards an automatic turing test: Learning to evaluate dialogue responses. *Preprint*, arXiv:1708.07149.

Benedetta Muscato, Praveen Bushipaka, Gizem Gezici, Lucia Passaro, Fosca Giannotti, and Tommaso Cucinotta. 2025. Embracing diversity: A multi-perspective approach with soft labels. *arXiv preprint arXiv:2503.00489*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Sanghyun Park, Boris Maciejovsky, and Phanish Puranam. 2025. Thinking with many minds: Using large language models for multi-perspective problem-solving. *Preprint*, arXiv:2501.02348.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*.

John Rawls. 1971. An egalitarian theory of justice. *Philosophical Ethics: An Introduction to Moral Philosophy*, pages 365–370.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Clemencia Siro. 2023. Evaluating task-oriented dialogue systems with users. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3495–3495.

Eric Michael Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. *arXiv preprint arXiv:2201.04723*.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

Guangzhi Sun, Xiao Zhan, and Jose Such. 2024. Building better ai agents: A provocation on the utilisation of persona in llm-based conversational agents. *Preprint*, arXiv:2407.11977.

Bryan Chen Zhengyu Tan and Roy Ka-Wei Lee. 2025. Unmasking implicit bias: Evaluating persona-prompted llm responses in power-disparate social scenarios. *arXiv preprint arXiv:2503.01532*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

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

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *Preprint*, arXiv:2307.05300.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Model Implementation Details

All experiments were conducted using the OpenRouter API [1]. Across all models, the results are averages over 5 runs with a temperature of 1.0 and a maximum number of tokens of 2048. The other parameters are set to their default values.

B Prompts Used in CINEMETRIC

We describe the prompts used for each step in the CINEMETRIC framework.

B.1 Perspective Making

B.1.1 Character Selection w/ Training Examples

```
### GOAL: ###
You are a movie character matching expert. Your
     goal is to suggest {num_characters} well-
     known movie characters that match a user
     profile.

### USER'S PROFILE: ###
{user_info}

### TRAINING EXAMPLES: ###
Here are some examples of how this user
     annotated comments/ answered questions:
{examples with ground truth}

### TASK: ###
Based on both the user's profile AND the
     examples shown above, suggest {
     num_characters} movie characters who would
     likely have similar perspectives on what
     content is toxic or harmful.

Consider:
- How this person's background might influence
     their views.
- Characters whose values and experiences would
     align with these judgment patterns.

### IMPORTANT RULES: ###
1. Only suggest real characters from real,
     widely known movies.
2. Choose characters from popular, mainstream
     movies that most people would recognize.
3. Make sure all {num_characters} characters
     are DIFFERENT.

### OUTPUT FORMAT ###
Format your response EXACTLY as follows:
Character Name:[character name 1]
Movie Title:[movie title 1]
Reasoning:[explain why you chose this character,
      considering the person's profile and the
     examples shown]
"""
```

B.1.2 Character Selection w/o Training Examples

```
### GOAL: ###
You are a movie character matching expert. Your
     goal is to suggest {num_characters} well-
     known movie characters that match a user
     profile.

### TASK: ###
Based on the following user profile, suggest {
     num_characters} well-known movie characters
      that match these characteristics:
{user_info}

### IMPORTANT RULES: ###
1. Only suggest real characters from real,
     widely known movies.
2. Choose characters from popular, mainstream
     movies that most people would recognize.
3. Make sure all {num_characters} characters
     are DIFFERENT.

### OUTPUT FORMAT ###
Format your response EXACTLY as follows:
Character Name:[character name 1]
Movie Title:[movie title 1]
```

B.1.3 Character Validation

```
### TASK: ###
You are a movie expert. You should verify if
     the character "{character_name}" is a
     character from the movie "{movie_title}".

### IMPORTANT RULES: ###
1. Only respond with "VALID" if you are 100%
     certain this character exists in this movie
     .
2. Respond with "INVALID" if:
   - The character doesn't exist in this movie
   - The movie doesn't exist
3. Do not provide any explanation, just respond
      with VALID or INVALID.
```

B.2 Perspective Taking

B.2.1 Evaluation w/ Reasoning

```
### GOAL: ###
You should role-play the following movie
     character and solve the task from the movie
      character's perspective, informed by how
     similar real people have approached these
     evaluations.

### MOVIE CHARACTER: ###
You should role-play {character_name} from the
     movie {movie_name}.
You must fully embody the role of {
     character_name} throughout the entire
     evaluation process. This means your
     judgment and reasoning must reflect how {
     character_name} behaves and thinks in {
     movie_name}.

### CHARACTER SELECTION RATIONALE: ###
The reason why this character was chosen is as
     follows:
{Reasoning}
```

[1] https://openrouter.ai/models

```
### TASK: ###
As {character_name}, evaluate ...

### INSTRUCTION: ###
Consider the character selection rationale
    above and the character's values and
    personality when making judgments. This
    means your evaluation must reflect what the
     movie character would genuinely prefer.
    You should not say your personal opinion or
     an objective analysis.

### OUTPUT FORMAT: ###
Your output should have the following format:
Reasoning: [The reason why you choose the
    response.]
Character's aspect: [The aspect of the movie
    character's perspective you use to make
    this judgment.]
Answer: [Your choice]
```

B.2.2 Evaluation w/o Reasoning

```
### GOAL: ###
You should role-play the following movie
    character and evaluate the toxicity of the
    following comment from the movie character'
    s perspective, informed by how similar real
     people have approached these evaluations.

### MOVIE CHARACTER: ###
You should role-play {character_name} from the
    movie {movie_name}.
You must fully embody the role of {
    character_name} throughout the entire
    evaluation process. This means your
    judgment and reasoning must reflect how {
    character_name} behaves and thinks in {
    movie_name}.

### TASK: ###
As {character_name}, evaluate ...

### OUTPUT FORMAT: ###
Your output should have the following format:
Reasoning: [The reason why you choose the
    response.]
Character's aspect: [The aspect of the movie
    character's perspective you use to make
    this judgment.]
Answer: [Your choice]
```

B.2.3 Evaluation w/o Movie Character

```
### TASK: ###
Your task is to evaluate ... from the
    perspective of a user who has the following
     characteristics: {user_info}

### USER RATIONALE: ###
{Reasoning}

### OUTPUT FORMAT: ###
Reasoning: [The reason why you choose the
    response.]
User's aspect: [The aspect of the user's
    perspective you use to make this judgment.]
Answer: [Your choice]
```

B.2.4 LLM-as-a-Personalized-Judge

```
### GOAL: ###
Based on the user profile provided below,
    evaluate ...

### User Profile: ###
{user_info}

### OUTPUT FORMAT: ###
Reasoning: [Explain your reasoning based on
    your profile]
Answer: [Your choice]
```

B.2.5 LLM-as-a-Judge

```
### TASK: ###
Evaluate ...

### OUTPUT FORMAT: ###
Answer: [Your choice]
```

C Case Study Example

To illustrate the workings of CINEMETRIC in a concrete scenario, we present an example involving a single real-world annotator and their three selected personas. The goal is to showcase how the model simulates different character perspectives during toxicity evaluation.

C.1 Persona Matching and Evaluation Walkthrough

Human Annotator Profile: The person is female. Racially, the person is Asian. She is 55 - 64 years old and has some college but no degree. The person views technology's impact as somewhat positive and uses social media, news media, video media, online forums. The person has personally seen toxic content online and considers toxic comments to be occasionally a problem. Politically, the person is independent and identifies as heterosexual. The person is a parent. Religion is very important to the person.

Selected Movie Characters and Rationales:

- **Mulan (Mulan, 1998)**: A disciplined, honor-driven Asian woman with strong moral values, aligned with the user's cultural identity and parenting role. Likely sensitive to personal attacks and sexually inappropriate content.
- **Helen Parr / Elastigirl (The Incredibles, 2004)**: A pragmatic parent and protector, inclined to flag comments that threaten familial or social wellbeing, while tolerating factual or benign competition-related statements.

- **Evelyn Wang (Everything Everywhere All At Once, 2022)**: A resilient Asian immigrant navigating chaos and family, with deep moral grounding. Strong aversion to disrespect, but tolerant of less consequential critiques or political slogans.

Task:
Rating the toxicity of the following comment:
`Comment:` *"Denying genocide is a crime!"*

Role-Played Evaluations:

- **Mulan**:
  - **Reasoning**: Mulan, as a soldier and symbol of justice, sees genocide denial as morally egregious. Upholding truth and defending the vulnerable are central to her character.
  - **Character Aspect**: Military background and moral absolutism.
  - **Toxicity Score**: **2**
- **Helen Parr**:
  - **Reasoning**: As a mother and protector, Helen supports the comment as a factual and justice-aligned assertion. She interprets it as a call to accountability rather than toxicity.
  - **Character Aspect**: Balanced, justice-oriented parental perspective.
  - **Toxicity Score**: **0**
- **Evelyn Wang**:
  - **Reasoning**: Evelyn, grounded in spiritual resilience and firsthand experience with injustice, perceives the comment as morally justified. Her empathy for victims drives her interpretation.
  - **Character Aspect**: Lived trauma, spiritual values, and parental protectiveness.
  - **Toxicity Score**: **0**

Towards a Perspectivist Understanding of Irony through Rhetorical Figures

Pier Felice Balestrucci[1*], Michael Oliverio[1*], Elisa Chierchiello[1], Eliana Di Palma[1], Luca Anselma[1], Valerio Basile[1], Cristina Bosco[1], Alessandro Mazzei[1], Viviana Patti[1],
[1]Computer Science Department, University of Turin, Italy,
Correspondence: pierfelice.balestrucci@unito.it

Abstract

Irony is a subjective and pragmatically complex phenomenon, often conveyed through rhetorical figures and interpreted differently across individuals. In this study, we adopt a perspectivist approach, accounting for the socio-demographic background of annotators, to investigate whether specific rhetorical strategies promote a shared perception of irony within demographic groups, and whether Large Language Models (LLMs) reflect specific perspectives. Focusing on the Italian subset of the perspectivist MultiPICo dataset, we manually annotate rhetorical figures in ironic replies using a linguistically grounded taxonomy. The annotation is carried out by expert annotators balanced by generation and gender, enabling us to analyze inter-group agreement and polarization. Our results show that some rhetorical figures lead to higher levels of agreement, suggesting that certain rhetorical strategies are more effective in promoting a shared perception of irony. We fine-tune multilingual LLMs for rhetorical figure classification, and evaluate whether their outputs align with different demographic perspectives. Results reveal that models show varying degrees of alignment with specific groups, reflecting potential perspectivist behavior in model predictions. These findings highlight the role of rhetorical figures in structuring irony perception and underscore the importance of socio-demographics in both annotation and model evaluation.

1 Introduction

Irony is a complex communicative phenomenon in which the intended meaning diverges from the literal interpretation of an utterance (Muecke, 1970). It often relies on pragmatic inference and contextual cues, making it a challenging target for computational modeling. Beyond its linguistic complexity, irony is also deeply subjective: its perception varies across individuals and is shaped by socio-demographic traits such as age, gender, or cultural background (Frenda et al., 2023a).

Linguistic studies distinguish several categories of irony, including hyperbole, exaggeration, and changes in register, conveyed through rhetorical figures (Karoui et al., 2017). These rhetorical figures can be seen as markers of different categories of irony, each relying on distinct communicative cues (Athanasiadou and Colston, 2020; Kühn and Mitrović, 2024). Recognizing such strategies may therefore aid in detecting irony and understanding how it is perceived across individuals.

At the same time, the subjectivity inherent in irony interpretation poses a challenge: what one person may find clearly ironic, another may interpret literally or fail to recognize altogether. This perspectivist dimension (Frenda et al., 2024) highlights the subjective variability in irony perception, posing challenges for both annotation and computational modeling.

In this paper, we study irony not as a uniform phenomenon, but as a set of rhetorical categories that shape its interpretation. Specifically, we investigate whether certain rhetorical figures promote a shared perception of irony categories among individuals who share socio-demographic traits—and whether such alignment can also be observed in the behavior of Large Language Models (LLMs).

Indeed, LLMs have emerged as powerful tools for natural language understanding and generation. Their ability to capture subtle patterns in language makes them promising candidates for modeling complex pragmatic phenomena such as irony (Balestrucci et al., 2024). Yet, LLMs are not neutral observers: their outputs reflect the data they were trained on, which may embed implicit cultural backgrounds, social perspectives, or biases (Kotek et al., 2023). When applied to subjective phenomena like irony, this raises the question of whether LLMs themselves adopt specific perspec-

*Equal contribution.

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 27–36

tives in how they interpret rhetorical and ironic content (Basile et al., 2024).

To this end, in the first part of the paper, we focus on the Italian subset of the perspectivist MultiPICo dataset (Casola et al., 2024), which contains short social media conversations annotated for irony (*ironic* versus *not-ironic*) by a diverse pool of annotators. So, we augment the MultiPICo annotation by manually annotating the rhetorical figures, adopting the taxonomy proposed by Karoui et al. (2017), into the replies that were labeled as ironic by majority vote in the original campaign. This process is carried out by annotators grouped by generation and gender, allowing us to examine patterns of agreement both within and across demographic groups.

In the second part of the study, we first train LLMs to automatically classify rhetorical figures in ironic replies. In order to improve classification performance, we fine-tune the models on TWITTIRÒ-UD (Cignarella et al., 2017), a corpus of ironic Italian tweets annotated with rhetorical figures. We then examine whether the predictions made by the models reflect the annotation patterns of particular demographic groups—thus highlighting potential perspectivist biases in how LLMs handle complex pragmatic phenomena like irony.

Our study is guided by the following research questions (RQs):

- **RQ1:** Do rhetorical figures promote a shared perception of irony categories across different demographic groups?
- **RQ2:** Do LLMs exhibit perspectivist behavior when classifying rhetorical figures in ironic texts?

The remainder of the paper is structured as follows. Section 2 reviews the literature on irony, rhetorical figures, and perspectivist annotation. Section 3 introduces the MultiPICo dataset. Section 4 outlines our experimental design, followed by the manual annotation campaign and result analysis in Section 5. In Section 6, we present the automatic classification experiments with LLMs. We conclude with a summary of findings in Section 7 and a discussion of limitations in Section 8.[1]

[1]All code and the manually annotated corpus used in this study are available at: https://github.com/MichaelOliverio/perspectivist-understanding-rhetorical-figures.

2 Related Works

Recent work in NLP has increasingly emphasized the importance of taking annotators' perspectives into account when dealing with subjective linguistic phenomena such as irony or hate speech. Instead of treating disagreement as a flaw to be minimized, the *perspectivist approach* (Basile et al., 2021; Frenda et al., 2025) considers it meaningful variation that reflects different ways of interpreting language. To support this view, several studies have proposed modeling annotations at the level of individuals (Davani et al., 2022) or groups defined by shared beliefs or demographic traits (Frenda et al., 2023b; Akhtar et al., 2019).

This line of research relies on disaggregated datasets, where annotations are linked to metadata such as age, gender, ideology, or cultural background (Cabitza et al., 2023; Sachdeva et al., 2022). These datasets allow researchers to investigate how socio-demographic traits influence linguistic judgments, and to build models that better capture the diversity of interpretations (Sap et al., 2021; Wan et al., 2023). Incorporating this information has been shown to improve not only fairness, but also classification performance.

In the domain of irony detection, several studies have started to explore the relationship between perspectivism and the perception of irony (Frenda et al., 2023a,b), revealing, for instance, that irony can be more polarizing depending on the annotators' generation (Casola et al., 2024). In line with this direction, the present work aims to further investigate the perspectivist nature of irony by considering it as a phenomenon that can be classified into rhetorical categories (Karoui et al., 2017). Specifically, we propose a study that seeks to explain and analyze the role of annotators' perspectives in the perception and classification of irony through rhetorical figures.

3 MultiPICo

MultiPICo (Casola et al., 2024) is a multilingual dataset of short social media conversations, each consisting of a post and its reply, annotated to indicate whether the reply is ironic in response to the post. It contains a total of 18,778 post–reply pairs collected from Reddit (8,956) and Twitter (9,822), spanning nine languages. The annotations were obtained through crowdsourcing from 506 individuals with diverse demographic profiles, resulting in 94,342 labels—an average of 5.02 annotations per post–reply pair. Each label is enriched with demographic metadata, including gender, age, ethnicity, student status, and employment.

In the Italian subset, 24 annotators provided 4,790 labels across 1,000 conversations.[2] Among them, 11 were female and 13 male. With respect to age groups, 11 annotators belonged to Gen Z (born between 1997 and 2012), 12 to Gen Y or Millennials (born between 1981 and 1996), and 1 to Gen X (born between 1965 and 1980).

4 Methodology

The first step of our methodology consists in the manual annotation of the Italian subset of MultiPICo by linguistically trained experts with specific knowledge of rhetorical figures. We adopt the taxonomy proposed by Karoui et al. (2017), which classifies irony into eight categories. Seven of these are grounded in rhetorical structures, while the eighth—OTHER —serves as an umbrella category encompassing situational irony and humor (Shelley, 2001; Niogret, 2004).

The seven rhetorical categories are as follows:

- ANALOGY (Ritchie, 2005; Burgers, 2010): involves similarity between two things that have different ontological concepts or domains, on which a comparison may be based.

- HYPERBOLE (Berntsen and Kennedy, 1996; Mercier-Leca, 2003; Didio, 2007): makes a strong impression or emphasizes a point.

- EUPHEMISM (Muecke, 1978; Seto, 1998): reduces the facts of an expression or an idea considered unpleasant in order to soften the reality.

- RHETORICAL QUESTION (Barbe, 1995; Berntsen and Kennedy, 1996): asks a question in order to make a point rather than to elicit an answer.

- CONTEXT SHIFT (Haiman, 1998; Leech, 2016): a sudden change of topic or frame; use of exaggerated politeness in a situation where it is inappropriate, etc.

- FALSE ASSERTION (Didio, 2007): a proposition, fact, or assertion that fails to make sense against reality.

- OXYMORON/PARADOX (Gibbs, 1994; Barbe, 1995; Tayot, 1984): equivalent to "False assertion" except that the contradiction is explicit.

All annotators belong to the same demographic groups considered in the original MultiPICo annotation campaign. For this study, we focus on two dimensions: gender and generation. A subset of 200 ironic Italian post–reply pairs was annotated by six individuals—three male and three female—balanced across generations: two from Gen X, two from Gen Y, and two from Gen Z.

We then analyze whether these groups show consistent patterns in the identification of rhetorical figures for ironic texts, both within and across demographic groups, in order to address our first research question.

In the second phase of the study, we fine-tune various LLMs on rhetorical figure classification. We then evaluate their capability to classify rhetorical figures in ironic post–reply pairs from MultiPICo. Finally, we investigate whether these LLMs exhibit specific perspectives in their classification outputs, analyzing potential alignment with human demographic groups.

5 MultiPICo Annotation

In this section, we describe the annotation of the Italian subset of MultiPICo using the taxonomy proposed by Karoui et al. (2017), which was specifically developed for the analysis of ironic texts. We focus exclusively on post–reply pairs annotated as ironic in MultiPICo, selected through a majority vote strategy. This yields a total of 278 ironic post–reply pairs.

The annotation was performed by six volunteer native Italian speakers, all with a strong academic background in linguistics, on 200 out of the 278 ironic post–reply pairs.

The annotation process follows these steps:

- We adopt the annotation guidelines released by Karoui et al. (2017) to ensure consistency with their framework.[3]

- We label the reply, using the post as contextual information to support the classification of rhetorical figures;

- We assign one or more labels to each reply, depending on the rhetorical figures identified.

[2] `https://huggingface.co/datasets/Multilingual-Perspectivist-NLU/MultiPICo`

[3] Guidelines available at: `https://github.com/Jihen-Karoui/Scheme`

Annotator Agreement across Rhetorical Figures Once the annotation phase was completed, we analyzed the level of agreement among annotators to understand whether certain rhetorical figures promote a more shared perception of irony.

Our hypothesis is that, if some rhetorical figures are more easily or intuitively recognized as markers of irony, they should yield higher agreement scores across annotators. To test this, we computed inter-annotator agreement for each figure using both Fleiss' κ (Fleiss, 1971) and Krippendorff's α (Krippendorff, 2011), as shown in Table 1.

The results reveal notable differences across labels: `RHETORICAL QUESTION` achieves the highest agreement ($\kappa = 0.426$, $\alpha = 0.426$), followed by `HYPERBOLE` and `ANALOGY`. This may be due to the fact that these figures often exhibit salient syntactic or lexical markers in Italian—such as the use of a question mark in rhetorical questions, or comparative structures introduced by *come* ("like/as") in analogies—making them more easily recognizable and less open to interpretive ambiguity. Other figures—such as `EUPHEMISM`, `OXYMORON`, and `CONTEXT SHIFT`—show much lower agreement scores.

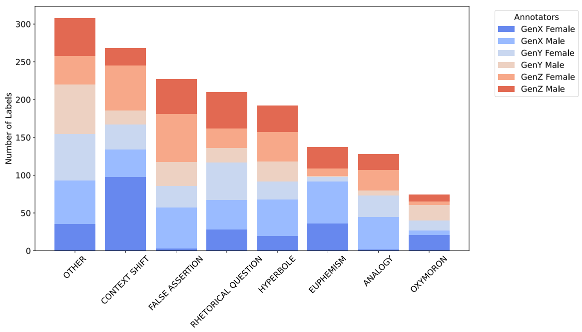


Figure 1: Distribution of Rhetorical Figures Annotated per Annotator

An analysis of label distribution (Figure 1) shows that the most frequent categories are `OTHER` and `CONTEXT SHIFT`, further confirming that annotator agreement is driven more by the presence of recognizable linguistic cues than by the predominance of any single category within the annotated sample.

To illustrate how certain rhetorical figures may be more easily and consistently identified, we report two representative examples from our dataset:

- **Post:** "@USER Not exactly good morning." *("@USER Non troppo buongiorno.")*
- **Reply:** "@USER Grandpa! Already awake???" *("@USER Nonnino! Già sveglio???")*

Five annotators labeled the reply as a *rhetorical question*. The ironic tone emerges from the contrast between the reply's exaggerated cheerfulness and the original negative tone. The question is not meant to be answered, but rather functions as a rhetorical device to underscore the mismatch in mood, making irony both recognizable and effective.

- **Post:** "If you find university easier than high school, I would seriously question your degree program. After all, that's how it should work—you grow, you mature, and gradually you deal with more difficult topics. But the truth is, many universities are just daycare 2.0 for people in their twenties." *("Se trovate più facile l'università che il liceo mi farei serie domande sulla vostra facoltà. D'altronde dovrebbe essere l'ordine naturale delle cose, si cresce, si matura e pian piano si affrontano argomenti più difficili. La verità è però che tante università non sono altro che un asilo 2.0 per ventenni.")*
- **Reply:** "Of course, everyone knows that in every RPG, the final boss is always the hardest one—especially if it's the biggest in the game." *("Del resto lo sanno tutti che in ogni GDR il boss più difficile in assoluto è quello finale, soprattutto se è il più grosso del gioco.")*

Also in this case, five out of six annotators labeled the reply as an *analogy*. The ironic intent is conveyed through a comparison between university education and video game dynamics, suggesting that an academic path should progressively become more challenging—just like in a role-playing game. The analogy is built around a clearly structured evaluative comparison, making the rhetorical figure relatively unambiguous and contributing to the high level of agreement among annotators.

While these examples show that some rhetorical figures can be consistently identified by different annotators, the overall picture remains more nuanced. The average agreement across all figures is modest ($\kappa = 0.198$, $\alpha = 0.199$), suggesting that only some rhetorical strategies promote a shared perception of irony categories.

Crucially, all annotators involved are trained linguists with expertise in rhetorical analysis, and were provided with detailed annotation guidelines. One might therefore expect a high level of objectivity and consistency. However, the observed variation indicates that the classification of rhetorical figures in ironic texts is not a straightforward or universally shared process, but rather a task that involves subjective interpretation—even among experts.

Label	**Fleiss' κ**	**Krippendorff's α**
ANALOGY	0.238	0.238
CONTEXT SHIFT	0.112	0.112
EUPHEMISM	0.089	0.090
FALSE ASSERTION	0.194	0.194
HYPERBOLE	0.304	0.304
OTHER	0.142	0.143
OXYMORON	0.084	0.085
RHETORICAL QUESTION	0.426	0.426
Average	0.198	0.199

Table 1: Inter-annotator agreement scores (Fleiss' κ and Krippendorff's α) for each rhetorical figure.

Annotators' Polarization Following the analysis proposed by Casola et al. (2024), we used the Polarization Index (P-index) introduced by Akhtar et al. (2019). This measure evaluates, for each instance—in our case, each post–reply pair—the polarization in annotations provided by annotators grouped according to specific sociodemographic characteristics. An example of such grouping, shown in Table 2, is by gender (male/female) or by generation (Gen X/Y/Z).

The P-index ranges between 0 and 1, where 0 indicates complete agreement across different groups (no polarization), and 1 indicates maximum internal agreement within each group but total disagreement between groups (maximum polarization).

Formally, the P-index for an instance i is defined as:

$$P(i) = \frac{1}{k} \sum_{w=1}^{k} a(G_w) \cdot (1 - a(G)) \tag{1}$$

where k is the number of groups (for example, 3 in the case of grouping by generation), $a(G_w)$ is the internal agreement level within group G_w for instance i, and $a(G)$ is the overall agreement level of all annotators on instance i. Following the original proposal, the agreement (a) is calculated using a normalized χ^2 statistic:

$$a(G) = \frac{\chi^2(G)}{|M|} \tag{2}$$

where $\chi^2(G)$ denotes the chi-square statistic for group G, and $|M|$ is the number of annotations for the corresponding instance.

We employed the P-index for groups defined by gender and generation. Due to the multi-label nature of our annotation scheme, where annotators can assign multiple rhetorical figures to a single instance, we compute the P-index independently for each rhetorical figure and report the average across all figures. An example of the P-index on an instance can be seen in Table 3.

To establish a baseline, we calculated the P-index for each rhetorical figure over all possible random combinations of annotators—pairs for gender grouping and triplets for generation grouping—and averaged the results accordingly.

Additionally, we also calculated the percentage difference (%Δ) between the real P-index and the random P-index, to highlight the degree of polarization actually observed compared to a random baseline.

Gender			**Generation**		
real	random	%Δ	real	random	%Δ
0.124	0.132	-6.10	0.191	0.146	31.13

Table 2: Polarization index values calculated for annotator groups based on gender and generation. The table shows the real P-index, the random P-index obtained by averaging over random permutations of annotators, and the relative percentage difference (%Δ) between the real and random values.

The results in Table 2 show that for the gender dimension, the real P-index value (0.124) is lower than the one expected by chance (0.132), with a negative percentage difference of -6.10%. This suggests that annotators do not tend to polarize based on gender; in fact, their annotations appear to be slightly less variable within gender groups than would be expected randomly. In contrast, for the generation dimension, the real P-index value (0.191) is higher than the random baseline (0.146), with a positive difference of 31.13%. This indicates that generation is a polarizing trait in the annotation of rhetorical figures. In other words, annotators within the same age group tend to agree more with each other, while differing more from those in other generational groups.

Post	Reply	Ann. Gen.	An	Cs	Eu	Fa	Hy	Ot	Ox	Rq	P-index
@USER It will be the first strong team they face.... (*@USER Sarà la prima squadra forte che affrontano....*)	@USER Which one of the two? ? (*@USER Quale delle due? ?*)	X	0	1	0	0	0	0	0	1	0.100
		X	0	1	0	0	0	0	0	0	
		Y	0	0	0	0	0	0	0	1	
		Y	0	0	0	0	0	0	0	1	
		Z	0	0	0	0	0	0	0	1	
		Z	0	0	0	0	0	0	0	1	

Table 3: Example of polarization in the annotations. While the reply *"Which one of the two?"* may appear as a rhetorical question, the table reveals disagreement among annotators from different generations. All Gen Z and Gen Y annotators labeled it as a *Rhetorical question* (Rq), whereas only one Gen X annotator agreed, with another opting for *Context shift* (Cs). Abbreviations: Ann. Gen. = Annotator Generation, An = Analogy, Cs = Context shift, Eu = Euphemism, Fa = False assertion, Hy = Hyperbole, Ot = Other, Ox = Oxymoron, Rq = Rhetorical question.

6 Rhetorical Figure Classification and Perspective Alignment in LLMs

In this section, we explore whether LLMs reflect specific perspectives when classifying rhetorical figures in ironic texts. As a first step, we fine-tuned a set of multilingual LLMs on the TWITTIRÒ-UD dataset, aiming to enhance their performance in the classification of rhetorical figures within ironic language. Indeed, while the TWITTIRÒ-UD dataset serves to fine-tune and evaluate the LLMs' classification abilities, the MultiPICo data instead allow us to assess whether model predictions align more closely with specific demographic perspectives.

TWITTIRÒ-UD TWITTIRÒ-UD is a corpus of ironic Italian tweets annotated with rhetorical figures and linguistic information following the Universal Dependencies (UD) framework.[4] It contains 1,424 tweets and over 28,000 tokens, originally collected for the fine-grained annotation of irony. Each tweet is labeled with the rhetorical figure used to convey irony, based on the taxonomy proposed by Karoui et al. (2017).

Model Setup and Fine-Tuning We fine-tuned four LLMs on TWITTIRÒ-UD using a reasoning instruction format, in which the model is prompted to first generate a short explanation before producing the final label, following the Chain-of-Thought prompting strategy (Wei et al., 2022). The models we used are:

- Llama-3.1-8B-Instruct[5],
- Ministral-8B-Instruct-2410[6],
- LLaMAntino-3-ANITA-8B-Inst-DPO-ITA[7],
- Minerva-7B-instruct-v1.0[8].

Model fine-tuning Fine-tuning was performed using the Low-Rank Adaptation (LoRA) method (Hu et al., 2021). All models were prompted in English and trained to output both the explanation and the final rhetorical figures using the labels from the original annotation schema. The training was conducted using the `transformers` and `peft` libraries. Table 4 summarizes the main parameters used in the `TrainingArguments` class and in the LoRA configuration.

Parameter	Value
LoRA configuration	
LoRA rank (r)	64
LoRA alpha	16
Dropout probability	0.1
TrainingArguments	
Number of training epochs	5
Enable fp16 training	`False`
Enable bf16 training	`True`
Batch size per GPU for training	1
Batch size per GPU for evaluation	1
Gradient accumulation steps	1
Maximum gradient norm	0.3
Initial learning rate	2e−4
Weight decay	0.001
Optimizer	`adamw_torch`
Learning rate schedule	`cosine`
Warmup ratio	0.03

Table 4: Configuration of hyperparameters used in the LoRA-based fine-tuning process.

The input prompt for the fine-tuning followed this format:

[4] https://github.com/UniversalDependencies/UD_Italian-TWITTIRO

[5] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[6] https://huggingface.co/mistralai/Ministral-8B-Instruct-2410

[7] https://huggingface.co/swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA

[8] https://huggingface.co/sapienzanlp/Minerva-7B-instruct-v1.0

Instruction: Given the ironic sentence (INPUT), identify and return the rhetorical figure it exemplifies in (OUTPUT). Explain your reasoning first, and then answer with the rhetorical figure.

Baselines To contextualize the performance of the fine-tuned models, we defined two baselines:

- **Random**: a naive classifier that assigns one of the eight possible rhetorical categories uniformly at random. This provides a sense of the task's inherent difficulty.
- **Zero-Shot prompting**: we prompted the best-performing model in its non–fine-tuned version using the same instruction and listing all rhetorical categories as candidate outputs. This baseline allows us to estimate how much LLMs know about rhetorical devices without fine-tuning.

Model	Precision	Recall	F1-score
Llama-3.1-8B	0.378	0.406	0.384
LLaMAntino-3-8B	0.382	0.397	0.385
Ministral-8B	**0.393**	**0.408**	**0.396**
Minerva-7B	0.367	0.385	0.372
Random	0.138	0.122	0.125
Zero-Shot	0.213	0.218	0.185

Table 5: Performance of fine-tuned models on the TWITTIRÒ-UD test set. Scores are reported as weighted averages of precision, recall, and F1-score across three runs.

Results on TWITTIRÒ-UD Table 5 reports the classification results on the TWITTIRÒ-UD test split. Each LLM was run three times per input using a temperature of 0.1. We report the results as the weighted average of Precision, Recall, and F1-score, in order to account for the different distribution of the rhetorical figures in the dataset.

The random baseline acts as a benchmark to evaluate the inherent difficulty of the task: given the presence of eight possible classes, it is very unlikely to achieve strong results through chance alone. Within this challenging setup, Ministral-8B achieves the highest performance, narrowly surpassing other fine-tuned models. Moreover, the zero-shot results obtained by prompting Ministral-8B reveal that LLMs possess some prior understanding of rhetorical figures and their use, as evidenced by their performance exceeding random chance. Finally, fine-tuning on the TWITTIRÒ-UD dataset leads to a substantial improvement in their classification performance.

Do LLMs Exhibit a Specific Perspective? To explore whether LLMs adopt a specific perspective when classifying rhetorical figures, we assessed their performance against gold references derived from different demographic groups. Specifically, for each group in the Italian subset of MultiPICo (Female, Male, Gen X, Gen Y, Gen Z), we computed the most frequently assigned rhetorical figure label across all instances, based on the annotations provided by human annotators belonging to that group in Section 5. These labels were then used as gold references to calculate precision, recall, and F1-scores for each model. We also computed an additional "Global" reference, using the most frequent label aggregated across all annotators, regardless of group.

Table 6 reports model performance under these different evaluation perspectives. The results show consistent variation depending on which group's labels are used as gold. For instance, Llama-3.1-8B performs notably better when evaluated against the Gen X labels (F1 = 0.215), suggesting a closer alignment with the rhetorical preferences of Gen X annotators. Minerva-7B shows a similar trend, also achieving its highest F1-score (0.260) with Gen X. In contrast, LLaMAntino-3-8B performs best when evaluated against the labels assigned by the Gen Z group (F1 = 0.241), while Ministral-8B performs best with the Female group (F1 = 0.261)

These findings suggest that LLMs may align more closely with certain annotation patterns, reflecting differences in how rhetorical figures are interpreted across demographic groups.

Error Analysis To better understand the classifications produced by the models, we conducted an analysis of the most frequent errors.

One of the most common issues involves the distinction between the post and the reply. In many cases, the models tend to assign the label to the post rather than the reply, which is actually the correct target for classification. For example, in the following pair:

- **Post**: "Do you think a MORTADELLA SANDWICH could be considered HOMEOPATHIC?" (*"Secondo voi il PANINO CON LA MORTADELLA si può considerare OMEOPATICO?"*)

Model	Group	Precision	Recall	F1-Score
Llama-3.1-8B	Female	0.236	0.214	0.204
	Male	0.220	0.199	0.177
	Gen X	**0.305**	0.199	**0.215**
	Gen Y	0.187	0.194	0.160
	Gen Z	0.161	0.159	0.138
	Global	0.217	**0.219**	0.195
LLaMAntino-3-8B	**Female**	**0.333**	0.174	0.187
	Male	0.271	0.189	0.202
	Gen X	0.251	0.179	0.193
	Gen Y	0.267	0.199	0.204
	Gen Z	0.275	**0.249**	**0.241**
	Global	0.311	0.204	0.224
Ministral-8B	**Female**	0.327	**0.244**	**0.261**
	Male	0.254	0.199	0.200
	Gen X	0.258	0.184	0.202
	Gen Y	0.275	0.224	0.218
	Gen Z	0.193	0.189	0.182
	Global	**0.346**	0.239	0.250
Minerva-7B	Female	0.305	0.214	0.220
	Male	0.327	0.184	0.183
	Gen X	**0.367**	**0.234**	**0.260**
	Gen Y	0.296	0.184	0.166
	Gen Z	0.181	0.184	0.167
	Global	0.314	0.209	0.202

Table 6: Performance of each model on the Italian subset of MultiPICo, reported as weighted averages of precision, recall, and F1-score. Gold labels correspond to the most frequent label assigned by human annotators for each demographic group (Female, Male, Gen X, Gen Y, Gen Z) and overall (Global).

- **Reply**: "@USER Yes" (*"@USER Si"*)

LLaMAntino-3-8B assigns the label `RHETORICAL QUESTION`, which is more appropriate for the post than for the reply. In this case, most human annotators labeled the reply as `FALSE ASSERTION`, a rhetorical figure that better reflects the content of the response.

Another critical issue is the presence of hallucinations in the models' responses. For instance:

- **Post**: "@USER No no, it's right, it has to be there, you feed it, cuddle it, keep it warm, it has to be there" (*"@USER No no è giusto, ce deve sta, la nutri la coccoli la tieni calda, ce deve sta"*)

- **Reply**: "@USER Actually, the other one handles it. I'm just a disruptive element." (*"@USER Veramente ce pensa quell'altro. Io sono un mero elemento di disturbo."*)

In this case, Llama-3.1-8B labels the reply as `SITUATIONAL IRONY`, which is not part of the label set used during fine-tuning. The appropriate label would be `OTHER`, which was in fact the most frequently assigned category by annotators in similar situations.

This analysis highlights the need for improvements in the fine-tuning phase of the models, particularly to ensure clarity that the classification should refer exclusively to the reply, with the post serving only as contextual information. Additionally, it is important to reinforce the alignment between the available labels and those used by the model, in order to avoid generating labels not included in the adopted taxonomy.

7 Conclusions

In this paper we investigated irony as a multifaceted phenomenon, structured by different rhetorical figures that guide its interpretation. By focusing on the Italian subset of the perspectivist MultiPICo dataset, we conducted a manual annotation campaign in which expert annotators labeled rhetorical figures in ironic replies. The annotators were balanced across gender and generation, allowing us to explore patterns of agreement both within and across demographic groups.

Our findings show that only some rhetorical figures—such as `RHETORICAL QUESTION`, `HYPERBOLE`, and `ANALOGY`—promote a more shared perception of irony categories. Others yielded lower agreement, highlighting the subjective nature of this task. Despite the linguistic expertise of the annotators and the use of detailed guidelines, the overall agreement remained modest, supporting the perspectivist view that irony interpretation is influenced by socio-demographic background.

We then trained and evaluated LLMs on rhetorical figure classification. While fine-tuned models outperformed baselines, their predictions showed variation depending on which group's annotations were used as gold labels. In particular, different models aligned more closely with different demographic perspectives—suggesting that LLMs may replicate specific patterns observed in human annotation.

These results emphasize the importance of incorporating socio-demographic information when modeling complex pragmatic phenomena such as irony, both to improve classification performance and to better account for variation in human interpretation.

8 Limitations

This study presents a first attempt to investigate the perspectivist nature of irony through the lens of rhetorical figures. However, it presents some limitations that open directions for future work.

First, our analysis is limited to the Italian subset of the MultiPICo dataset. While this choice enabled a controlled and linguistically grounded study, future work will extend the approach to other languages and cultural contexts, to assess whether similar perspectivist patterns emerge cross-linguistically.

Second, the annotation was carried out by a small group of six annotators. This limited sample size may restrict the generalizability of our findings. Nonetheless, we opted for a small but expert group of annotators—all with a background in linguistics—to ensure a high-quality annotation of complex rhetorical phenomena. Relying on larger but less specialized crowdsourcing platforms could have introduced noise and inconsistencies, particularly in the classification of fine-grained rhetorical strategies.

Third, to improve model performance in the automatic classification task, we fine-tuned the LLMs on the TWITTIRÒ-UD dataset. While this resource provides valuable rhetorical annotations for ironic content, its use may introduce a potential source of bias, as the labels reflect the interpretative choices of a different group of annotators.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In *AI* IA 2019–Advances in Artificial Intelligence: XVIIIth International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, November 19–22, 2019, Proceedings 18*, pages 588–603. Springer.

Angeliki Athanasiadou and Herbert L Colston. 2020. *The Diversity of Irony*, volume 65. Walter de Gruyter GmbH & Co KG.

Pier Felice Balestrucci, Silvia Casola, SODA Lo, Valerio Basile, Alessandro Mazzei, and 1 others. 2024. I'm sure you're a real scholar yourself: Exploring ironic content generation by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14480–14494. Association for Computational Linguistics.

Katharina Barbe. 1995. Irony in context.

Valerio Basile, Silvia Casola, Simona Frenda, and Soda Marem Lo. 2024. PERSEID - perspectivist irony detection: A CALAMITA challenge. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1074–1081, Pisa, Italy. CEUR Workshop Proceedings.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, and 1 others. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics.

Dorthe Berntsen and John M Kennedy. 1996. Unresolved contradictions specifying attitudes—in metaphor, irony, understatement and tautology. *Poetics*, 24(1):13–29.

Christian Frederik Burgers. 2010. *Verbal irony: Use and effects in written discourse*. sl: sn.

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

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Alessandra Cignarella, Cristina Bosco, and Viviana Patti. 2017. *TWITTIRÒ: a Social Media Corpus with a Multi-layered Annotation for Irony*, pages 101–106.

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.

Lucie Didio. 2007. *Une approche sémantico-sémiotique de l'ironie*. Ph.D. thesis, Limoges.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

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*, pages 1–28.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist approaches to natural language processing: a survey. *Lang. Resour. Evaluation*, 59(2):1719–1746.

Simona Frenda, SODA Lo, Silvia Casola, Bianca Scarlini, Cristina Marco, Valerio Basile, Davide Bernardi, and 1 others. 2023a. Does anyone see the irony here? analysis of perspective-aware model predictions in irony detection. In *CEUR WORKSHOP PROCEEDINGS*, volume 3494, pages 1–11. CEUR-WS.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023b. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Raymond W Gibbs. 1994. *The poetics of mind: Figurative thought, language, and understanding*. Cambridge University Press.

John Haiman. 1998. *Talk is cheap: Sarcasm, alienation, and the evolution of language*. Oxford University Press.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Jihen Karoui, Farah Benamara, Véronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles. 2017. Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 262–272, Valencia, Spain. Association for Computational Linguistics.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA. Association for Computing Machinery.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Ramona Kühn and Jelena Mitrović. 2024. The elephant in the room: Ten challenges of computational detection of rhetorical figures. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 45–52, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Geoffrey N Leech. 2016. *Principles of pragmatics*. Routledge.

Florence Mercier-Leca. 2003. *L'ironie*. Hachette Éducation.

D. C. Muecke. 1970. *Irony and the Ironic*. Methuen, London.

Douglas C Muecke. 1978. Irony markers. *Poetics*, 7(4):363–375.

Philippe Niogret. 2004. Les figures de l'ironie dans" a la recherche du temps perdu".

David Ritchie. 2005. Frame-shifting in humor and irony. *Metaphor and Symbol*, 20(4):275–294.

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.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Ken-ichi Seto. 1998. On non-echoic irony. *PRAGMATICS AND BEYOND NEW SERIES*, pages 239–256.

Cameron Shelley. 2001. The bicoherence theory of situational irony. *Cognitive Science*, 25(5):775–818.

Claudine Tayot. 1984. *L'ironie*. Ph.D. thesis.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14523–14530.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

From Disagreement to Understanding: The Case for Ambiguity Detection in NLI

Chathuri Jayaweera and Bonnie J. Dorr
University of Florida, Gainesville, FL,USA
{chathuri.jayawee, bonniejdorr}@ufl.edu

Abstract

This position paper argues that annotation disagreement in Natural Language Inference (NLI) is not mere noise but often reflects meaningful variation, especially when triggered by ambiguity in the premise or hypothesis. While underspecified guidelines and annotator behavior contribute to variation, content-based ambiguity provides a process-independent signal of divergent human perspectives. We call for a shift toward ambiguity-aware NLI that first identifies ambiguous input pairs, classifies their types, and only then proceeds to inference. To support this shift, we present a framework that incorporates ambiguity detection and classification prior to inference. We also introduce a unified taxonomy that synthesizes existing taxonomies, illustrates key subtypes with examples, and motivates targeted detection methods that better align models with human interpretation. Although current resources lack datasets explicitly annotated for ambiguity and subtypes, this gap presents an opportunity: by developing new annotated resources and exploring unsupervised approaches to ambiguity detection, we enable more robust, explainable, and human-aligned NLI systems.

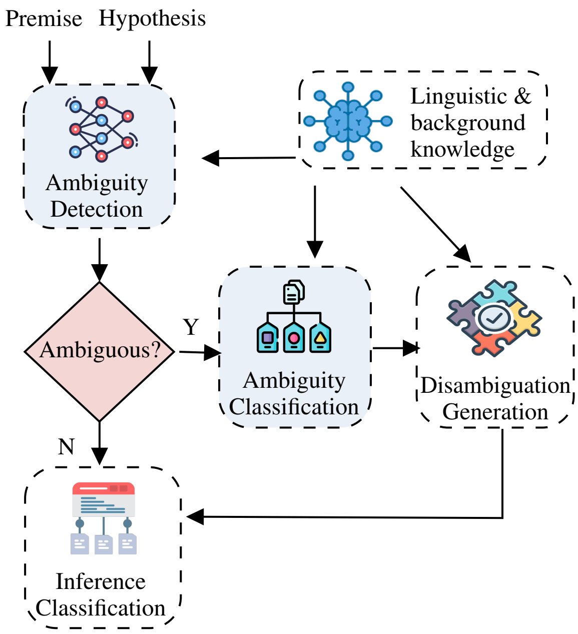


Figure 1: Framework for ambiguity-aware NLI: First detect whether a (P)remise or (H)ypothesis is ambiguous and, if so, classify the ambiguity type. Generate disambiguated versions and pass these to the inference classifier. Linguistic and other relevant background knowledge inform each stage. Gray = focus of this paper; white = supporting stages.

1 Introduction

This paper takes a position on how disagreement in Natural Language Inference (NLI) is best understood and modeled. While prior work has often treated annotator disagreement as noise—something to be minimized or resolved (Snow et al., 2008; Bowman et al., 2015)—we argue that such disagreement can reflect meaningful, coexisting interpretations grounded in linguistic ambiguity.

NLI, also known as Recognizing Textual Entailment (RTE) (Dagan et al., 2005), aims to classify the relationship between a premise (P) and a hypothesis (H). Suppose P1=*John likes Mary*, P2=*John lives near Mary*, H1=*John knows Mary*, H2=*John doesn't know Mary*. Standard inference labels would assign *entailment* to (P1, H1), *contradiction* to (P1, H2), and *neutral* to (P2, H1). However, humans may diverge: (P1, H2) could also be neutral if *likes* is interpreted as distant admiration (e.g., of a celebrity).

We adopt a perspectivist reframing of NLI that treats variation in inference classifications not as a flaw but as an inherent feature of natural language understanding. We emphasize content-based ambiguity as a central source of disagreement and advocate for deeper exploration of its role in shaping inference judgments.

We situate our analysis within a framework for handling ambiguous NLI instances (Figure 1). This framework first determines whether an input pair is ambiguous; if so, it disambiguates the pair into dis-

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 37–46
November 8, 2025 ©2025 Association for Computational Linguistics

tinct yet plausible human interpretations, enabling predictions aligned with each interpretation.

NLI is pivotal in understanding semantic relationships and is central to evaluating how well language models process natural language. NLI benchmarks are typically constructed using human-annotated entailment labels (Bowman et al., 2015; Williams et al., 2018). Despite frequent disagreements among annotators on the "correct" label for a given premise-hypothesis pair, most NLI research assumes a single "true" inference for each case.

Instances that deviate from this assumption are either filtered out during dataset construction (Bayer et al., 2005) or handled via majority vote (Bowman et al., 2015), based on the belief that annotation disagreements reflect random error rather than systematic variation. However, this approach contradicts the original purpose of NLI: to model what a reasonable, attentive, and informed human would plausibly infer from text (Manning, 2006).

Recent studies have challenged the assumption that annotation disagreements are mere noise, demonstrating instead that such disagreements exhibit reproducible patterns grounded in legitimate interpretive differences (Pavlick and Kwiatkowski, 2019). This recognition has motivated efforts to model the full distribution of plausible human inferences (Chen et al., 2020; Meissner et al., 2021).

While these efforts are important, understanding *why* such differences arise is equally critical for developing systems that reflect multiple human interpretations. This position paper argues for an NLI modeling goal that centers on **identifying and categorizing ambiguity into recurring interpretive patterns, rather than merely modeling annotator distributions to capture coexisting human perspectives**. We support this position through a review and analysis of existing research.

The next section reviews work on modeling annotator label distributions and their limitations, motivating a closer look at sources of disagreement in NLI. Section 3 examines prior categorizations of these sources, Section 4 highlights the unique role of ambiguity in premise-hypothesis pairs, and Section 5 surveys current ambiguity-focused NLI research and future directions.

2 Modeling annotator distribution

Most NLI benchmarks are constructed using human-annotated entailment labels, which often result in cases where multiple annotators assign different labels to the same premise-hypothesis pair. From the early days of NLI research, scholars have expressed concerns about how to handle such disagreements (Bayer et al., 2005).

Re-annotation of the RTE1 development and training sets reveals substantial discrepancies between the original and new labels (Bayer et al., 2005). Even after filtering out problematic examples, human judges only achieve a 91% agreement rate. Similar disagreements in the Stanford Natural Language Inference (SNLI) dataset complicate the process of learning robust decision boundaries for each entailment label (Pan et al., 2018).

Such cases are typically treated as "annotation noise," resolved by assigning a majority label under the assumption that one "true" inference exists for each premise-hypothesis pair. However, growing evidence suggests that these disagreements reflect systematic, reproducible variation rather than random error (Pavlick and Kwiatkowski, 2019). In many cases, divergent annotations signal the existence of multiple plausible interpretations.

Current NLI models, trained on majority-labeled benchmarks, struggle to capture the full distribution of human judgments and tend to perform better when annotator agreement is high (Nie et al., 2020). This highlights both a dependence on agreement and a failure to model collective human reasoning. Meissner et al. (2021) further show that models trained on soft labels—distributions over annotator responses—better approximate human judgments and improve single-label prediction accuracy.

These findings have inspired a growing line of research focused on modeling human opinion distributions. For example, the Uncertain Natural Language Inference (UNLI) framework (Chen et al., 2020) proposes predicting subjective probabilities of entailment rather than coarse categorical labels. While UNLI captures a more probabilistic notion of inference, it targets average responses and does not attempt to model the full range of interpretations.

Zhang and de Marneffe (2021) contrast systematic inference (high agreement) and ambiguous cases (high disagreement). They build artificial annotators using BERT (Devlin et al., 2019) to simulate annotation variation, enabling downstream models to determine if a given premise-hypothesis pair is likely to elicit disagreement. Zhou et al. (2022) further improve modeling of opinion distributions beyond standard softmax assumptions.

Together, these studies mark a shift from a prescriptive view—assuming a single correct label—

toward a descriptive approach that acknowledges interpretive variations. They help pave the way for systems that capture ambiguity inherent in natural language. However, simply modeling disagreement alone does not explain *why* interpretations diverge. To advance beyond descriptive modeling, we argue that NLI systems must also systematically identify and categorize the sources of disagreement—especially content-based ambiguity—as a foundation for more perspective-sensitive inference (Plank, 2022). We next examine the sources that give rise to divergent judgments in NLI.

3 Disagreement sources in NLI

According to the "Triangle of Reference" (Aroyo and Welty, 2015), disagreement in annotation arises from three main sources: (1) interpretative ambiguity in the *input content* itself (Uncertainty in sentence meaning); (2) unclear annotation guidelines (Underspecification in guidelines); and (3) differences in annotators' background knowledge or task understanding (Annotator behavior). This framework maps directly onto annotation workflows in NLI benchmarks. Building on this foundation, Jiang and de Marneffe (2022) propose a more fine-grained taxonomy for NLI, refining each category into subtypes that reflect recurring premise-hypothesis patterns (Figure 2).

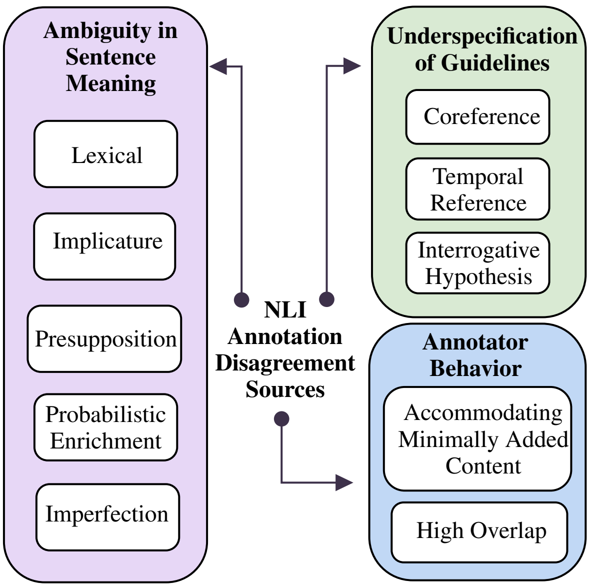


Figure 2: The taxonomy of disagreement sources developed by Jiang and de Marneffe (2022), building on Aroyo and Welty (2015). While these frameworks classify sources of disagreement, we argue for *reframing* such disagreement as a signal of coexisting interpretations to model—not noise to be resolved.

In the subsections below, we adopt a variant of this taxonomy, reframing *Uncertainty in Sentence Meaning* as *Ambiguity in Sentence Meaning* (Section 3.1), a shift already noted in Figure 2. This distinction is central to our position. For completeness, we also briefly describe the roles of guideline underspecification (Section 3.2) and annotator behavior (Section 3.3), though these are not the central emphasis of our position.

In addition, while we view Jiang and de Marneffe's taxonomy as a valuable classification framework, we go further: rather than treating disagreement as noise to be explained or resolved, we reframe it as a meaningful signal of coexisting interpretations—something to be modeled directly as part of the NLI task.

3.1 Ambiguity in Sentence Meaning

Ambiguity in sentence meaning—manifesting as multiple plausible interpretations—is a major source of disagreement in NLI annotations. In the taxonomy introduced by Jiang and de Marneffe (2022), this form of content-based ambiguity is further divided into five subtypes: Lexical, Implicature, Presupposition, Probabilistic Enrichment, and Imperfection. Together, these categories reflect the range of interpretative uncertainty that arises from the language content itself, independent of annotator knowledge or instructions specified in annotation guidelines.

`Lexical` arises when a word or phrase in the premise or hypothesis has multiple possible senses or is underspecified. `Implicature` refers to cases where the hypothesis expresses a logical or pragmatic implication of the premise, leaving room for divergent judgments depending on the reader's perspective. `Presupposition` covers instances where the hypothesis draws on background presuppositions introduced by the premise, which may or may not be universally shared. `Probabilistic Enrichment` denotes cases where the inference relationship is not categorical, but depends on plausibility or likelihood, producing variation in individual perception (Figure 3). `Imperfection` includes typos, grammatical errors, or fragmented phrasing that impede clear interpretation.

While this categorization is based on a manually analyzed sample and shaped by the judgments of linguistically trained annotators, it offers a valuable foundation for surfacing and organizing patterns of interpretative variation in NLI. Although it does not capture the full range of ambiguities present in

Probabilistic Enrichment
Premise: "I think this report shows that we have had an inordinately productive and successful year."
Hypothesis: "The report shows that we need to be productive to have a successful year"

Figure 3: Probabilistic Enrichment ambiguity: Annotator's label choice—`Entailment` or `Neutral`—depends on whether the relationship between productivity and success mentioned in the premise is considered plausible or not.

premise-hypothesis pairs, it provides an important starting point for tracing the roots of disagreement in NLI annotations and for recognizing how such divergences arise from legitimate differences in interpretation rather than annotation error.

3.2 Underspecification of Guidelines

Guideline underspecification is another source of annotation disagreement, but unlike content-based ambiguity, it often reflects task design flaws that can be addressed through clearer instructions. Even when the premise-hypothesis pairs are unambiguous, annotators may diverge in how they interpret or apply the labeling instructions if those instructions lack sufficient precision or fail to address edge cases. Jiang and de Marneffe (2022) identify three specific subtypes under this category: Coreference, Temporal Reference, and Interrogative Hypothesis.

Temporal Reference
Premise: "You wake up one bright autumn morning and you're halfway to the subway when you decide to walk to work instead."
Hypothesis: "You wake up early and decide to walk instead of take the subway."

Figure 4: Temporal Reference disagreement: Annotator's label choice—`Contradiction` or `Entailment`—depends on whether the decision is interpreted as happening before or after the commute, respectively.

`Coreference` cases involve unclear assumptions about whether entities in the premise and hypothesis refer to the same thing. Without guidance on how strongly to assume shared reference, annotators may reach inconsistent labels.

`Temporal Reference` arises when it is unclear when the hypothesis should be evaluated. In Figure 4 one annotator might interpret the decision as occurring **before** the commute (`Contradiction`), while another may align it with a decision made **during** the commute (`Entailment`).

The `Interrogative Hypothesis` covers cases where the hypothesis is phrased as a question. Since questions are not truth-apt (i.e., not directly true or false), annotators must infer an implied assertion. Jiang and de Marneffe (2022) focus only on interrogative hypotheses, while others (e.g., Gubelmann et al. (2023)) argue that interrogative premises can cause similar confusion.

These sources of disagreement are worth recognizing, but they stem from instruction gaps rather than genuine interpretive variation—and thus lie outside this paper's primary focus.

3.3 Annotator Behavior

The third pillar of disagreement, according to the Triangle of Reference, is variation in annotation behavior: differences in background knowledge, beliefs, attention, or task interpretation across annotators. While often treated as noise in NLI pipelines, such variation can reflect meaningful differences in how people reason with language. Two annotators may bring different contextual assumptions to the same premise-hypothesis pair, leading to divergent but reasonable judgments.

Jiang and de Marneffe (2022) identify two specific behavioral tendencies that contribute to such disagreement: `Accommodating Minimally Added Content` and `High Overlap`. The first involves hypotheses that add a small amount of plausible but unstated information. Some annotators accept this as implied, while others reject it based on stricter entailment criteria. The second reflects a tendency to judge `Entailment` based on surface-level similarity—lexical or structural. This can lead some to overestimate entailment based on form rather than meaning, while others focus on more subtle semantic distinctions (Figure 5).

High Overlap
Premise: "The sunlight, piercing through the branches, turned the auburn of her hair to quivering gold."
Hypothesis: "The auburn of her hair became golden then the sunlight hit it."

Figure 5: High Overlap disagreement: Annotators may infer `Entailment` between the premise and hypothesis due to the high lexical overlap, while the meaning of the two sentences suggests `Contradiction`.

Some annotation tendencies stem not from errors, but from genuine interpretive variation. For example, `Accommodating Minimally Added Content` reflects meaningful differences, whereas `High Overlap` more likely signals annotation error. This underscores the importance of evaluating annotator behavior carefully, rather than assuming that all variation reflects valid perspectives.

Among the various sources, content-based ambiguity is the most direct and reliable indicator of genuine interpretive divergence. While understanding annotator behavior is useful, our focus is on detecting ambiguity in the language itself. Even so, recognizing behavioral patterns can inform future perspectivist NLI systems that accommodate multiple interpretations. Next, we motivate why content-based ambiguity merits special attention relative to guideline and annotator effects.

4 Why Does Content-Based Ambiguity Deserve Special Attention?

Among the disagreement sources outlined above, content-based ambiguity stands out as the only type that can be systematically addressed through computational modeling without relying on additional information about the annotators or the guidelines they follow. This form of ambiguity originates from the text itself, independent of the annotation process, yet it remains a fundamental driver of divergent interpretations. As such, it represents a root cause of disagreement inherent to natural language, posing a persistent challenge for inference systems aiming for consistent and reliable predictions.

Implicature Ambiguity
Premise: "It hopes to bring on another 25 or 35 people when the new building opens next fall." **Hypothesis:** "They already have a waiting list for the new building"

Figure 6: Implicature Ambiguity: Annotators may infer a waiting list from *hopes to bring on another 25 or 35 people*. If so, they label it `Entailment`; if not, they label `Neutral`.

Consider the `Implicature` ambiguity in Figure 6. Some annotators interpret *hopes to bring on another 25 or 35 people* as implying a *waiting list* and choose `Entailment`. Others focus strictly on what is stated and select `Neutral`. This interpretative variability stems from linguistic ambiguity rather than annotator background or faulty instructions. Such cases underscore the importance of treating content-based ambiguity as central to analysis, rather than dismissed as noise.

Jiang and de Marneffe (2022)'s findings indicate that the most common sources of disagreement fall under content-based ambiguity, underscoring its prevalence. In contrast, issues related to underspecified guidelines can typically be resolved through clearer instructions and better annotation practices, meaning they do not strongly reflect genuine differences in human interpretation.

Similarly, while some annotator behaviors—such as `Accommodating Minimally Added Content`—reflect natural variation, others like `High Overlap` may undermine the goals of the NLI task. Disagreements stemming from annotator behavior or guideline underspecification therefore warrant scrutiny before being treated as meaningful. Not all disagreements are noise, though some clearly reflect error (Weber-Genzel et al., 2024). In contrast, content-based ambiguity arises from the language itself and requires no external filtering or supervision, making it a uniquely reliable source of interpretive variation. Identifying such ambiguity supports the creation of disambiguated versions, allowing NLI benchmarks to better capture the range of plausible interpretations. Beyond benchmarking, ambiguity-aware modeling has practical consequences in downstream settings.

Identifying content-based ambiguity in NLI data has significant real-world implications. NLI frequently serves as a core component of fact-verification pipelines, where it is used to assess the relationship between claims and supporting evidence (Thorne et al., 2018; Jayaweera et al., 2024). Effectively pinpointing potential ambiguities—including those intentionally introduced—strengthens such pipelines by improving their capacity to flag potentially misleading content in real-world settings (Liu et al., 2023).

However, there are currently no established methods for disentangling disagreements caused by content-based ambiguity from those arising due to underspecified annotation guidelines or annotator behavior. As a result, most existing work focuses primarily on detecting premise-hypothesis pairs with high annotation disagreement, rather than investigating the underlying types of disagreement, particularly those stemming from ambiguity. Therefore, there is a necessity to build models that: (1) identify ambiguous premise-hypothesis pairs and (2) classify the respective ambiguity type. These observations motivate two concrete tasks—ambiguity detection and ambiguity classification—which we discuss next.

5 Understanding Ambiguity in NLI

NLI systems aim to determine the inference relationship between a given premise and hypothesis, but ambiguity in either can complicate that process (Figure 6). This often leads to discrepancies among annotators, who may assign different inference labels based on their individual interpretations. In some cases, annotators may even agree on the same label while interpreting the text differently—a phenomenon known as within-label variation (Jiang et al., 2023). Further complicating matters, ambiguity may arise in the premise, the hypothesis, or both, increasing the complexity of inference decisions (Liu et al., 2023).

While some efforts have been made to develop models that detect instances with high annotator disagreement (Jiang and de Marneffe, 2022; Jiang et al., 2023; Park and Kim, 2025), there are no existing implementations that specifically identify or classify ambiguous instances in NLI—underscoring the need for systems designed to address this gap.

5.1 Ambiguity Detection in NLI

We define *ambiguity detection* in NLI as identifying instances that elicit divergent interpretations due to input ambiguity—whether in the premise, the hypothesis, or both.

Jiang and de Marneffe (2022) explore the detection of high-disagreement instances in NLI using multi-label prediction and a four-class classification scheme (`Entailment`, `Contradiction`, `Neutral`, and `Complicated`). However, their work does not go further to distinguish ambiguity as a specific cause of disagreement. Jiang et al. (2023) build on this by incorporating explanations for disagreement but still focus solely on identifying highly contested instances.

Park and Kim (2025) attempt to detect ambiguous cases in NLI benchmarks using hidden layer representations of Large Language Models (LLMs), but their training data includes disagreements from all categories, making the system a general disagreement detector rather than a model focused on ambiguity. Liu et al. (2023) assess language models' ability to detect ambiguous instances using the Ambient dataset, but their results show that model performance remains below human-level accuracy.

These studies reflect the current state of ambiguity detection in NLI, highlighting the need for further investigation. A key challenge in developing systems to identify ambiguous premise-hypothesis pairs is the lack of datasets annotated for ambiguity. Creating annotated datasets and exploring unsupervised methods are essential next steps.

To address the current scarcity of ambiguity-type annotated NLI data, we leverage existing datasets that already incorporate disambigutations (Liu et al., 2023) and explanations (Jiang et al., 2023) as assistive cues to annotate ambiguity types. This approach would help create a more cohesive dataset that integrates insights across the various taxonomies discussed in Section 5.

At the same time, the limited scale of these resources highlights the need for additional strategies. Promising directions include data augmentation techniques such as paraphrasing, the continued use of manual annotation to ensure high-quality gold standards, and the strategic use of large language models (LLMs) as evaluators. Together, these methods can substantially expand the availability of annotated data, enabling both broader coverage of ambiguity types and more robust evaluation of ambiguity-aware NLI systems.

5.2 Ambiguity Classification in NLI

Ambiguity classification identifies the exact type(s) of ambiguity present in a premise-hypothesis pair. Several taxonomies have been developed to categorize the various forms of ambiguity found in NLI inputs. As noted and illustrated in Figure 2, Jiang and de Marneffe (2022) present a taxonomy comprising five ambiguity types. These have been identified in samples from the ChaosNLI (Nie et al., 2020) and MNLI (Williams et al., 2018) datasets.

Liu et al. (2023) introduce a taxonomy based on expert linguistic annotations of the Ambient dataset, which contains both curated and generated ambiguous premise-hypothesis pairs. They identify additional ambiguity types in NLI data, including `Syntactic`, `Pragmatic`, `Scopal`, and `Figurative` ambiguities, while grouping others under a residual `Other` category.

Building on this, Li et al. (2024) refine the classification by proposing finer-grained types such as Type/Token and Collective/Distributive, and aligning with Jiang and de Marneffe (2022) through the inclusion of Presupposition and Implicature. These refinements reveal further unexplored ambiguities that enhance the understanding of human interpretations. Drawing on these developments, we present a unified taxonomy,[1] that organizes ambiguity types into four broad categories—Lexical, Syntactic, Semantic, and Pragmatic—to support a more comprehensive view (Figure 7).

[1]Refer to (Li et al., 2024) for the definitions of each ambiguity type not described in this paper.

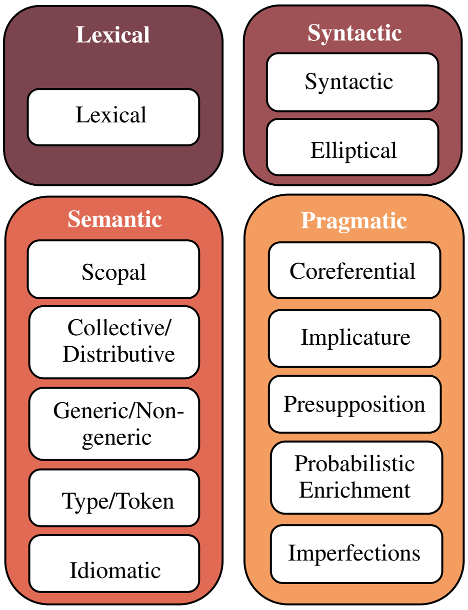


Figure 7: Unified ambiguity type taxonomy: We build on prior taxonomies (Jiang and de Marneffe, 2022; Liu et al., 2023; Li et al., 2024), organizing them into four broad types—Lexical, Syntactic, Semantic and Pragmatic—to support tailored detection strategies based on common characteristics.

However, to our knowledge, no existing system automatically identifies the ambiguity type(s) present in a given premise–hypothesis pair. This gap highlights the practical importance of our framework: by organizing ambiguity types into linguistically grounded categories, we lay the foundation for developing detection methods tailored to each type. In doing so, we advance toward more nuanced, interpretable NLI models that not only detect ambiguous input, but also explain *how* and *why* human interpretations diverge.

6 Call to Action: Recognizing Ambiguity as Signal, Not Noise

Disagreement among annotators in Natural Language Inference (NLI) is often treated as noise—something to minimize or discard. However, many of these disagreements reflect genuine interpretive differences, often triggered by ambiguity in the premise, hypothesis, or both. Our analysis suggests that while underspecified annotation guidelines and inconsistent annotator behavior can lead to label disagreement, such cases must be carefully scrutinized to distinguish between annotation errors and genuine differences in human interpretation.

In contrast, ambiguity in the NLI input itself—whether lexical, syntactic, semantic, or pragmatic—serves as a clear, process-independent signal of interpretive variation, providing a basis for understanding how meaning can diverge across readers. This highlights a needed shift: from optimizing for annotator consensus to explicitly identifying and characterizing ambiguity as a central feature of natural language.

To support this shift, we outline the following two key directions:

- **Identify Ambiguous Pairs:** Develop robust methods to detect premise–hypothesis pairs that exhibit inherent ambiguity, using cues from linguistic theory, annotation patterns, and interpretability tools.
- **Classify Ambiguity Types:** Design strategies for distinguishing among different types of ambiguity. A unified classification framework that groups ambiguity types based on their shared characteristics can offer a foundation for designing targeted identification methods.

A key novelty of this work lies in articulating a unified framework (Figure 1) that extends beyond prior approaches focused solely on modeling annotation distribution. Whereas earlier efforts largely stop at detecting high-disagreement instances, this framework explicitly distinguishes ambiguity from other sources of variation by leveraging linguistic features and pertinent background knowledge.

The framework consists of four stages. The first determines whether a premise-hypothesis pair is inherently ambiguous, thereby distinguishing between genuine interpretive variation from annotation noise and other sources of disagreement. If an instance is ambiguous, the second classifies the type(s),creating systematic linkages to linguistic background knowledge.

The third stage generates relevant disambiguated versions, after which the fourth models inference classification for both ambiguous and non-ambiguous instances, based on predictions from earlier stages. The framework's strength lies in offering a structured and operational foundation for ambiguity-aware NLI, moving the field from descriptive accounts of disagreement toward a principled methodology that can be empirically tested.

By pursuing these goals, we can build NLI models that are not only more aligned with human interpretation, but also more explainable in predictions. Ambiguity-aware systems better align with human interpretation and produce more consistent, interpretable, and robust predictions.

This reframing is not only timely—it is essential for developing NLI systems that reflect the complexity of human understanding, rather than abstracting it away.

While we advocate for a shift toward ambiguity-aware NLI systems, realizing this vision is currently constrained by a key limitation: the lack of datasets that are explicitly annotated for ambiguity and categorized by ambiguity type. Most existing NLI datasets are not designed with interpretive variation or ambiguity classification in mind, making it difficult to systematically identify and analyze ambiguous instances or to evaluate models on their ability to handle them.

This gap limits the development and benchmarking of methods for detecting and classifying ambiguity. The absence of gold-standard annotations for different ambiguity types hinders progress in training and evaluating models that aim to align more closely with human interpretive processes.

To address this, we suggest two complementary directions. First, there is a clear need for the **creation of new datasets** specifically annotated for ambiguity presence and type. Such resources would lay the groundwork for both empirical analysis and model development. Second, we see promise in **exploring unsupervised or weakly supervised methods** that can surface potential ambiguities without requiring extensive manual labeling. Techniques leveraging patterns of annotator disagreement, discourse features, or model uncertainty could offer scalable alternatives in the absence of annotated data.

Despite current limitations, these strategies offer a promising path toward building NLI systems that better reflect the complexity and nuance of human language understanding.

Limitations

The framework we articulate for ambiguity-aware NLI systems establishes a theoretical foundation, with empirical validation remaining an important next step. As a position paper, our aim is to stimulate discussion and motivate future empirical work. The framework's logical rigor and integration of existing taxonomies offer a strong basis for future experimentation and evaluation.

Future research must address the creation of datasets explicitly annotated for ambiguity, alongside the development and evaluation of systems to identify ambiguous instances in NLI. Such efforts will contribute to a deeper understanding by identifying indicators of different ambiguity types, and characterizing how they shape inference judgments. We also anticipate exploring hybid approaches that combine linguistic analysis with large language models to advance ambiguity detection for NLI.

Acknowledgments

This work would not have been possible without the generous startup support provided by Dr. Herbert Wertheim through the Herbert Wertheim College of Engineering at the University of Florida.

References

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. Number: 1.

Samuel Bayer, John Burger, Lisa Ferro, John Henderson, and Alexander Yeh. 2005. MITRE's Submissions to the EU Pascal RTE Challenge. In *Proceedings of the Pattern Analysis, Statistical Modelling, and Computational Learning (PASCAL) Challenges Workshop on Recognising Textual Entailment*, pages 41–44.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain Natural Language Inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Proceedings of the First international conference on Machine Learning Challenges: evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, pages 177–190, Berlin, Heidelberg. Springer-Verlag.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Reto Gubelmann, Aikaterini-lida Kalouli, Christina Niklaus, and Siegfried Handschuh. 2023. When Truth Matters - Addressing Pragmatic Categories in

Natural Language Inference (NLI) by Large Language Models (LLMs). In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 24–39, Toronto, Canada. Association for Computational Linguistics.

Chathuri Jayaweera, Sangpil Youm, and Bonnie J Dorr. 2024. AMREx: AMR for Explainable Fact Verification. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 234–244, Miami, Florida, USA. 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. Place: Cambridge, MA Publisher: MIT Press.

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

Margaret Y. Li, Alisa Liu, Zhaofeng Wu, and Noah A. Smith. 2024. A Taxonomy of Ambiguity Types for NLP. *arXiv preprint*. ArXiv:2403.14072 [cs].

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023. We're Afraid Language Models Aren't Modeling Ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.

Christopher D. Manning. 2006. LOCAL TEXTUAL INFERENCE : IT'S HARD TO CIRCUMSCRIBE , BUT YOU KNOW IT WHEN YOU SEE IT - AND NLP NEEDS IT.

Johannes Mario Meissner, Napat Thumwanit, Saku Sugawara, and Akiko Aizawa. 2021. Embracing Ambiguity: Shifting the Training Target of NLI Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 862–869, Online. 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.

Boyuan Pan, Yazheng Yang, Zhou Zhao, Yueting Zhuang, Deng Cai, and Xiaofei He. 2018. Discourse Marker Augmented Network with Reinforcement Learning for Natural Language Inference. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 989–999, Melbourne, Australia. Association for Computational Linguistics.

Hancheol Park and Geonmin Kim. 2025. Where do LLMs Encode the Knowledge to Assess the Ambiguity? In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 445–452, Abu Dhabi, UAE. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694. Place: Cambridge, MA Publisher: MIT Press.

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.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for Fact Extraction and VERification. *arXiv preprint*. ArXiv:1803.05355.

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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

Xiang Zhou, Yixin Nie, and Mohit Bansal. 2022. Distributed NLI: Learning to Predict Human Opinion Distributions for Language Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 972–987, Dublin, Ireland. Association for Computational Linguistics.

Balancing Quality and Variation: Spam Filtering Distorts Data Label Distributions

Eve Fleisig*[1], Matthias Orlikowski*[2], Philipp Cimiano[2], and Dan Klein[1]

[1]UC Berkeley, {efleisig,klein}@berkeley.edu
[2]Bielefeld University, {morlikowski,cimiano}@techfak.uni-bielefeld.de

Abstract

For datasets to accurately represent diverse opinions in a population, they must preserve variation in data labels while filtering out spam or low-quality responses. How can we balance annotator reliability and representation? We empirically evaluate how a range of heuristics for annotator filtering affect the preservation of variation on subjective tasks. We find that these methods, designed for contexts in which variation from a single ground-truth label is considered noise, often remove annotators who disagree *instead* of spam annotators, introducing suboptimal tradeoffs between accuracy and label diversity. We find that conservative settings for annotator removal (<5%) are best, after which all tested methods increase the mean absolute error from the true average label. We analyze performance on synthetic spam to observe that these methods often assume spam annotators are less random than real spammers tend to be: most spammers are distributionally indistinguishable from real annotators, and the minority that are distinguishable tend to give fixed answers, not random ones. Thus, tasks requiring the preservation of variation reverse the intuition of existing spam filtering methods: spammers tend to be *less* random than non-spammers, so metrics that assume variation is spam fare worse. These results highlight the need for spam removal methods that account for label diversity.

1 Introduction

Because spam responses are common on crowdsourcing sites, researchers need reliable ways to filter out low-quality data. Many of these methods aim to find annotators with unusual labeling behavior. However, a growing body of work has found that information from annotators with minority opinions can be a valuable source of information, since this disagreement helps to understand variability in the opinions of a population, identify cases where some annotators may be better- or worse-informed, or reveal ambiguity in the task. How can we preserve the opinions of annotators who disagree, while still removing spam annotations?

We examine the effects of applying several common methods for discounting spam annotators based on their labeling behavior. Despite the existence of spam removal methods that use attention checks or metadata (e.g., time spent on task), filtering based on labeling behavior remains common practice (Klie et al., 2024); thus, weaknesses in these methods risk affecting a wide range of common machine learning tasks. We test three of these methods—MACE (Hovy et al., 2013), CrowdTruth (Aroyo and Welty, 2014), and inter-annotator agreement metrics—on relatively subjective tasks and analyze effects on variability in the filtered data. We find that, although many methods are near-indistinguishable in terms of their *accuracy* at classifying spam annotators, some are far more likely to remove non-spam annotators who disagree. Furthermore, we find that under most tested methods, removing more annotators degrades the variety of opinions expressed, without improving accuracy at removing spam annotators; thus, these methods seem most effective only when a very low number of annotators are removed.

We also find that assumptions about the distribution of spam annotations can hinder the effectiveness of these methods. We examine performance on synthetic distributions of spam annotations to analyze whether these methods effectively remove spam annotations, or simply remove annotations farther from the mean. Performance on synthetic spam indicates that most methods perform far better for random spam (e.g., randomly clicking answers) than fixed spam (e.g., always answering "No"). Yet true spammer behavior exhibits the opposite trend: most spammers are distributionally

*Equal contribution; order determined by coin flip.

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 47–62
November 8, 2025 ©2025 Association for Computational Linguistics

similar to high-quality annotators; the minority that can be reliably identified tends to have fixed spamming behavior. As a result, methods that perform poorly on fixed spam tend to also perform poorly on real spam.

Our results indicate that spam detection for subjective problems flips model assumptions: spam annotators are often less random than non-spam ones. Thus, attempts to remove spam can backfire by instead removing annotators with minority opinions who are not spammers. As a result, existing methods work best when only low percentages of annotators are removed based on their labeling behavior. When over-filtering for spam, these methods risk distorting the distributions of labels.

2 Related Work

Methods for spammer removal impact the variation in resulting disaggregated datasets, as discussed in work on *spammer detection and aggregation methods* and studies underscoring the role of *subjectivity and variation in annotation*.

Defining Spammers in Annotation. Drawing a conceptual boundary between spammers and genuine annotators is complex; definitions vary on what range of intentional, inattentive, or low-effort behaviors should be filtered out; and on whether spammers are posited as too random or too fixed. Buchholz and Latorre (2011) highlight that spammers are incentivized to earn more money faster, leading them to ignore task instructions or participation requirements. Rothwell et al. (2015) argue that spammers act with intention, unlike other types of low-quality annotators, and show *repeated patterns* in an attempt to complete tasks fast. In contrast, Raykar and Yu (2012) posit that spammers *assign labels randomly*, because they do not follow labeling criteria, skip reading the instances or might use automation. Gadiraju et al. (2015) present a nuanced taxonomy of annotator types and underscore that genuine annotators' behavior might overlap with spammers, e.g., failing attention checks for innocuous reasons. The datasets used in our study excluded annotators if they failed data quality checks combining multiple sources of information, thus following a wider definition of spam (Aroyo et al., 2023; Huang et al., 2023, see Section 3). To summarize, any definition of "spammer" includes or excludes different subsets of annotators. These ambiguous boundaries suggest that different subsets of spammers may exhibit different behaviors, potentially raising challenges in distinguishing spammers from non-spammers.

Spammer Detection and Gold Label Aggregation. Data quality and questionable trust in non-expert raters are longstanding problems in crowdsourced annotation (Snow et al., 2008). Attempts to improve data quality may modify tasks to attract less spam before data collection (Eickhoff and de Vries, 2013) or use quality control afterwards (Difallah et al., 2012). Methods for a posteriori detection of low-quality raters and spammers often use intrinsic metrics based on the labeling behavior itself (Buchholz and Latorre, 2011). Intrinsic metrics used for spammer detection include clustering on a post-processed annotation matrix (Traganitis and Giannakis, 2021), rater similarity and agreement scores (Ak et al., 2021), or distance between sequential spamming behaviors (Ba et al., 2024), among others (Ipeirotis et al., 2010; Raykar and Yu, 2012; Gadiraju et al., 2015). Other methods analyze labeling behavior with the goal of *aggregating to the true label* while accounting for varying annotator reliability. Dawid and Skene (1979) model annotator error rates to estimate the true labels and are foundational to many subsequent aggregation methods (Whitehill et al., 2009; Welinder et al., 2010), including in NLP (Wiebe et al., 1999). Passonneau and Carpenter (2014) present a probabilistic variant of the Dawid & Skene model, and many other extensions of this basic model exist (Paun et al., 2018, 2022). In particular, Hovy et al. (2013) present MACE, a probabilistic model tailored towards estimating annotator competence by modeling spamming behaviors. In contrast, CrowdTruth, a non-probabilistic paradigm, derives quality metrics from vector space representations of annotators, annotated examples and annotations (Aroyo and Welty, 2014; Dumitrache et al., 2018b). We evaluate MACE and CrowdTruth as they underwent widespread adoption in NLP and have reference implementations available (see Sections 4.1, 4.2).

Subjectivity and Variation in Annotation. There is a growing body of work researching informative disagreement, diversity of perspectives, and label variation in human annotation (Plank, 2022; Leonardelli et al., 2023; Sandri et al., 2023; Frenda et al., 2024; Fleisig et al., 2024). These works agree that aggregating labels into a single truth is an oversimplification for many tasks (Aroyo and Welty, 2015; Uma et al., 2021; Basile et al., 2021) and might not represent perspectives fairly (Abercrombie et al., 2022). Instead, studies release annotator-level labels (Prabhakaran et al., 2021) to enable alternative approaches, such as modeling individual annotators' rating behaviors (Fleisig et al., 2023; Orlikowski et al., 2023; Heinisch et al., 2023; Orlikowski et al., 2025). Our work is motivated by studies on *rating distributions in a given population* as an alternative to single ground truth prediction (Sorensen et al., 2024; Meister et al., 2025). Among these, Prabhakaran et al. (2024) study systematic disagreement using similar metrics to ours, but on the level of demographic subgroups. In this context, the issue of how capturing labeling variation intersects with annotation quality is largely unexplored. One exception is VariErr (Weber-Genzel et al., 2024), an annotation methodology to differentiate between annotation errors and plausible variation in annotation. In contrast, we study properties of methods that determine annotator reliability, not individual annotation errors.

3 Datasets

We selected two datasets for the basis of our experiments: DICES 350 (Aroyo et al., 2023) and Huang et al. (2023)'s survey of Amazon Mechanical Turk workers. We present each dataset's statistics and discuss our dataset selection process below.

DICES-350 DICES-350 (Aroyo et al., 2023), a harmful language dataset, consists of 43,050 annotations on a 3-point scale across 350 items, each of which was labeled by every participant. 123 annotators participated, of whom 19 annotators were labeled as spam (15% of annotators).

MTurk From Huang et al. (2023)'s survey, we used 16 questions on a 7-point scale, each of which was answered by every participant, for a total of 3,312 annotations. 207 annotators participated, of whom 40 were labeled as spam (19% of annotators).

Dataset selection. Our experiments require datasets that retain (a) responses from multiple annotators per question, permitting measurement of disagreement statistics, and (b) responses from known spammers. Despite increasing availability of annotator-level data,[1] most public datasets do not include spammer information. For papers that report spammer removal, we contacted authors for access to the unfiltered datasets, but spammer responses are regularly lost over time (e.g., Buchholz and Latorre, 2011; Dumitrache et al., 2017; Paun et al., 2018). Even for published data, maintaining data access is not always possible; some datasets with verified spammer information were no longer available (e.g., Soberón et al., 2013; Gadiraju et al., 2015). Similarly, many studies on spammer detection evaluate only on downstream performance or exclusively use synthetic data, so they do not provide metadata on known natural spammers (e.g., Raykar and Yu, 2012; Ak et al., 2021). See Appendix D for details on all 22 considered datasets, including spammer metadata and data availability. In summary, DICES-350 and the MTurk survey are, to the best of our knowledge, the only available datasets meeting our criteria. Nevertheless, these datasets do represent two representative use cases in which preserving rater variation is essential. DICES-350 collects annotations on AI safety to preserve variation on diverse perspectives regarding high-stakes topics; the MTurk dataset polls workers on their personal opinions about their crowdwork experiences in order to best understand the range of opinions of the community.

4 Methods for Spammer Detection

We study a number of established methods and baselines to calculate scores of annotator reliability. To perform spammer detection, we rank annotators using the respective reliability score and identify the k lowest-scoring annotators as spammers for a given value of k.

4.1 Multi-Annotator Competence Estimation (MACE)

MACE (Hovy et al., 2013) is based on a probabilistic model of annotation. We highlight a few aspects of MACE that are important to our study and refer to the original paper for full details. The model includes a parameter θ for each annotator which encodes the probability that they give the true answer (*competence*). Specifically, for each instance i and annotator j, the binary variable S_{ij} indicates whether an annotator is spamming. S_{ij} is drawn from a Bernoulli distribution with parameter $1 - \theta_j$. If the annotator is spamming, i.e., $S_{ij} = 1$, then the assigned label A_{ij} is sampled from a multinomial distribution with a parameter vector ζ_j that encodes each annotator's spamming strategy. Otherwise, if $S_{ij} = 0$, the model assumes that the

[1] For example, https://github.com/mainlp/awesome-human-label-variation

annotator simply assigns the correct label—an intentional simplification to focus on modeling spam behavior. Only the annotations A_{ij} are observed; the other parameters are inferred when updating the model from data.

Usually, when applying MACE for label aggregation, the model would weigh all annotations to estimate the correct labels without discarding specific annotators. But the learned parameters can also be used to identify spamming annotators: the competence θ correlates more strongly than agreement measures with an annotator's fraction of correctly annotated examples (Hovy et al., 2013) and both learned annotator parameters (θ, ζ) were shown to encode characteristic spamming behaviors (Paun et al., 2018). Consequently, other studies have used MACE to exclude spammers during dataset construction based on an empirically chosen threshold for competence (Pei and Jurgens, 2023). In our experiments, we also use the competence parameter to score annotators.

4.2 CrowdTruth

The CrowdTruth framework (Aroyo and Welty, 2014) computes several interdependent quality metrics that use vector representations of annotations to measure disagreement and ambiguity, including a worker quality score. The metrics follow the aim of ambiguity-aware label aggregation, so that, for example, disagreement on ambiguous instances discounts worker quality less. The worker quality score (WQS) for an annotator i is computed as the product of two other scores $WQS(i) = WUA(i) \cdot WWA(i)$, the worker-unit agreement (WUA) and the worker-worker agreement (WWA). Conceptually, WWA measures how similar a given worker's annotations are to other workers, weighted by the workers' quality and the instances' ambiguity. WUA measures how much a worker agrees with the aggregate label over all their annotated instances, weighted by the instances' ambiguity. (See Appendix A for details on how these metrics are computed.)

The CrowdTruth metrics were explored on various tasks (Dumitrache et al., 2017, 2018a) and have been explicitly used for spammer removal (Dumitrache et al., 2021). In a related study, Soberón et al. (2013) report an accuracy of 0.88 for removing spam annotators using CrowdTruth metrics. In our experiments, we use the worker quality score (WQS) to score annotators.

4.3 Cohen's Kappa

As a representative example of using inter-annotator agreement metrics to filter annotators, we compute each annotator's pair-wise agreement as measured by Cohen's kappa (Cohen, 1968) with each other annotator. We then use the averaged agreement to score annotators.

4.4 Random Baseline

We assign scores to annotators (from 0.0 to 1.0) by drawing from a uniform distribution.

5 Results

We applied MACE, Crowdtruth, the Cohen's kappa filter, and a random baseline on both datasets, as the threshold for number of annotators removed increases. For studies in spammer detection and gold label aggregation (see Section 2) the primary metric to optimize is downstream classification performance, often based on synthetic spam annotations, whereas we focus on tasks where preserving labeling variation is key. We measured the change in standard deviation, entropy, and accuracy of spammer detection for the DICES-350 and Mturk datasets, as well as the KL-divergence and mean absolute error of the filtered labels from the labels of non-spammers.

5.1 Accuracy vs. Preserved Variation for Spam Detection

Across methods, increasing the number of removed annotators gradually decreases the accuracy of classifying annotators as spammers (Figure 1, top). For the MTurk dataset, the accuracy of spam classification never rises above the accuracy of not removing any annotators. For the DICES dataset, Cohen's kappa and MACE outperform removing zero annotators when <10% of annotators are removed, while CrowdTruth and random removal quickly fall below baseline accuracy (Figure 1, bottom). The best accuracy is achieved when only focusing on the lowest-scoring annotators (lowest 2-4%).

We also measured the change in entropy and standard deviation of the filtered dataset, finding that these methods typically reduce variance in the distribution of annotator opinions, discarding information about annotator disagreement (Figures 2 and 3). Except for the random baseline, the tested methods generally decrease the entropy of the distributions as more raters are removed. This is especially true of CrowdTruth, which quickly decreases

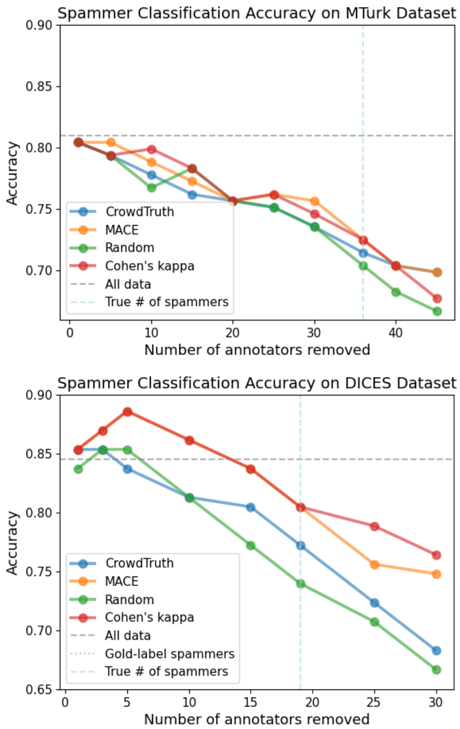


Figure 1: Across methods, increasing the number of removed annotators gradually decreases the accuracy of spam classification when over 2-4% of annotators are removed. Cohen's kappa and MACE increase the spam classification accuracy up to 4% of annotators removed on DICES; otherwise, the spam classification accuracy rarely rises above the baseline of not removing any annotators. The blue line indicates the true number of spammers in the data; the gray line indicates the baseline classification accuracy before removing any spammers.

the entropy; MACE and Cohen's kappa also decrease the entropy to a lesser extent. CrowdTruth also consistently decreases the standard deviation of the data. MACE and Cohen's kappa decrease the standard deviation on the MTurk dataset, but not on DICES.

To understand whether these methods affect how well the filtered datasets represent the true distribution of non-spam annotators' ratings, we also measured the mean absolute error (MAE) per example between the filtered annotators and the true non-spam annotators (i.e., the difference between their average labels on a given example; Figure 5) and the KL-divergence between the filtered and non-spam annotators (Figure 6). All tested methods eventually increase the mean absolute error, indicating that the mean label of the filtered data drifts away from that of the true non-spam annotators as more labels are removed. However, the

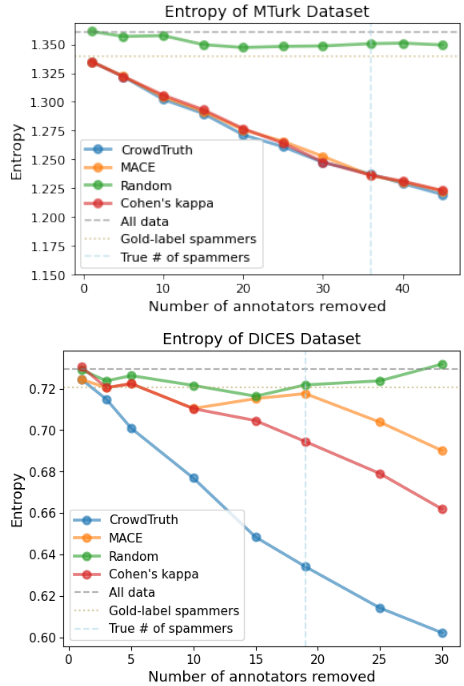


Figure 2: Entropy of each instance's label distribution, averaged over all instances. Most methods decrease the entropy of the dataset as more raters are removed. CrowdTruth especially decreases the entropy.

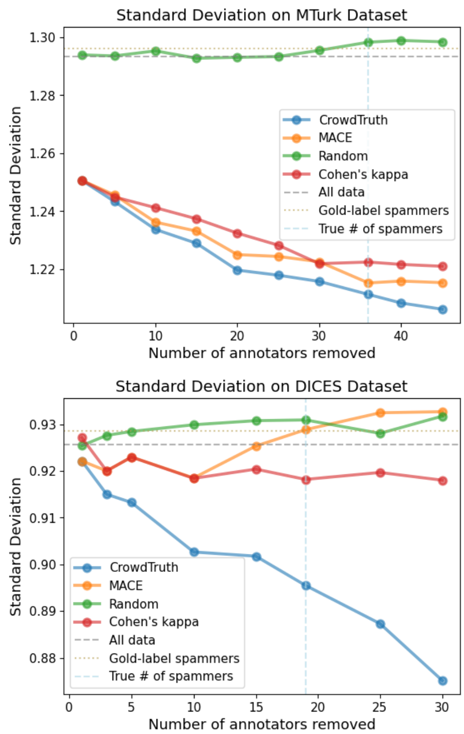


Figure 3: On the MTurk dataset, all methods except random removal decrease the standard deviation of the dataset. Among the tested methods, CrowdTruth decreases the standard deviation most.

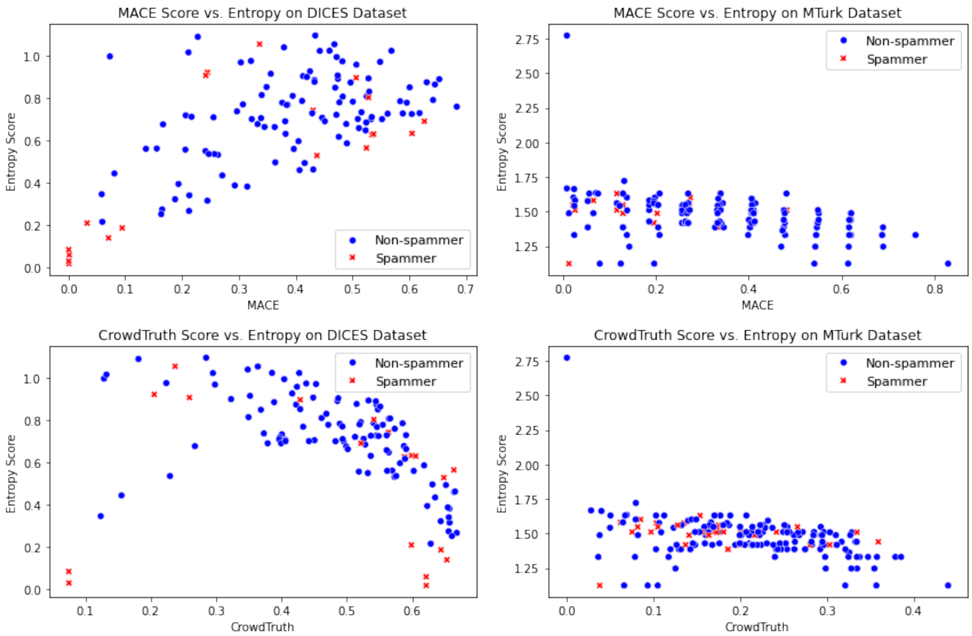


Figure 4: Entropy of each annotator's labeling distribution over all instances vs. score under filtering metrics (CrowdTruth and MACE). While many spam annotators are indistinguishable from non-spam ones under these metrics, those that are often have very low entropy: they are *less* random than non-spam annotators, not more.

extent of this varies by method and dataset: on the MTurk dataset, all non-random methods have relatively little change in MAE when <5% of annotators are removed, but increases after that; on the DICES dataset, CrowdTruth worsens the MAE much faster than other tested methods. The KL divergence remains relatively steady, but eventually increases on the MTurk dataset for all non-random methods, and fluctuates widely across methods on the DICES dataset.

Why might these methods fail to capture all spammers? Comparing the entropy of the responses given by each annotator with their scores under these metrics helps to understand where the assumed spammer behavior, as modeled by these metrics, differs from the spammer behavior seen in practice (Figure 4). Most annotators lie well within the distribution of non-spammers in terms of entropy, MACE score, and CrowdTruth score. However, a subset of annotators are distinguishable as spammers (best seen on the DICES dataset) because they have especially low entropy. MACE captures many of these annotators, but CrowdTruth only captures some of them, perhaps explaining the difference in these metrics.

Since a cluster of spam annotators that can be reliably distinguished tends to have especially fixed behavior, perhaps models perform best at capturing spam if they can identify annotators with unusually fixed annotation patterns. To investigate this, we next studied model performance using synthetic spam.

5.2 Synthetic Spam Analysis

To understand what factors affect spam detection methods' accuracy at classifying spam, and propensity to misclassify annotators who disagree as spam, we experiment with several kinds of synthetic data. *Random spam* experiments simulate spam annotators whose annotations are random; *fixed spam* experiments simulate spam annotators who always give the same answer, which is set to the mode response for the dataset.

Fixed spam. Because these methods tend to filter out annotators who are farther from the mean, most of them struggle to filter out annotators whose behavior is fixed to the mode value (e.g., answering "No" to every question). MACE performs much better than the other methods on fixed spam for DICES, but all methods are worse than the baseline for the fixed spammers on the MTurk data (Figure 7). MACE's higher accuracy on DICES can partially be explained by how well the method can capture fixed spamming behavior given how it is set up (see Section 4.1): A spammer would have low competence θ, so that the assigned label is frequently sampled from the annotator's spamming strategy ζ. As the spammer assigns always the same label,

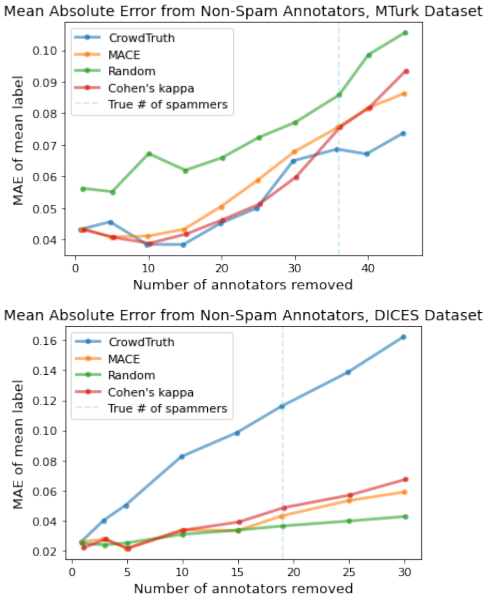


Figure 5: Mean absolute error of filtered ratings. Difference between average label on an example of non-spam annotators and filtered annotators, then averaged across examples.

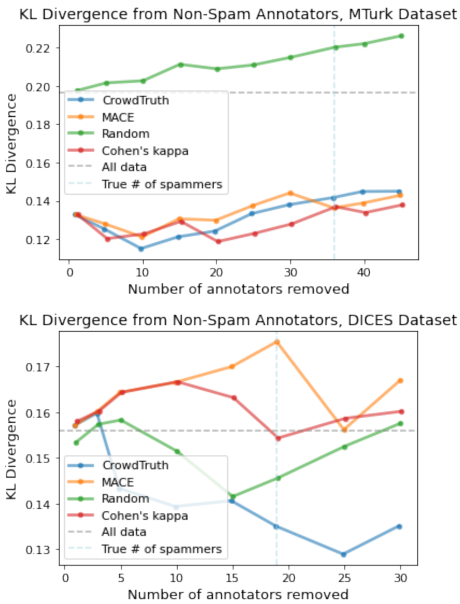


Figure 6: KL-divergence of filtered annotators per data item vs. true non-spam annotators, averaged across examples.

the parameter vector would encode high probability for that particular label and low probability for all others. In contrast, CrowdTruth factors in ambiguity but is ultimately based on agreement (see Section 4.2). As a spammer who always assigns the mode label can score relatively high agreement in subjective tasks with stronger labeling variability, fixed spam annotators are not filtered out by CrowdTruth. This result about agreement for fixed spam is in line with the accuracy scores by Cohen's kappa filtering, which are identical to CrowdTruth. Notably, this observation does not transfer to real spammer behavior (Figure 1), where CrowdTruth is often more accurate than Cohen's kappa.

MACE's poor accuracy on MTurk is surprising given its perfect accuracy on DICES. This result is likely caused by answers in the MTurk dataset mostly following a normal distribution with the same mode, so that MACE overestimates the competence of fixed spammers (in contrast to DICES; see Appendix C).

Because the spammers all give the same ratings, we expect accurate spam classification to increase the standard deviation and the entropy, as happens for MACE on DICES; by contrast, Crowdtruth and Cohen's kappa filtering on DICES (and MACE on MTurk) decrease the standard deviation and the entropy without ever increasing spam classification accuracy above the baseline (Appendix B).

Random spam. On the random data (Figure 7, right), CrowdTruth, MACE, and Cohen's kappa have similar accuracies (peaking when the number of annotators removed equals the number of spam annotators). This suggests that random spam is closest to the spam behavior for which these methods work optimally.

In this case, we expect accurate spam classification to decrease the entropy, which indeed happens for both datasets across methods (Appendix B); the standard deviation also decreases for MTurk, and is more random for DICES, likely because DICES has a smaller set of possible answer values.

Together, these results suggest that real spam annotators are less random than the imagined spammer behavior under CrowdTruth and inter-annotator agreement filtering. This makes these methods vulnerable to removing annotators who are further from the mean rather than actual spammers. MACE, which is more robust to filtering out fixed spammers, also performs better at filtering out real spammers.

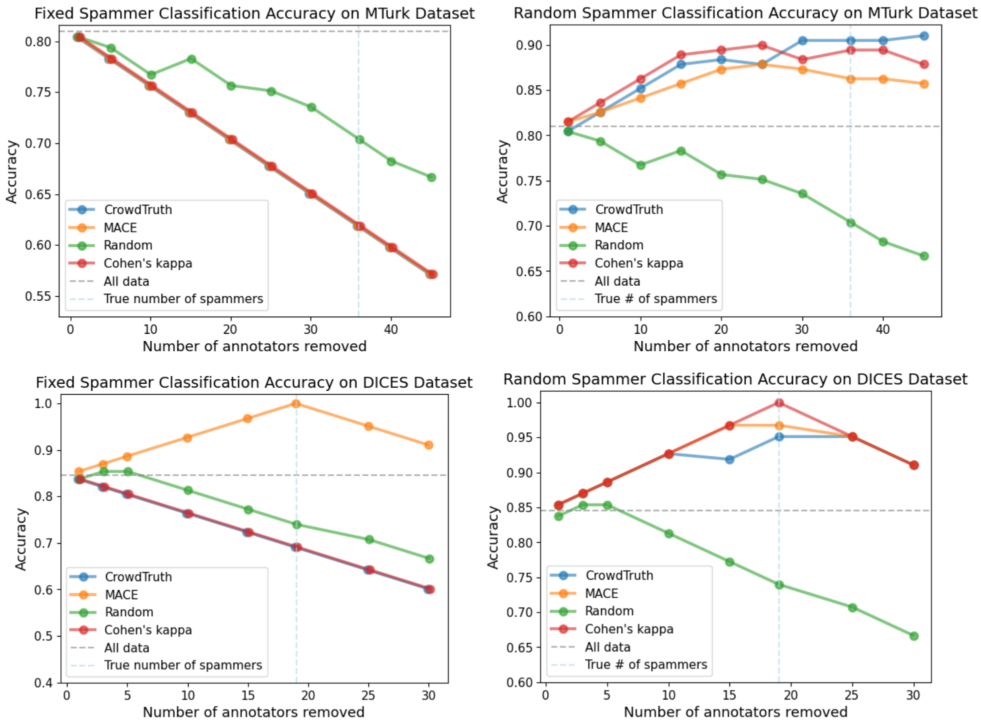


Figure 7: Accuracy with fixed-spam and random-spam synthetic annotators. For DICES, MACE performs best on fixed data; the other methods universally struggle. For random spam, all methods outperform the baseline, with Cohen's kappa performing optimally on DICES.

6 Discussion and Conclusion

Spam detection for subjective problems flips model assumptions: spam annotators are often *less* random than non-spam ones. Longstanding paradigms of annotation, focused on improving downstream model accuracy under the assumption of a single ground truth, often assume that disagreement indicates low-quality annotations. However, in problems where disagreement is expected, and preserving this variation *is* the goal, this intuition is flipped. We find that many spam annotators are indistinguishable from non-spam annotators, and those that are identifiable are in fact those with very low entropy. Examining the performance of tested methods on completely random vs. completely fixed spam reveals that many methods struggle to identify fixed spam. In particular, as a fixed mode response results in relatively high agreement in datasets with substantial variation. These models also struggle on real-world spam in our tested datasets, suggesting that, where preserving variation is paramount, models assuming that spam annotators are more random are not as well suited.

Existing methods work best only when removing few annotators, and distort distributions afterwards. Tested methods (particularly MACE) are effective at identifying spam annotators for low n (<2-4% of tested annotators). When more annotators are removed, we see issues across a range of metrics: increased mean absolute error; lower accuracy at spam detection, lower standard deviation, and lower entropy. These issues mean that overfiltering data can lead to labels that do not fully represent the variation in the original distribution.

Detecting spammers vs. detecting low-quality raters. Since different types of low-quality annotators behave differently, annotators that need to be excluded can exhibit varied behaviors beyond simple patterns such as always selecting the same answer. Consequently, while annotator reliability scoring can often single out spammers showing these stereotypical behaviors, many genuine annotators will be scored similarly to low-quality raters. This result highlights that in addition to the labeling behavior, additional signals should be included in spammer removal. These can be *metadata*, such as when and how much time is spent on annotation (Rothwell et al., 2015) or previous acceptance rates of annotators (Difallah et al., 2012). Similarly, *verifiable test questions* could be used, that is, unambiguous cases where comparison to known answers is possible (gold standard or attention checks, Difallah et al., 2012; Rothwell et al., 2015).

Future work. Existing methods struggle to distinguish spam from non-spam annotators in contexts where variation in opinion is expected and desirable. This gap highlights the need for spam filtering methods that are robust to variation in labeling behavior.

In addition, the scarcity of available metadata on removed spam data makes it difficult to characterize spammer behavior across a range of contexts. Difallah et al. (2012) highlight a "need for new benchmarks on which to evaluate and compare existing and novel spam detection techniques for crowdsourcing platforms" that still persists. Datasets often do not report spam filtering techniques or preserve the spam responses; however, this data is extremely helpful for more fine-grained characterization of spam behavior, especially in complex contexts where variation is expected. Thus, making this data available would be a valuable resource for future research.

Limitations

Due to data scarcity, we only used a narrow range of datasets. While the used datasets represent two important use cases where capturing variation matters (AI safety annotations, survey questions), more datasets are needed, especially with different levels of subjectivity, languages and use cases. As such our results represent only a fraction of relevant scenarios.

Categorizing raters as "spammers" is based on varying definitions and procedures. So "gold spammers" are not ground truth the same way that other data might be. Importantly, self-reported spammer information, where spammers disclose themselves, is largely not even gathered (see for an exception, Paun et al., 2018) and not publicly available. Consequently, the "gold spammer" labels used in our study are based on external categorizations. While these are reported to be based on manual checks and multiple data types (labeling behavior, metadata, and attention checks), there remains a risk of wrong categorizations.

We scoped to spam filtering methods that only look at the labeling behavior, given our research question on how this (widely adopted) type of filtering changes the captured variation in labeling. However, there are approaches based on metadata that we could expect to be more effective, perhaps in combination with the evaluated methods using intrinsic metrics based on labeling behavior.

Acknowledgments

Matthias Orlikowski and Philipp Cimiano were funded by Volkswagen Foundation as part of the "Bots Building Bridges (3B)" project in the "Artificial Intelligence and the Society of the Future" programme. Eve Fleisig is partly supported by an NSF GRFP grant.

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.

Ali Ak, Mona Abid, Matthieu Perreira da Silva, and Patrick Le Callet. 2021. On Spammer Detection in Crowdsourcing Pairwise Comparison Tasks: Case Study on Two Multimedia QoE Assessment Scenarios. In *ICME 2021 - First International Workshop on Quality of Experience in Interactive Multimedia*, Virtual, China.

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342.

Lora Aroyo and Chris Welty. 2014. The Three Sides of CrowdTruth. *Human Computation*, 1(1). Number: 1.

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. Number: 1.

Yang Ba, Michelle V. Mancenido, Erin K. Chiou, and Rong Pan. 2024. Data Quality in Crowdsourcing and Spamming Behavior Detection. *arXiv preprint*. ArXiv:2404.17582 [cs].

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.

Sabine Buchholz and Javier Latorre. 2011. Crowdsourcing Preference Tests, and How to Detect Cheating. In *12th Annual Conference of the International Speech Communication Association, INTERSPEECH 2011, Florence, Italy, August 27-31, 2011*, pages 3053–3056. ISCA.

Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2016. Phrase Detectives Corpus 1.0 Crowdsourced Anaphoric Coreference. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2039–2046, Portorož, Slovenia. European Language Resources Association (ELRA).

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

A. P. Dawid and A. M. Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28. Publisher: Royal Statistical Society, Oxford University Press.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2012. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *CrowdSearch*, volume 842 of *CEUR Workshop Proceedings*, pages 26–30. CEUR-WS.org.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2017. False Positive and Cross-relation Signals in Distant Supervision Data. In *6th Workshop on Automated Knowledge Base Construction (AKBC)*.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018a. Crowdsourcing Semantic Label Propagation in Relation Classification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 16–21, Brussels, Belgium. Association for Computational Linguistics.

Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018b. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement. *Preprint*, arXiv:1808.06080.

Anca Dumitrache, Oana Inel, Benjamin Timmermans, Carlos Ortiz, Robert-Jan Sips, Lora Aroyo, and Chris Welty. 2021. Empirical methodology for crowdsourcing ground truth. *Semantic Web*, 12(3):403–421. Publisher: SAGE Publications.

Carsten Eickhoff and Arjen P. de Vries. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*, 16(2):121–137.

Şeyda Ertekin, Cynthia Rudin, and Haym Hirsh. 2014. Approximating the crowd. *Data Mining and Knowledge Discovery*. Kluwer Academic Publishers.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. 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 2279–2292, Mexico City, Mexico. 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*.

Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1631–1640, New York, NY, USA. Association for Computing Machinery.

Philipp Heinisch, Matthias Orlikowski, Julia Romberg, and Philipp Cimiano. 2023. Architectural sweet spots for modeling human label variation by the example of argument quality: It's best to relate perspectives! In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11138–11154, Singapore. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Olivia Huang, Eve Fleisig, and Dan Klein. 2023. Incorporating worker perspectives into MTurk annotation practices for NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1010–1028, Singapore. Association for Computational Linguistics.

Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 64–67, New York, NY, USA. Association for Computing Machinery.

Aiqi Jiang, Nikolas Vitsakis, Tanvi Dinkar, Gavin Abercrombie, and Ioannis Konstas. 2024. Re-examining sexism and misogyny classification with annotator attitudes. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15103–15125, Miami, Florida, USA. Association for Computational Linguistics.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

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.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2025. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 24–49, Albuquerque, New Mexico. Association for Computational Linguistics.

Lester James Validad Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A. Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. 2025. Hybrid preferences: Learning to route instances for human vs. AI feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7162–7200, Vienna, Austria. Association for Computational Linguistics.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are Large Language Models Reliable Argument Quality Annotators? In *Robust Argumentation Machines*, pages 129–146, Cham. Springer Nature Switzerland.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2111, Vienna, Austria. Association for Computational Linguistics.

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Silviu Paun, Ron Artstein, and Massimo Poesio. 2022. *Statistical Methods for Annotation Analysis*. Synthesis Lectures on Human Language Technologies. Springer International Publishing, Cham.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Aida Mostafazadeh Davani, Alicia Parrish, Alex Taylor, Mark Diaz, Ding Wang, and Gregory Serapio-García. 2024. GRASP: A disagreement analysis framework to assess group associations in perspectives. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3473–3492, Mexico City, Mexico. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vikas C. Raykar and Shipeng Yu. 2012. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *J. Mach. Learn. Res.*, 13:491–518.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR.

Spencer Rothwell, Ahmad Elshenawy, Steele Carter, Daniela Braga, Faraz Romani, Michael Kennewick, and Bob Kennewick. 2015. Controlling quality and handling fraud in large scale crowdsourcing speech data collections. In *Proceedings of Interspeech 2015*, pages 2784–2788. ISCA.

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.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Guillermo Soberón, Lora Aroyo, Chris Welty, Oana Inel, Hui Lin, and Manfred Overmeen. 2013. Measuring crowd truth: disagreement metrics combined with worker behavior filters. In *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web - Volume 1030*, CrowdSem'13, pages 45–58, Aachen, DEU. CEUR-WS.org.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Panagiotis A. Traganitis and Georgios B. Giannakis. 2021. Identifying Spammers to Boost Crowdsourced Classification. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2840–2844. ISSN: 2379-190X.

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

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2025. HelpSteer2: open-source dataset for training top-performing reward models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37 of *NeurIPS '24*, pages 1474–1501, Red Hook, NY, USA. Curran Associates Inc.

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.

Shira Wein and Nathan Schneider. 2022. Accounting for language effect in the evaluation of cross-lingual AMR parsers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3824–3834, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. 2010. The Multidimensional Wisdom of Crowds. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. 2009. Whose vote should count more: optimal integration of labels from labelers of unknown expertise. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, pages 2035–2043, Red Hook, NY, USA. Curran Associates Inc.

Janyce M. Wiebe, Rebecca F. Bruce, and Thomas P. O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 246–253, College Park, Maryland, USA. Association for Computational Linguistics.

Dengyong Zhou, Sumit Basu, Yi Mao, and John Platt. 2012. Learning from the Wisdom of Crowds by Minimax Entropy. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

A Computing the CrowdTruth Worker Quality Score

As highlighted in Section 4.2, WWA measures how similar a given worker's annotations are to other workers, weighted by the workers' quality and the instances' (or units') ambiguity. WUA measures how much a worker agrees with the aggregate label over all their annotated instances, weighted by the instances' ambiguity. WWA and WUA are roughly computed as follows (ignoring the normalization terms for clarity, full details in Dumitrache et al., 2018b):

$$WWA(i) = \sum_{j,u} sim(i,j,u) \cdot WQS(j) \cdot UQS(u)$$

$$WUA(i) = \sum_{u \in units(i)} sim(i,u) \cdot UQS(u)$$

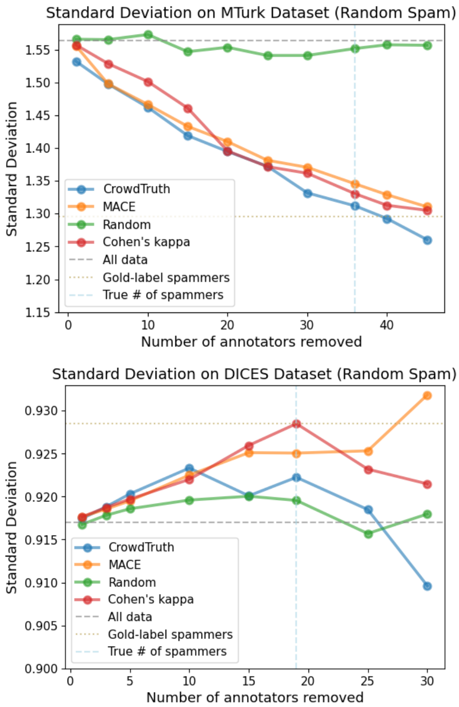


Figure 8: Standard deviation on random spammers.

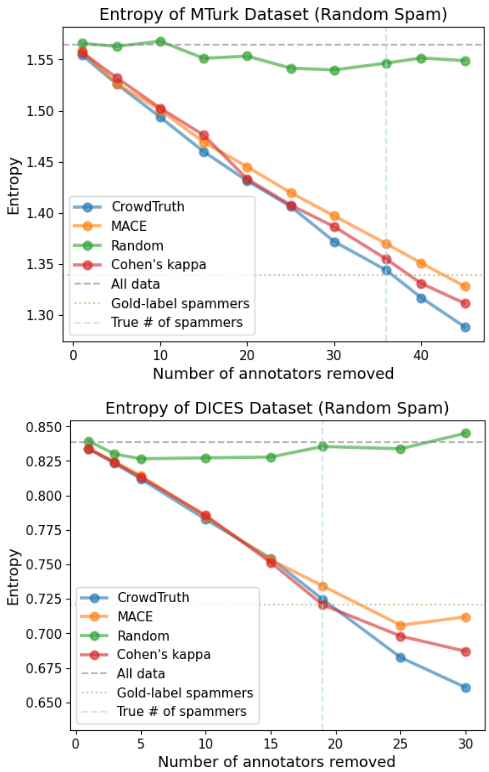


Figure 9: Entropy on random spammers. Entropy generally decreases as more spammers are removed, as expected for accurate spam classification.

Here, $sim(i, j, u)$ is the cosine similarity between the annotation vectors of workers i and j on an instance u. Similarly, $sim(i, u)$ is the cosine similarity between the annotation vector by worker i and the instance vector for instance u (i.e., summed annotation vectors of all other annotators). It is computed over all instances annotated by annotator i, denoted $units(i)$. Additionally, $UQS(u)$ measures how much workers agree on an instance u (how ambiguous it is) and is also connected to the workers' quality scores. Due to their inter-dependent nature, the CrowdTruth metrics are re-calculated iteratively until convergence.

B Details of Synthetic Spam Results

Standard deviation and entropy for the random and fixed spammers are shown in Figure 8, Figure 9, Figure **??**, and Figure **??**.

C Why does MACE fail to recognize fixed spammers on the MTurk dataset?

On fixed spammers, who always respond with the mode (the most frequent label in each dataset), MACE gets perfect accuracy on DICES, while on MTurk it performs as poorly as all other methods, failing to reach baseline performance (see Section 5.2). This result is likely due to the peculiarities of the survey data in the MTurk dataset, where answers follow a normal distribution and the mode is the same for most questions. Here, fixed spammers' seem competent because they always respond with the ground truth as estimated by MACE. Because of this perfect answering behavior of spammers, their average difference to the estimated ground truth is zero, as shown in Figure 10, so that naturally non-spammers are further away from the estimated ground truth, looking less competent to MACE. In contrast, on DICES, which has more varied examples of labeling behavior, non-spammers are on average closer to the estimated ground truth than spammers (see Figure 10).

D Dataset Selection Table

A total of 22 datasets were considered to be included in our study, mostly informed by related work. Table 1 lists all of these datasets, including the corresponding references. The table details for each dataset if gold spammer data was collected in principle and if that data was still available. As described in Section 3, we were only able to include two out of these 22 datasets in our experiments.

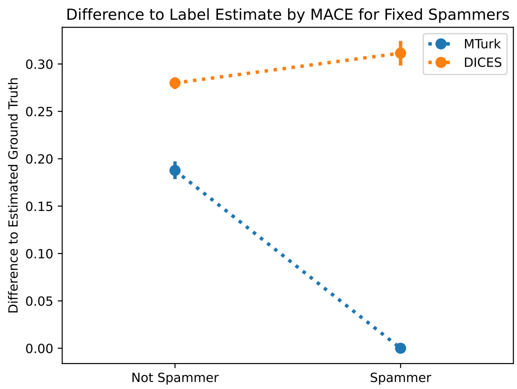


Figure 10: Distance to the ground truth estimated by MACE on fixed spammers vs non-spammers (lower is better). Shows the averaged absolute difference between annotations and the estimated ground truth label. Before averaging, distances are normalized using min-max normalization for each dataset, scaling distances into the range of zero to one.

Dataset	Reference	Gold spammers?	Included?	If excluded, why?
DICES	Aroyo et al. 2023	Yes	Yes	
MTurk Survey	Huang et al. 2023	Yes	Yes	
MHS corpus	Sachdeva et al. 2022	Yes	No	Raters excluded (details in their paper), but data not available.
AdultContent3 ("Get Another Label" datasets)	Ipeirotis et al. 2010	No	No	No gold spammers. Experiments in paper use synthetic data and simply report impact on the collected dataset
HITspam	Discussed in Ertekin et al. 2014	No	No	Despite the name, does not contain spammers. Instead, the task is to judge whether a task on MTurk itself should be considered spam (e.g., because it asks workers to follow a specific social media account).
EDOS-DOM	Jiang et al. 2024	Yes	No	Only one annotator removed after labels were collected. That annotator had annotated only 8 examples (first author vial email).
Argument Quality	Mirzakhmedova et al. 2024	No	No	Excludes a number of disagreeing annotations per example. Does not exclude on the level of the annotator.
MultiPref	Miranda et al. 2025	No	No	No gold spammers.
HelpSteer2	Wang et al. 2025	No	No	No gold spammers.
CrowdTruth Corpus for Open Domain Relation Extraction	Dumitrache et al. 2017	No	No	Emailed first author, full data not available anymore.
AMR / Sentence Similarity Data	Wein and Schneider 2022	Yes	No	Only one annotator removed out of three in total.
Phrase Detectives	Chamberlain et al. 2016	Yes, self-reported	No	Spammer data not available anymore according to authors.
Crowd-Sourced Preference Tests	Buchholz and Latorre 2011	Yes, inferred	No	Data not available anymore according to first author.
VariErr NLI	Weber-Genzel et al. 2024	No	No	Data has annotator IDs and individual decisions plus error judgments (error = no self-validations), but no excluded raters. Not a crowd-sourced study (four annotators).
Malicious Worker Survey Dataset	Gadiraju et al. 2015	Yes	No	Dataset is not available anymore.
Dog (Imagenet Subset)	Deng et al. 2009	No	No	No spammer information. Used by Traganitis and Giannakis (2021), but only evaluated by accuracy of resulting classifier.
ImageNetV2	Recht et al. 2019	No	No	No spammer information.

Bluebird	Welinder et al. 2010	No	No	No spammer information. Used by Traganitis and Giannakis (2021), but only evaluated by accuracy of resulting classifier.
Web	Zhou et al. 2012	Unlikely	No	Could not find reference to data.
WSD	Snow et al. 2008	Unlikely	No	Data not available anymore
RTE	Snow et al. 2008	Unlikely	No	Data not available anymore
TEMP	Snow et al. 2008	Unlikely	No	Data not available anymore
POPQUORN	Pei and Jurgens 2023	Yes	No	Only a single annotator was removed.

Table 1: Dataset Selection. Shows which datasets where considered and why 20 out of 22 datasets were not included in our study.

Consistency is Key: Disentangling Label Variation in Natural Language Processing with Intra-Annotator Agreement

Gavin Abercrombie[1] and **Tanvi Dinkar**[1,] and **Amanda Cercas Curry**[2]
and **Verena Rieser**[1*] and **Dirk Hovy**[3]
[1]Heriot-Watt University [2]CENTAI Institute [3]Bocconi University
g.abercrombie@hw.ac.uk

Abstract

We commonly use agreement measures to assess the utility of judgements made by human annotators in Natural Language Processing (NLP) tasks. While *inter*-annotator agreement is frequently used as an indication of label *reliability* by measuring consistency between annotators, we argue for the additional use of *intra*-annotator agreement to measure label *stability* (and annotator *consistency*) over time. However, in a systematic review, we find that the latter is rarely reported in this field. Calculating these measures can act as important quality control and could provide insights into *why* annotators disagree. We conduct exploratory annotation experiments to investigate the relationships between these measures and perceptions of subjectivity and ambiguity in text items, finding that annotators provide inconsistent responses around 25% of the time across four different NLP tasks.

1 Introduction

Agreement measures are commonly used to assess the utility of judgements made by human annotators for Natural Language Processing (NLP) tasks. Indeed, the reporting of *inter*-annotator agreement (or inter-rater reliability) has long been the standard to indicate dataset quality (Carletta, 1996) and frequently serves as an upper bound for model performance on a task (Boguslav and Cohen, 2017).

While inter-annotator agreement is frequently used in NLP to determine the *reliability* of labels or the processes used to produce them (Artstein, 2017), *intra*-annotator agreement is rarely, if ever, reported. However, we can use it to measure the temporal *consistency* of the annotators who chose the labels and, hence, the *stability* of the labels and data that they generate.[1] Consistency and label stability are important because, without them, annotation schemes are unlikely to be repeatable or reproducible (Teufel et al., 1999).[2]

Such measures of intra-rater agreement are frequently reported in areas of medicine such as physiotherapy (e.g. Bennell et al., 1998; Meseguer-Henarejos et al., 2018), and speech pathology (e.g. Capilouto et al., 2005; Rose and Douglas, 2003). Intra-rater measures are also reported in other fields as diverse as economics (Hodgson, 2008), software engineering (Grimstad and Jørgensen, 2007), and psychology (Ashton, 2000).

However, reporting intra-annotator agreement is so far extremely uncommon in NLP, as we show in a systematic review in Section 2.

Disagreement and label variation in NLP In addition, we argue that the use of inter- and intra-annotator agreement allows us to distinguish and measure different sources of observed label variation (Rottger et al., 2022; Plank, 2022). This is important as NLP researchers have increasingly recognised that, for many tasks, different points of view may be equally valid (Aroyo and Welty, 2015; Basile et al., 2021a; Plank, 2022; Rottger et al., 2022), and that their aggregation can erase minority perspectives (Basile et al., 2021a; Blodgett, 2021).

One of the main challenges in implementing this new paradigm is the interpretation of disagreement. Disagreement between annotators may be due to two sources: 1) genuine differences in their subjective beliefs/perspectives, which can be desirable under this paradigm, or 2) task difficulty, ambiguity, or annotator error, all of which are undesirable. While agreement measures *between* annotators can give us an idea of task **subjectivity**, they provide little insight as to its **difficulty**, **ambiguity,** or the quality and attentiveness of the annotators themselves (Rottger et al., 2022).

*Now at Google DeepMind

[1]We apply the term *consistency* to annotator behaviour and *stability* to labels and datasets.

[2]Although there may be situations in which annotation consistency is not expected, such as longitudinal studies of attitudinal change.

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 63–74
November 8, 2025 ©2025 Association for Computational Linguistics

In the following, we propose the use of *intra*-annotator agreement as a measure of subjectivity.

The reliability-stability agreement matrix What then, does it mean when individual annotators' interpretations are not stable, i.e., internally inconsistent? In addition to providing an additional layer of quality control, we suggest that measurement of label stability can help to interpret potential causes of *inter*-annotator disagreement. To this end, we propose the reliability-stability matrix, a framework for mapping and interpreting the relationship between inter- and intra-annotator agreement in labelled datasets (Table 1).

		Reliability (between annotators)	
		Low *inter*	High *inter*
Stability (temporal within annotator)	High *intra*	Variable perspectives/ High subjectivity	Straight-forward/ Good quality
	Low *intra*	Ambiguous or difficult/ Poor quality	Systematic errors/ Value changes

Table 1: The reliability-stability matrix for *inter*- and *intra*- annotator agreement.

Under this framework, *inter*-annotator agreement and *intra*-annotator agreement, taken together, indicate the task's ambiguity or complexity and its subjectivity level. *Inter*-annotator agreement measures reliability, while *intra*-annotator agreement measures stability. The resulting axes form a confusion matrix that describes four cases.

If both measures are high, we assume the task is unambiguous and simple, and the annotator group relatively homogonous. Presumably, the quality of the guidelines and textual data is also good (Ide and Pustejovsky, 2017). In this scenario, the task or item should be relatively straightforward.

Where both agreement measures are low, we are likely to be faced with a highly ambiguous or difficult task or item–perhaps with multiple equally valid responses–or the annotation quality is poor.

If reliability is low, but consistency is high, the labels likely reflect the annotators' varied but potentially equally valid subjective perspectives.

We do not foresee many situations where reliability is high yet stability/consistency is low. Any agreement between inconsistent annotators would presumably be purely by chance or mass random spamming, i.e., systematic errors. Exceptions could include population-level value shifts over longer time intervals arising from awareness-raising events such as the #MeToo (Szekeres et al., 2020) and #BLM (Sawyer and Gampa, 2018) movements.

Our framework can be applied at the dataset- or item-level by computing any standard agreement metrics.We illustrate this in exploratory annotation experiments described in Section 3.

Our contributions 1) We conduct a systematic review, finding that **a tiny fraction of NLP publications report intra-annotator agreement**; (2) we suggest addition of intra-annotator agreement as a standard measure, and show how **measuring annotator stability could complement existing reliability measures to distinguish reasons for label variation**; and (3) we conduct exploratory longitudinal annotation experiments across four NLP tasks, finding that **annotators provide inconsistent responses for more than 25% of items**, calling into question the implicit assumption that differences in annotation behaviour are seen only between and not within individuals.[3]

2 Intra-Annotator Agreement in the NLP Community

To get a snapshot of the extent to which intra-annotator agreement is reported in the NLP community, we conducted a systematic review of papers published in the *Anthology* of the Association for Computational Linguistics (ACL).[4] Here, we wish to discover for which tasks and what purposes NLP researchers collect and report on repeat annotations and evidence for how and when repeat items should be presented to annotators. Full details of the review methodology are available in Appendix A.

To what extent and why is intra-annotator agreement reported in NLP? When we conducted our study, the search and filtering process returned only 56 relevant publications out of more than 80,000 papers listed in the Anthology. In other words, a tiny fraction (less than 0.07%) of computational linguistics and NLP publications in the repository report measurement of intra-annotator agreement.[5]

The only area of NLP in which intra-annotator agreement is somewhat regularly reported is machine translation (MT), which accounts for

[3] Data available at `https://github.com/HWU-NLP/consistency`.

[4] `https://aclanthology.org/`

[5] We acknowledge that intra-annotator agreement is irrelevant to many papers, but highlight that the number of publications which report it is nevertheless extremely low.

more than half of the included publications. Most of these were agreement measures on human evaluation of translation quality, with one on word alignment annotation for MT (Li et al., 2010). Several other publications on evaluating natural language generation also report measurement on human evaluation tasks (e.g. Belz and Kow, 2011; Belz et al., 2016, 2018; Jovanovic et al., 2005). Other included fields are semantics (e.g. Cao et al., 2022; Hengchen and Tahmasebi, 2021), syntax (e.g. Baldridge and Palmer, 2009; Lameris and Stymne, 2021), affective computing (including sentiment analysis (Kiritchenko and Mohammad, 2017) and emotion detection (Vaassen and Daelemans, 2011)), and automatic text grading (Cleuren et al., 2008; Downey et al., 2011). There is also one paper on abusive language detection (Cercas Curry et al., 2021).[6]

Where the authors motivate the collection of repeat annotations, they usually mention quality control or annotator consistency. Notably, no papers mention the possibility that intra-annotator inconsistency could be valid or informative beyond these factors, as we propose.

Best practice for measuring intra-annotator agreement: how long should the label-relabel interval be? When designing annotation tasks (such as ours in Section 3), it would be helpful to know when to present repeated items, thus avoiding annotators labelling from memory, which may not be an actual test of their consistency.

Over a quarter of the papers (15/56) do not provide enough information to determine the interval between initial and repeat annotations. In most other cases, either it can be inferred, or the authors explicitly state that re-annotations are conducted in the same session as the original annotation. Those that report more extended time before re-annotation leave intervals varying from a few minutes (Kiritchenko and Mohammad, 2017) to a year (Cleuren et al., 2008; Hamon, 2010).

Two papers do specifically investigate the effects of time on annotator consistency. Li et al. (2010) experimented with intervals of one week, two weeks, and one month, comparing intra-annotator agreement for these and finding that consistency on their word alignment annotation degraded steadily over time. Kiritchenko and Mohammad (2017) performed a similar study, comparing intra-annotator agreement on ratings (on a scale) that were conducted with intervals from a few minutes to a few days between the initial and repeat judgements. They too found that inconsistencies increased as a function of increase in interval.

[6]We provide a full list of included papers in Appendix B.

3 Exploratory annotation experiments

We conduct an exploratory annotation experiment to investigate the relationships between agreement measures and the possible reasons for disagreements and inconsistencies. We also investigate whether, as is commonly believed, specific task types are generally more subjective than others.

Hypotheses

At the individual annotation item level, for a given task and dataset:

H1.1 *Subjective* annotation items have lower *inter*-annotator agreement than *straightforward* items, but higher *intra*-annotator agreement than *ambiguous* items.

H1.2 *Ambiguous* annotation items have lower *inter*- and *intra*-annotator agreement than both *straightforward* and *ambiguous* items.

At the dataset/task level:

H2 *Social* tasks—such as offensive language detection and sentiment analysis—are more *subjective* than *linguistic* tasks, like textual entailment or anaphora resolution. That is, stability is higher for social tasks than linguistic tasks.

	Task	Dataset	Labels
Social	Offensive language detection	Leonardelli et al. (2021)	*Offensive/not offensive*
	Sentiment analysis	Kenyon-Dean et al. (2018)	*Positive/negative/objective*
Linguistic	Natural language inference/ textual entailment	Williams et al. (2018)	*Entailment/contradiction/ neutral*
	Anaphora resolution	Poesio et al. (2019)	*Referring/non-referring*

Table 2: Datasets used in the annotation experiments.

Data We use subsets of four English language datasets, see Table 2: two social tasks that are commonly assumed to be subjective, and two linguistic tasks, thought of as objective (Basile et al., 2021b). These were selected because they (1) have limited label sets (of two or three classes), allowing for comparison across tasks; and (2) have been published with non-aggregated (i.e. annotator specific) labels, allowing us to include items with known inter-annotator disagreement in our subsamples. From each dataset, we selected 50 items with high disagreement in the original label sets for re-annotation.

Methodology We recruited crowdworkers from Prolific[7] to annotate a subset of fifty items from each of the tasks/datasets. As much of the text data is primarily sourced from the United States of America and, in some cases,[8] concerns American news stories such as the controversy surrounding the killing of George Floyd,[9] we recruited only annotators located in the US. To obtain high quality annotations, we prescreened participants to ensure that (1) their first language was English, and that (2) they had a 100% approval rate on Prolific.

Based on the evidence of our review (Li et al., 2010; Kiritchenko and Mohammad, 2017), and of more recent work by Abercrombie et al. (2023), we left an interval of two weeks before we recall the annotators to collect a second round of annotations in order to measure their consistency. Of 30 annotators that began the first task, 16 completed both rounds of all four tasks, and we base our results on the labels they provided. All annotators were L1 English speakers; nine were male and eight female; 11 identified as 'White', four as 'Black', one 'Asian', and one 'Mixed'; and ages ranged from 20 to 67; ($\mu = 43.9; s = 14.0$). Annotators were provided with the original instructions pertaining to each task.

We then recruited a second set of expert annotators to annotate the examples that demonstrate internal and or external disagreement with rationalisations for these disagreements, using the labels *ambiguous*. *subjective*, or *straightforward*.

4 Results

We report agreement for each task, and examine differences between the groups of items labelled as *subjective*, *ambiguous*, and *straightforward*.

Overall agreement As *intra*-annotator agreement is typically assumed to be 100% (i.e. by omitting to consider it (Abercrombie et al., 2023)), we measure and raw report percentage agreement as a primary metric to examine whether this holds. For *inter*-annotator agreement, we calculate these pairwise across annotators and report the means. For completeness, we also report Cohen's kappa scores in Appendix C.

	Reliability (*Inter-*) %		**Stability** (*Intra-*) %	
	μ	σ	μ	σ
Offence	68.3	15.4	74.4	15.0
Sentiment	63.6	21.7	69.2	19.5
Entailment	58.6	21.4	72.6	15.1
Anaphora	76.2	14.3	80.5	13.0
Overall	66.7	19.6	74.2	16.3

Table 3: Pairwise reliability and stability of the collected labels measured with mean (μ) and standard deviations (σ) across items for raw percentage inter- and intra-annotator agreement scores.

Agreement scores are presented in Table 8. As expected, agreement is higher for stability than reliability for all tasks, although considerably lower than perfect agreement—just 74.2% overall, and no higher than 80.5% for any task. Individual annotators all have very similar levels of stability: $\mu = 74.2\%; \sigma = 4.3\%; max = 81.5\%; min = 67.5\%$. These results are also remarkably similar to those of Abercrombie et al. (2023), who reported mean intra-annotator agreement of 74.5% on a hate speech identification task conducted over a comparable time frame and on the same recruitment and annotation platforms.

Agreement by task The distribution of annotation items on the reliability-stability matrix is shown in Figure 1. A multivariate Kruskal-Wallis test indicates statistically significant differences between tasks for both variables: for inter-annotator agreement, $H{-}statistic : 12.42, p{-}value : 0.01$; and for intra-annotator agreement, $H - statistic : 10.76, p - value : 0.01$.[10]

[7]https://www.prolific.co/

[8]Particularly in the offensive language dataset.

[9]*The Guardian* April 20 2021 (McGreal, 2021).

[10]Post-hoc pairwise Dunn's tests with Bonferroni correction reveal that only *sentiment-anaphora* and *entailment anaphora* have significantly different distributions for reliability, and only *sentiment-anaphora* for stability.

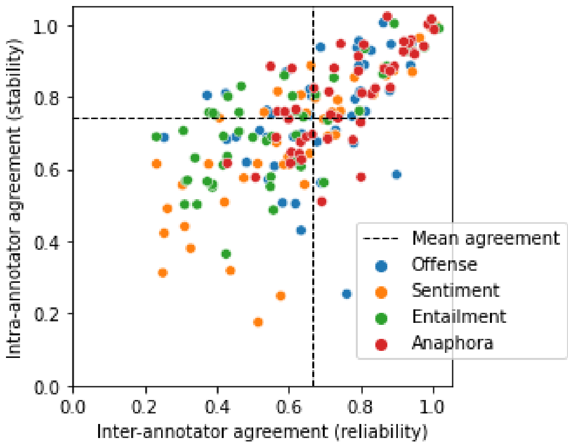


Figure 1: By task raw percentage agreement on individual items for reliability (pairwise) and stability.

However, these differences do not confirm the view that social tasks are more subjective than linguistic tasks (H2). Rather, the *offense* and *anaphora* tasks obtain higher agreement (both *inter* and *intra*) than the *sentiment* and *entailment tasks*, suggesting that, for the particular items in these data samples, the former are simply easier to agree and be consistent on than the latter.

	Bottom-left (*Amb.*)	Top-left (*Subj.*)	Top-right (*Straight.*)	Bottom-right (*Errors*)
Offense	30.0	18.0	38.0	14.0
Sentiment	46.0	10.0	38.0	6.0
Entailment	48.0	20.0	28.0	4.0
Anaphora	22.0	10.0	56.0	12.0
Overall	36.5	14.5	40.0	9.0

Table 4: Percentage of annotation items in each quadrant of the plot in Figure 1.

As Figure 1 and Table 4 show, while the annotation items are predominantly distributed across the bottom-left and top-right quadrants, *sentiment* and *entailment* are skewed to the bottom left, indicating greater ambiguity, and offensive language and entailment tend towards the top-right (*subjectivity*). With 68% of items on the left-hand side, *entailment* is the least, and *anaphora*, with 56% in the top-right, the most *straightforward* task.

Anaphora resolution seems to be the most *straightforward* task, with most items in the upper-right quadrant, while sentiment analysis and entailment are the most ambiguous/difficult, both having almost 50% of examples fall in the bottom left quadrant. As expected, the lowest number of items fall in the bottom right section of the plot.

Rationalisation In an attempt to validate the reliability-stability matrix and to test H1.1 and H1.2, rationalisation labels were applied by two postdoctoral researchers with backgrounds in NLP and computational linguistics. They were asked to read the annotation instructions and items and provide each example with a label: *subjective*, *ambiguous*, or *straightforward*. Disagreements were resolved by discussion between these and a third author. Inter-annotator agreement (before resolution) is shown in Table 5, indicating that this in itself was a very difficult task to reach agreement on.

Offence	Sentiment	Entailment	Anaphora
0.26	0.11	0.47	0.02

Table 5: Inter-annotator agreement on the rationalisation labelling task, measured with Cohen's *kappa*.

To quantitatively examine the relationship between the perceived reason for agreement/disagreement and the reliability and stability measurements, we applied a multivariate Kruskal-Wallis test to the independent categorical variable *rationale* (*straightforward*, *subjective*, and *ambiguous*) and the two dependent continuous variables *inter-* and *intra-annotator agreement*.

The test showed that there is only a very small and non-significant difference in the dependent vectors between the different groups, with an H-statistic of $2.734, p = 0.26$, indicating that the assigned rationale labels do not explain the inter- and intra annotator agreement rates.

5 Discussion and conclusion

We have examined the role and use of intra-annotator agreement measures in NLP research. Calculation of such measures can act as an important quality control and could potentially provide insights into the reasons for disagreements between annotators. However, in a systematic review, we found that they are rarely reported in this field.

We have proposed a framework for the interpretation of inter- and intra-annotator agreement, the *reliabilty-stability agreement matrix*. Exploratory annotation experiments failed to validate our theory that this framework can be used to tease apart subjectivity and ambiguity, and it proved to be very hard to recognise or agree on these, even for trained annotators. However, we have shown how comparing both *inter-* and *intra-* annotator agreement enables quantification of the difficulty of particular tasks and/or annotation items. Strikingly, we found that, across four different tasks, crowdsourced annotators were consistently inconsistent, calling into question the implicit assumption that labels provided by individual annotators are stable, and reinforcing the need to collect within-annotator labels for NLP tasks, including those typically considered to be ‘objective’.

Limitations

We acknowledge that the scope of our exploratory experiments is quite small at 50 items per task and 16 annotators, and that larger studies may produce different results. While we took some measures to ensure the quality of recruited annotators (section 3), there are known issues with crowdworker quality for annotation (e.g. Hovy et al., 2013; Weber-Genzel et al., 2024), and some annotator inconsistency may due to inattention—another factor that should be considered and further reason to measure and report intra-annotator agreement.

Ethical considerations

Because we recruit humans to work on data labelling, we obtained approval to undertake this study from the Institutional Review Board (IRB) of the School of Mathematics & Computer Science at Heriot-Watt University, reference 2023-4926-7368. Additionally, we took the following measures:

Compensation We paid the annotators above the Living Wage in our jurisdiction (higher than the legal minimum wage, as recommended (as a minimum) by Shmueli et al. (2021).

Welfare As some of the data to be labelled included offensive language, we:

- avoided recruiting members of vulnerable groups by restricting annotators to those aged over 18, provided them with comprehensive warnings prior to consenting to participate, and asked them to self-declare that they would not be adversely affected by participating;
- allowed annotators to leave the study at any time and informed them that they would be paid for their time regardless;
- kept the annotation task short to avoid lengthy exposure to material which may exceed ‘*minimal risk*’ (Shmueli et al., 2021).

Privacy All personal data of recruited annotators was collected anonymously.

Acknowledgements

We would like to thank the reviewers for their insightful comments which we have tried to incorporate into this version of the paper.

Gavin Abercrombie and Tanvi Dinkar were supported by the EPSRC project ‘Equally Safe Online’ (EP/W025493/1). Dirk Hovy was supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR). He is a member of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis (BIDSA).

References

Gavin Abercrombie, Dirk Hovy, and Vinodkumar Prabhakaran. 2023. Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 96–103, Toronto, Canada. Association for Computational Linguistics.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Abdulrahman Alosaimy and Eric Atwell. 2018. Web-based annotation tool for inflectional language resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

Ron Artstein. 2017. *Inter-annotator Agreement*, pages 297–313. Springer Netherlands, Dordrecht.

Robert H. Ashton. 2000. A review and analysis of research on the test–retest reliability of professional judgment. *Journal of Behavioral Decision Making*, 13(3):277–294.

Jason Baldridge and Alexis Palmer. 2009. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305, Singapore. Association for Computational Linguistics.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021a. Toward a perspectivist turn in ground truthing for predictive computing. In *Conference of the Italian Chapter of the Association for Intelligent Systems (ItAIS 2021)*.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021b. 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.

Anja Belz and Eric Kow. 2011. Discrete vs. continuous rating scales for language evaluation in NLP. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 230–235, Portland, Oregon, USA. Association for Computational Linguistics.

Anja Belz, Adrian Muscat, Pierre Anguill, Mouhamadou Sow, Gaétan Vincent, and Yassine Zinessabah. 2018. SpatialVOC2K: A multilingual dataset of images with annotations and features for spatial relations between objects. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 140–145, Tilburg University, The Netherlands. Association for Computational Linguistics.

Anja Belz, Adrian Muscat, Brandon Birmingham, Jessie Levacher, Julie Pain, and Adam Quinquenel. 2016. Effect of data annotation, feature selection and model choice on spatial description generation in French. In *Proceedings of the 9th International Natural Language Generation conference*, pages 237–241, Edinburgh, UK. Association for Computational Linguistics.

Kim Bennell, Richard Talbot, Henry Wajswelner, Wassana Techovanich, David Kelly, and AJ Hall. 1998. Intra-rater and inter-rater reliability of a weight-bearing lunge measure of ankle dorsiflexion. *Australian Journal of Physiotherapy*, 44(3):175–180.

Luisa Bentivogli, Marcello Federico, Giovanni Moretti, and Michael Paul. 2011. Getting expert quality from the crowd for machine translation evaluation. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Jan Berka, Martin Černý, and Ondřej Bojar. 2011. Quiz-based evaluation of machine translation. *The Prague Bulletin of Mathematical Linguistics*, 95:77.

Su Lin Blodgett. 2021. *Sociolinguistically Driven Approaches for Just Natural Language Processing*. Ph.D. thesis, University of Massachusetts Amherst.

Mayla Boguslav and Kevin Bretonnel Cohen. 2017. Inter-annotator agreement and the upper limit on machine performance: Evidence from biomedical natural language processing. *Studies in health technology and informatics*, 245:298–302.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Houda Bouamor, Hanan Alshikhabobakr, Behrang Mohit, and Kemal Oflazer. 2014. A human judgement corpus and a metric for Arabic MT evaluation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 207–213, Doha, Qatar. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics-MATR*, pages 17–53, Uppsala, Sweden. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland. Association for Computational Linguistics.

Yang Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

Gilson Capilouto, Heather Harris Wright, and Stacy A. Wagovich. 2005. CIU and main event analyses of the structured discourse of older and younger adults. *Journal of Communication Disorders*, 38(6):431–444.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

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.

Leen Cleuren, Jacques Duchateau, Pol Ghesquière, and Hugo Van hamme. 2008. Children's oral reading corpus (CHOREC): Description and assessment of annotator agreement. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Awantee Deshpande, Dana Ruiter, Marius Mosbach, and Dietrich Klakow. 2022. StereoKG: Data-driven knowledge graph construction for cultural knowledge and stereotypes. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 67–78, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Ryan Downey, David Rubin, Jian Cheng, and Jared Bernstein. 2011. Performance of automated scoring for children's oral reading. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 46–55, Portland, Oregon. Association for Computational Linguistics.

Jennifer D'Souza, Sören Auer, and Ted Pedersen. 2021. SemEval-2021 task 11: NLPContributionGraph - structuring scholarly NLP contributions for a research knowledge graph. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 364–376, Online. Association for Computational Linguistics.

Annemarie Friedrich and Alexis Palmer. 2014. Situation entity annotation. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 149–158, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop* *and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Stein Grimstad and Magne Jørgensen. 2007. Inconsistency of expert judgment-based estimates of software development effort. *Journal of Systems and Software*, 80(11):1770–1777.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. 2015. Human evaluation of grammatical error correction systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470, Lisbon, Portugal. Association for Computational Linguistics.

Olivier Hamon. 2010. Is my judge a good one? In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Yifan He, Yanjun Ma, Johann Roturier, Andy Way, and Josef van Genabith. 2010. Improving the post-editing experience using translation recommendation: A user study. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Simon Hengchen and Nina Tahmasebi. 2021. SuperSim: a test set for word similarity and relatedness in Swedish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 268–275, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Aurelie Herbelot and Ann Copestake. 2010. Annotating underquantification. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 73–81, Uppsala, Sweden. Association for Computational Linguistics.

Limor Hochberg, Cecilia Ovesdotter Alm, Esa M. Rantanen, Caroline M. DeLong, and Anne Haake. 2014a. Decision style in a clinical reasoning corpus. In *Proceedings of BioNLP 2014*, pages 83–87, Baltimore, Maryland. Association for Computational Linguistics.

Limor Hochberg, Cecilia Ovesdotter Alm, Esa M. Rantanen, Qi Yu, Caroline M. DeLong, and Anne Haake. 2014b. Towards automatic annotation of clinical decision-making style. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 129–138, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Robert T. Hodgson. 2008. An examination of judge reliability at a major U.S. wine competition. *Journal of Wine Economics*, 3(2):105–113.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Nancy Ide and James Pustejovsky. 2017. *Handbook of linguistic annotation*, volume 1. Springer.

Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt. 2005. A corpus for studying addressing behavior in multi-party dialogues. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, pages 107–116, Lisbon, Portugal. Special Interest Group on Discourse and Dialogue (SIGdial).

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: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Julia Kreutzer, Nathaniel Berger, and Stefan Riezler. 2020. Correct me if you can: Learning from error corrections and markings. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 135–144, Lisboa, Portugal. European Association for Machine Translation.

Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. 2018. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Melbourne, Australia. Association for Computational Linguistics.

Ivana Kruijff-Korbayová, Klára Chvátalová, and Oana Postolache. 2006. Annotation guidelines for Czech-English word alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Harm Lameris and Sara Stymne. 2021. Whit's the richt pairt o speech: PoS tagging for Scots. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 39–48, Kiyv, Ukraine. Association for Computational Linguistics.

Samuel Läubli, Mark Fishel, Manuela Weibel, and Martin Volk. 2013. Statistical machine translation for automobile marketing texts. In *Proceedings of Machine Translation Summit XIV: Posters*, Nice, France.

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.

Xuansong Li, Niyu Ge, Stephen Grimes, Stephanie M. Strassel, and Kazuaki Maeda. 2010. Enriching word alignment with linguistic tags. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Wanqiu Long, Xinyi Cai, James Reid, Bonnie Webber, and Deyi Xiong. 2020. Shallow discourse annotation for Chinese TED talks. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1025–1032, Marseille, France. European Language Resources Association.

Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Rui Li, Jeff B. Pelz, Pengcheng Shi, and Anne Haake. 2012. Annotation schemes to encode domain knowledge in medical narratives. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 95–103, Jeju, Republic of Korea. Association for Computational Linguistics.

Chris McGreal. 2021. Derek Chauvin found guilty of George Floyd's murder. *The Guardian*.

Ana-Belén Meseguer-Henarejos, Julio Sánchez-Meca, José-Antonio López-Pina, and Ricardo Carles-Hernández. 2018. Inter- and intra-rater reliability of the Modified Ashworth Scale: a systematic review and meta-analysis. *European journal of physical and rehabilitation medicine*, 54(4):576—590.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and the PRISMA Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *Annals of Internal Medicine*, 151(4):264–269.

Barbara Plank. 2022. The 'problem' of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of EMNLP*. ACL.

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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789, Minneapolis, Minnesota. Association for Computational Linguistics.

Miranda Rose and Jacinta Douglas. 2003. Limb apraxia, pantomine, and lexical gesture in aphasic speakers: Preliminary findings. *Aphasiology*, 17(5):453–464.

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.

Dana Ruiter, Thomas Kleinbauer, Cristina España-Bonet, Josef van Genabith, and Dietrich Klakow. 2022. Exploiting social media content for self-supervised style transfer. In *Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media*, pages 11–34, Seattle, Washington. Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, and Juan Antonio Pérez-Ortiz. 2012. Source-language dictionaries help non-expert users to enlarge target-language dictionaries for machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3422–3429, Istanbul, Turkey. European Language Resources Association (ELRA).

Jeremy Sawyer and Anup Gampa. 2018. Implicit and explicit racial attitudes changed during Black Lives Matter. *Personality and Social Psychology Bulletin*, 44.

Claudia Schulz, Christian M. Meyer, Jan Kiesewetter, Michael Sailer, Elisabeth Bauer, Martin R. Fischer, Frank Fischer, and Iryna Gurevych. 2019. Analysis of automatic annotation suggestions for hard discourse-level tasks in expert domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2761–2772, Florence, Italy. Association for Computational Linguistics.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.

Hanna Szekeres, Eric Shuman, and Tamar Saguy. 2020. Views of sexual assault following #MeToo: The role of gender and individual differences. *Personality and Individual Differences*, 166:110203.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway. Association for Computational Linguistics.

Frederik Vaassen and Walter Daelemans. 2011. Automatic emotion classification for interpersonal communication. In *Proceedings of the 2nd Workshop on*

Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011), pages 104–110, Portland, Oregon. Association for Computational Linguistics.

Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. 2021. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 613–623, Online. Association for Computational Linguistics.

Mihaela Vela and Josef van Genabith. 2015. Re-assessing the WMT2013 human evaluation with professional translators trainees. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 161–168, Antalya, Turkey.

Abigail Walsh, Teresa Lynn, and Jennifer Foster. 2020. Annotating verbal MWEs in Irish for the PARSEME shared task 1.2. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 58–65, online. Association for Computational Linguistics.

Chaojun Wang, Christian Hardmeier, and Rico Sennrich. 2021. Exploring the importance of source text in automatic post-editing for context-aware machine translation. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 326–335, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Lu Wang, Hema Raghavan, Claire Cardie, and Vittorio Castelli. 2014. Query-focused opinion summarization for user-generated content. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1660–1669, Dublin, Ireland. Dublin City University and 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.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Deniz Zeyrek, Amália Mendes, and Murathan Kurfalı. 2018. Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A Systematic review methodology

For this review, we followed the established systematic review guidelines of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement (Moher et al., 2009), as recommended by van Miltenburg et al. (2021):

1. Develop search query terms
2. Conduct search
3. Apply inclusion/exclusion criteria
4. Code included publications
5. Measure inter- and intra-annotator agreement (re-code subset of publications)
6. Synthesise results

The review covers all results retrieved from the Anthology's search facility. The searches were conducted on September 14 2022. Following retrieval of the resulting publications, we applied the inclusion/exclusion criteria shown in Table 6.

Include	**Exclude**
Human annotation studies	No human annotation study is conducted (e.g. surveys/reviews of other work)
Repeated annotations are collected	Repeated annotations are not collected
Intra-annotator measurement is reported	Intra-annotator measurement not reported
Measurement conducted on manual labels applied by human annotators	Labelling is performed automatically
'Intra-' refers to repeat annotations of the same items by the same annotator	Term 'intra-' is used, but refers to agreement measurements between different items and/or annotators
Publication is a full paper	Posters, proceedings, proposals, technical system descriptions etc.

Table 6: Criteria for in/exclusion in/from the review.

The searches returned 138 publications. After removing duplicates, and applying the inclusion criteria we were left with 56 relevant publications in the Anthology.

Publication	NLP sub-field	Publication	NLP sub-field
Akhbardeh et al. (2021)	Machine Translation	Graham et al. (2013)	Machine Translation
Alosaimy and Atwell (2018)	Syntax	Grundkiewicz et al. (2015)	Syntax
Baldridge and Palmer (2009)	Machine Translation	Hamon (2010)	Machine Translation
Belz and Kow (2011)	NLG	He et al. (2010)	Machine Translation
Belz et al. (2016)	NLG	Hengchen and Tahmasebi (2021)	Semantics
Belz et al. (2018)	NLG	Herbelot and Copestake (2010)	Semantics
Bentivogli et al. (2011)	Machine Translation	Hochberg et al. (2014a)	Cognitive psychology
Berka et al. (2011)	Machine Translation	Hochberg et al. (2014b)	Cognitive psychology
Bojar et al. (2013)	Machine Translation	Jovanovic et al. (2005)	NLG
Bojar et al. (2014)	Machine Translation	Kiritchenko and Mohammad (2017)	Affective computing
Bojar et al. (2015)	Machine Translation	Kreutzer et al. (2020)	Machine Translation
Bojar et al. (2016)	Machine Translation	Kreutzer et al. (2018)	Machine Translation
Bojar et al. (2017)	Machine Translation	Kruijff-Korbayová et al. (2006)	Machine Translationn
Bojar et al. (2018)	Machine Translation	Lameris and Stymne (2021)	Syntax
Bouamor et al. (2014)	Machine Translation	Läubli et al. (2013)	Machine Translation
Callison-Burch et al. (2007)	Machine Translation	Li et al. (2010)	Machine Translation
Callison-Burch et al. (2010)	Machine Translation	Long et al. (2020)	Semantics
Callison-Burch et al. (2011)	Machine Translation	McCoy et al. (2012)	Cognitive psychology
Callison-Burch et al. (2012)	Machine Translation	Ruiter et al. (2022)	NLG
Callison-Burch et al. (2009)	Machine Translation	Sánchez-Cartagena et al. (2012)	Machine Translation
Callison-Burch et al. (2008)	Machine Translation	Schulz et al. (2019)	Semantics
Cao et al. (2022)	Semantics	Vaassen and Daelemans (2011)	Affective computing
Cercas Curry et al. (2021)	Abuse detection	Vela and van Genabith (2015)	Machine Translation
Cleuren et al. (2008)	Automatic text grading	Walsh et al. (2020)	Syntax
D'Souza et al. (2021)	Sematics	Wang and Sennrich (2020)	Machine Translation
Deshpande et al. (2022)	Semantics	Wang et al. (2021)	Machine Translation
Downey et al. (2011)	Automatic text grading	Wang et al. (2014)	NLG
Friedrich and Palmer (2014)	Semantics	Zeyrek et al. (2018)	Semantics

Table 7: Publications in the ACL Anthology in which intra-annotator agreement is reported.

B Included papers

A list of included publications from the ACL Anthology that report intra-annotator agreement is presented in Table 7.

C Cohen's kappa scores

	Reliability (*Inter-*) κ		**Stability** (*Intra-*) κ	
	μ	σ	μ	σ
Offence	0.05	0.28	0.27	0.28
Sentiment	0.02	0.25	0.17	0.29
Entailment	0.02	0.25	0.28	0.28
Anaphora	0.07	0.31	0.22	0.35
Overall	0.04	0.28	0.23	0.30

Table 8: Pairwise reliability and stability of the collected labels measured with mean (μ) and standard deviations (σ) across items for inter- and intra-annotator agreement scores measured with Cohen's kappa (κ).

Revisiting Active Learning under (Human) Label Variation

Cornelia Gruber[*1] **Helen Alber**[*1,2] **Bernd Bischl**[1,2]
Göran Kauermann[1,2] **Barbara Plank**[2,3] **Matthias Aßenmacher**[1,2]
[1] LMU Munich, Department of Statistics, Germany
[2] Munich Center for Machine Learning (MCML), Germany
[3] LMU Munich, Center for Information and Language Processing (CIS), Germany
[*]Equal contribution **Correspondence:** cornelia.gruber@lmu.de, helen.alber@stat.uni-muenchen.de

Abstract

Access to high-quality labeled data remains a limiting factor in applied supervised learning. Active learning (AL), a popular approach to optimizing the use of limited annotation budgets in training ML models, often relies on at least one of several simplifying assumptions, which rarely hold in practice when acknowledging human label variation (HLV). Label variation (LV), i.e., differing labels for the same instance, is common, especially in natural language processing. Yet annotation frameworks often still rest on the assumption of a single ground truth, overlooking HLV, i.e., the occurrence of plausible differences in annotations, as an informative signal. In this paper, we examine foundational assumptions about truth and label nature, highlighting the need to decompose observed LV into signal (e.g., HLV) and noise (e.g., annotation error). We survey how the AL and (H)LV communities have addressed—or neglected—these distinctions and propose a conceptual framework for incorporating HLV throughout the AL loop, including instance selection, annotator choice, and label representation. We further discuss the integration of large language models (LLM) as annotators. Our work aims to lay a conceptual foundation for (H)LV-aware active learning, better reflecting the complexities of real-world annotation.

1 Introduction

Prediction algorithms play a central role in many natural language processing (NLP) tasks, like hate speech detection (Basile, 2020), sentiment analysis (Kenyon-Dean et al., 2018), or natural language inference (NLI; Pavlick and Kwiatkowski, 2019). For training such supervised machine learning (ML) models, a notable amount of labeled training data is necessary. However, acquiring high-quality labels is expensive as human crowd workers or, even more expensive, domain experts need to annotate the data. A popular scheme to efficiently guide the annotation process and allocate annotation budgets is active learning (AL; Abney, 2007; Settles, 2009). AL aims to maximize the expected predictive performance of the resulting model while minimizing the required number of annotations; often done by iterating the following three steps: (1) Training the ML model on available labeled data. (2) Selecting new instances for labeling from a pool of unlabeled data, usually based on an acquisition function. (3) Labeling these with an oracle. Those steps, which are repeated until the available annotation budget is depleted or the model has reached its target accuracy, rest on the following assumptions:

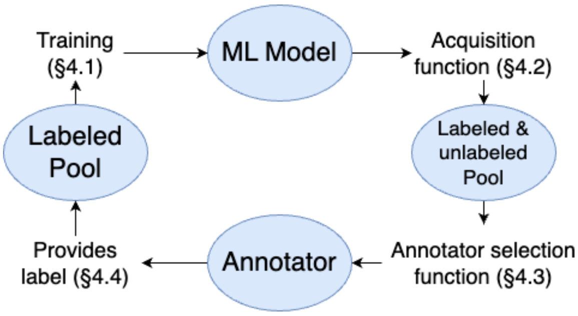


Figure 1: The traditional AL loop with possible adaptations in different steps, leading to generalized label variation aware AL

A1 There exists a single ground truth label per instance.

A2 The oracle provides the ground truth labels without any noise.

A3 The annotation difficulty or cost is equal for all instances.

Equal annotation cost is not, strictly speaking, a critical assumption for AL, but is becoming increasingly important to consider. However, in NLP, those assumptions often are not or cannot be fulfilled. Especially in the presence of human label variation (HLV), i.e., differences in human anno-

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 75–86
November 8, 2025 ©2025 Association for Computational Linguistics

tations that are plausible variability due to subjectivity or ambiguity and explicitly no sign of error (Plank, 2022; cf. §3), even the existence of such an omniscient oracle is questionable.

When we move away from these assumptions and acknowledge HLV, the AL loop is extended: an annotator selection function is introduced to choose among multiple annotators with varying perspectives or expertise, rather than assuming a single infallible oracle (cf. Figure 1).

Contributions In this work, we examine the consequences for the AL cycle when its conventional assumptions, i.e., A1 – A3, are violated due to plausible variation in labels, often coined HLV. We begin by discussing foundational assumptions about truth in annotation (§2), laying out different perspectives on label nature and emphasizing the need for a signal-noise decomposition of label variation (LV) into plausible variation (e.g., HLV) and noise. In what follows, we provide an overarching survey of the literature of both fields, i.e., (H)LV and AL, that reveals an emerging line of research integrating aspects of (H)LV into AL (§3), but simultaneously also uncovers shortcomings and misunderstandings between the fields. We then identify and categorize the adaptations required in the AL loop (§4), including modifications to the annotator selection function and considerations for incorporating LLMs. Altogether, we offer a holistic perspective on AL in the presence of (H)LV, aiming to establish a more structured ground for discussion and future empirical investigation by bridging ongoing debates across NLP, empirical ML, statistics, and philosophy.

2 Assumptions about Truth in Annotation

When observing LV in human annotations, it is important to recognize that this variation may arise from both error and HLV (Weber-Genzel et al., 2024), which can be present simultaneously. Throughout this work, we use LV to refer to the observed differences in annotation, which can be decomposed into signal, such as HLV, and noise, such as actual annotation error. Reflecting on the underlying assumptions about the true labels is crucial, as it helps to distinguish between these sources of LV, or, in other words, aids the "interpretation of any observed annotator disagreement" (Röttger et al., 2022, p. 3).

Task Dependence and Subjectivity The extent to which observed LV is attributed to HLV is often judged based on the (assumed) subjectivity of the task (Basile et al., 2021). In domains such as specific image classification tasks in computer vision (e.g., distinguishing between images of cats or dogs), lower levels of HLV may be expected, as the real-world categories constituting the data-generating process (e.g., actual cats or dogs) are typically less subjective. In such cases, higher shares of the observed label variation may be attributable to various types of errors, such as issues arising from imprecise measurement, the compression of real-world information into data, or noise, e.g., introduced during data collection like blurriness in images (Gruber et al., 2025), rather than to HLV.

The notion of inferring task subjectivity from observed LV introduces a certain circularity: LV is intuitively taken as evidence of subjectivity, while assumptions about subjectivity, in turn, inform how much of the variation is attributed to HLV. A more thorough discussion and a systematic approach to operationalizing subjectivity appear essential for future work when aiming to disentangle signal and noise in observed LV.

Worldviews and Nature of Truth Many NLP tasks, as well as certain computer vision tasks (e.g., image segmentation in medicine; Zhang et al., 2020), are assumed to involve a higher degree of subjectivity. Particularly when addressing such tasks, different underlying philosophical assumptions on the nature of truth and the closely related nature of reality can lead to varying methodological implications. For example, adopting a monistic worldview—drawing on the discussion of monism by Russell (1907)—may involve the assumption of a single underlying reality, with different annotations merely being different perspectives on it. In this context, no observed annotation could be fully true or false, and taking individual annotations into account as a distribution on the instance level may be a reasonable approach.

Label Non-Determinism and Levels Whether label variation is viewed from the annotator's perspective (annotator level) or the instance's perspective (instance level) can help clarify certain complexities. For example, on the annotator level, label non-determinism, defined as a probabilistic mapping between a real-world instance and a set of labels, can vary in degree between both subjective and less subjective cases, and may even include label-deterministic subjective settings. In contrast, on the instance level, greater subjectivity inherently results in more label non-determinism. Ambiguity, here clearly distinguished from subjectivity, is linked to higher label non-determinism at both levels. While factors like these—label non-determinism, subjectivity, ambiguity, and annotator level vs. instance level—can, in principle, be treated separately, we assume substantial dependencies between them. For instance, even at the annotator level, tasks that are assumed to be more subjective may be likely prone to exhibiting a greater degree of label non-determinism.

Types of Label Nature Approaching the discussion from a more applied perspective, we provide an overview of possible types of labels: (a) discrete class label (also known as "hard label"), (b) label as probability for discrete classes (sometimes referred to as "soft label", Uma et al., 2021, or "human judgment distribution"), and (c) label as continuous distribution for underlying fixed number of classes (cf. Figure 2). Note, that while the illustration depicts only scenarios with $k \in \{2, 3\}$ classes for simplicity, this schema is generally also applicable to settings with $k > 3$ classes. When viewing the annotation process from a statistical perspective, i.e. making assumptions about the data generating process, each label y_i can be regarded as a realization of a random variable Y. For discrete labels (a), an example in the binary setting is $y_i = 1$, with $Y \sim \text{Bin}(1, p)$; in the ternary case, i.e., three classes, an example is $y_i = [1, 0, 0]$, with $Y \sim \text{Multinom}(1, \boldsymbol{p})$, $\boldsymbol{p} = (p_A, p_B, p_C)$. Moving to probability labels (b), the label itself represents a probabilistic belief over class membership. For instance, $y_i = 0.75$ may arise from $Y \sim \text{Beta}(\alpha, \beta)$, and $y_i = [0.6, 0.2, 0.2]$ may be a realization from $Y \sim \text{Dir}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} = (\alpha_A, \alpha_B, \alpha_C)$. Finally, in the case of distribution labels (c), y_i takes the form of a full probability distribution—for example, $y_i = \text{Beta}(8, 3.5)$ in the binary case or $y_i = \text{Dir}(8, 3, 4)$ in the ternary case. Here, the label y_i is itself a distribution over class probabilities. The distribution of Y is modeled hierarchically by placing priors on the parameters of this distribution, e.g., on α, β in the Beta case or on $\boldsymbol{\alpha}$ in the Dirichlet case (Hechinger et al., 2024a).

We here challenge the common assumption of the first type (discrete class labels, sometimes also referred to as "single ground truths") by proposing the consideration of the latter two types, both as assumed true labels and requested annotations. The third label type appears to be the least studied of the ones listed; however, some work in uncertainty quantification has begun to explore different label representations (Bengs et al., 2022; Hechinger et al., 2024a; Sale et al., 2024; Wimmer et al., 2023).

In practice, a discrepancy can occur between the type of label assigned by the annotator and the assumed nature of the true label. This mismatch is especially likely when true labels are assumed to be continuous distributions over classes (cf. case (c) in Figure 2), as human annotators are not inherently equipped to give non-discrete annotations (cf. §4.4 for further discussion of the "oracle" in the AL cycle). This discrepancy introduces an irreducible uncertainty and may result in the interpretation that the observed label variation does not necessarily equate to HLV. This again emphasizes the importance of distinguishing between assumptions about the true labels and assumptions that may be required for practical reasons during annotation and the AL loop.

3 Views on Label Variation and Active Learning

In what follows, we outline key stages in how different fields have approached label variation, a phenomenon discussed under various terminologies and theoretical perspectives, and illustrate them with literature examples. Starting from work documenting its occurrence across diverse tasks, we move from approaches that neglect or seek to mitigate LV, to studies that measure variation mainly to steer away from high LV instances. Subsequently, we summarize recent perspectives that embrace (H)LV as a valuable signal, integrating it into learning objectives through distributional labels and motivating its decomposition. We then examine how the field of AL has responded to, incorporated, or overlooked these diverse understandings of LV in its methodological developments.

3.1 Label Variation

Supervised ML depends fundamentally on annotated data, making the quality and nature of labels a central part of the learning process. The phenomenon of LV, i.e., the occurrence of differing annotations for the same instance, both between and within annotators, is not limited to subjective tasks

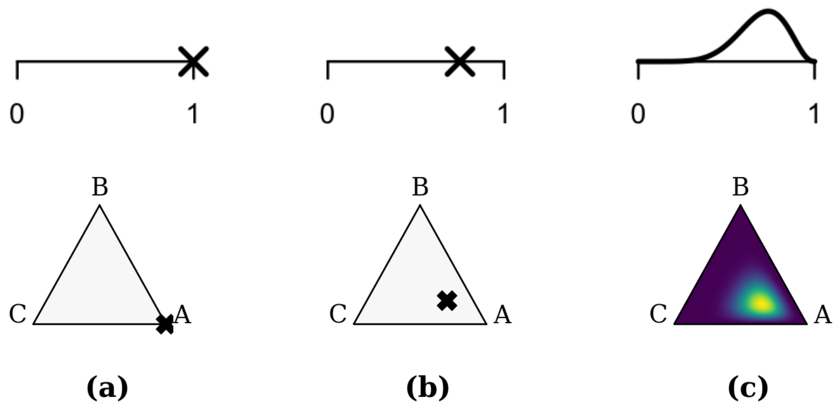


Figure 2: Types of labels visualized. Each label y_i is a realization of a random variable Y.
Top row: binary classes; bottom row: three classes.
(a) Discrete label: $y_i = 1$ with $Y \sim \text{Bin}(1, p)$ (top) and $y_i = [1, 0, 0]$ with $Y \sim \text{Multinom}(1, \boldsymbol{p})$ (bottom),
(b) probability label: $y_i = 0.75$ with $Y \sim \text{Beta}(\alpha, \beta)$ (top) and $y_i = [0.6, 0.2, 0.2]$ with $Y \sim \text{Dir}(\boldsymbol{\alpha})$ (bottom),
(c) distributional label: $y_i = \text{Beta}(8, 3.5)$ (top) and $y_i = \text{Dir}(8, 3, 4)$ (bottom) with a hierarchical model for the distribution of Y with priors on the parameters of the respective distributions. In the bottom row, ternary plots visualize the relative proportions of three classes as positions within a triangle. Each cross represents a single label, with its location indicating the class composition: the closer a point is to a corner, the higher the class proportion.

but has been found across a wide range of applications. In NLP, this includes tasks such as sentiment analysis (Kenyon-Dean et al., 2018), hate speech detection (Basile, 2020), veridicality judgments (De Marneffe et al., 2012), argumentation mining (Trautmann et al., 2020), natural language inference (Pavlick and Kwiatkowski, 2019), and even tasks traditionally considered “objective” like part-of-speech tagging (Plank et al., 2014b), word sense disambiguation (Passonneau et al., 2012), semantic role labeling (Dumitrache et al., 2019), and named entity recognition (Inel and Aroyo, 2017). Similar variation has also been observed in computer vision tasks like medical image classification and object identification (Uma et al., 2021), or remote sensing (Hechinger et al., 2024b), where annotator disagreement arises from ambiguity and subjectivity in visual interpretation. While most existing works listed treat this as *either* signal or noise, we refrain from exclusively assigning observed label variation to either category in the first place.

Mitigating Label Variation The assumption of a single ground truth label has long dominated ML practice, as reflected in foundational ML literature (Goodfellow et al., 2016; Hastie et al., 2009; Mitchell, 1997). Within this framework, LV is typically regarded as erroneous and to be minimized or corrected (Alm, 2011; Aroyo and Welty, 2015) with Cabitza et al. (2023), for example, documenting widespread practices of “disagreement removal”.

Treating “Hard” Cases Moving beyond the traditional view of LV, early work has begun to explore LV as a potential source of information. Exemplary, Reidsma and Op Den Akker (2008) advocate for analyzing patterns of disagreement, providing an overview of the various factors that may underlie annotator disagreement. However, this line of work uses information from LV to steer ML models away from “hard” cases (i.e., items with high LV), by, e.g., enabling classifiers to abstain from making predictions. Plank et al. (2014a) propose incorporating inter-annotator agreement measures into a cost-sensitive loss function, thereby explicitly integrating LV into the learning process as a signal of uncertainty. The next paragraph discusses approaches seeking to embrace LV more directly by explicitly modeling it, for instance, through adjustments to the nature and interpretation of the labels.

Human Label Variation There are two main bodies of literature relevant to this work addressing differences in human annotations: one that predominantly uses the term *variation* and another that refers to *disagreement*. We adopt the terminology of Plank (2022), who introduced the notion of HLV to conceptualize such differences as plausible and meaningful variations rather than as annotation errors. This perspective has been particularly motivated by developments in NLP, where subjectivity, leading to HLV, is recognized as an inherent property of many language-related tasks (Alm, 2011).

This framework further aligns with the concept of *perspectivism* introduced by Cabitza et al. (2023), which emphasizes that, rather than seeking a single ground truth, collecting multiple labels offers a way to sample the range of perceptions, opinions, and judgments present in a population.

The related body of literature that adopts the term disagreement rather than variation is more heterogeneous in its interpretation and evaluation of annotation differences. While some contributions view such disagreement as plausible or informative (Uma et al., 2021), others primarily treat it as a source of noise or error (Beigman Klebanov and Beigman, 2009). Throughout this section, we review work from both terminological traditions.

Distributional Labels Several contributions have moved beyond discrete labels by aggregating multiple annotations into distributional labels (cf. Figure 2 for the different types of label nature), aligning with a (strong) perspectivist stance. De Marneffe et al. (2012) frame veridicality assessment as a "distribution-prediction task", using judgments from 10 annotators per instance. Similarly, Aroyo and Welty (2015) view disagreement as a signal and introduce the "Crowd Truth" framework, which incorporates distributional labels through annotation aggregation and addresses factors like the design of annotation guidelines and differing annotator expertise. In computer vision, Peterson et al. (2019) show that training convolutional neural networks on soft labels derived from multiple annotators improves generalization under distributional shift. For NLI, Pavlick and Kwiatkowski (2019) use slider-based annotations to capture uncertainty and argue for models that predict distributions over judgments. More recently, Chen et al. (2022) and Gruber et al. (2024) investigated whether to prioritize more annotators per instance or more annotated instances when working on the distributional level via label aggregation.

However, these contributions primarily address HLV by aggregating multiple annotations per instance, thereby treating distributional labels as post-hoc constructions rather than as *distributional by nature*—as in case (c) above—i.e., labels deliberately designed from the outset to capture uncertainty directly as a characteristic of the label.

Decomposing Label Variation Furthermore, the above contributions tend to conflate LV with HLV, overlooking the simultaneous presence of *both* noise and signal within LV. Incorporating this conceptual distinction, Palomaki et al. (2018) highlight the need to distinguish between actual annotation errors and "disagreement that falls within the acceptable range", introducing the concept of *acceptable variation*, which may differ across subsets of instances and has direct implications for task design. Weber-Genzel et al. (2024) extend this conceptual distinction to NLI. They address the challenge of identifying annotation error by incorporating validated annotator *labels with explanations* through a second round of validity judgments, rather than relying on post-hoc interpretation alone. This builds on earlier work by Jiang et al. (2023), who identify the phenomenon of *within-label variation*, where, even when the same label is assigned, annotators may vary in their explanations.

Data annotation remains a labor-intensive and complex process, particularly when aiming to analyze or leverage a signal–noise decomposition of LV. The following section, therefore, turns to the field of active learning, which focuses on strategies for optimizing annotation budgets and minimizing annotation effort.

3.2 Active Learning

Active Learning has been a vivid field of research for over 30 years (Aggarwal et al., 2014; Lewis and Catlett, 1994; Settles, 2009; Seung et al., 1992). Settles (2011) already discussed practical issues arising in active learning, including querying in batches, noisy oracles, and variable labeling costs. Zhang et al. (2022) provide a survey on AL for NLP, while Rauch et al. (2023) propose a tailored NLP benchmark for AL.

Annotation Costs and Quality The true costs of annotation are explored in Krishnamurthy et al. (2019); Margineantu (2005); Settles et al. (2008); Tomanek and Hahn (2010); Xie et al. (2018), challenging assumption **A3** (*"The annotation difficulty or cost is equal for all instances."*) by modeling variation in annotation effort. Gao and Saar-Tsechansky (2020); Donmez et al. (2009) extend this by accounting for annotators with varying accuracies, while Donmez and Carbonell (2008) acknowledge that even oracles might be incorrect depending on task difficulty, both relaxing **A2** (*"The oracle provides the ground truth labels without any noise."*). Yan et al. (2011, 2012) suggest jointly selecting an instance and an annotator. Furthermore, Zhang and Chaudhuri (2015) and Chakraborty (2020) incorporate both low-cost and expert annotators by assuming a trade-off between cost and label quality. However, these approaches still assume a single ground truth label per instance (reliance on **A1**; *"There exists a single ground truth label per instance."*) and treat label variation as noise.

In contrast, we highlight the underexplored setting where (H)LV is inherent and may carry an informative signal, arguing that its integration into the AL framework requires rethinking core components such as acquisition and annotation strategies.

Relabeling Relabeling, i.e., collecting additional annotations for previously labeled instances to reduce noise or correct errors, is explored in Chen et al. (2022); Goh and Mueller (2023); Lin et al. (2016); Yuan et al. (2024). These approaches implicitly challenge assumptions **A2** and **A3** by acknowledging annotation errors and varying difficulty. However, they treat disagreement as an error rather than a potentially meaningful signal.

HLV-aware AL A few recent studies have begun to explore how AL can be adapted to account for HLV. Wang and Plank (2023) and van der Meer et al. (2024) suggest strategies to choose *which* human annotator should label an instance. Furthermore, Baumler et al. (2023) suggest aligning model uncertainty with annotator uncertainty. While these works offer valuable insights, they address specific assumptions or propose targeted adaptations to the AL process. In §4, we build on these efforts by systematically analyzing their contributions and organizing them into a broader framework. There, we formalize and categorize key adaptations required for making AL effective in the presence of HLV, and point to open challenges and directions for future research.

4 The Active Learning Loop Revisited

In the following, we discuss the consequences of the assumptions about truth in annotation and the nature of the labels (§2) on each of the steps of the AL loop (as visualized in Figure 1).

4.1 Training Measure

Traditional AL assumes a single ground truth label provided by an oracle. This aligns naturally with classic supervised ML, where models are optimized based on hard-label measures like Bernoulli loss or cross-entropy. However, in cases where label variation is not due to error but comes from plausible causes, different soft-label measures are necessary. In such cases, alternative loss measures based on label distributions, such as Kullback-Leibler (KL) divergence (Koller et al., 2024), Jensen-Shannon divergence, or label embeddings (Schweden et al., 2025) have been proposed. Baumler et al. (2023) offer solutions by comparing the predicted and observed label distribution, thus directly optimizing for a trustworthy representation of LV.

> **(C1) Consequence:** In the presence of HLV, distributional measures must be used for optimizing and evaluating the classifier.

4.2 Acquisition Function

The acquisition function ranks all unlabeled instances by their usefulness if they were to be labeled. The oracle then provides labels to the most instructive cases. Traditional AL (Zhang et al., 2022) uses querying strategies based on either *informativeness* or *representativeness*, or *hybrid* approaches (Ash et al., 2020). Informative querying often uses uncertainty sampling (Lewis and Gale, 1994), where the samples with the highest predicted label entropy get labeled first, thus the ones with the highest uncertainty. However, with HLV, high entropy can also be integral to the task, and thus not necessarily a sign of uncertainty. This shows that classic entropy sampling is not suitable for AL in the presence of HLV. Representativeness sampling favors samples that represent the unlabeled pool well. However, classical representativeness sampling ignores the option of labeling some instances multiple times to represent HLV properly and is thus also unsuitable for HLV. Further, defining representativeness in distributions is not trivial. One option to take HLV into account is to precede the AL loop by training a prediction model for annotator disagreement (entropy) and then changing the acquisition function to query samples where the predicted annotator entropy and model entropy diverge the most (Baumler et al., 2023).

> **(C2) Consequence:** In the presence of HLV, classical informativeness or representativeness sampling are unsuitable, as they ignore the option of labeling instances multiple times and fail to process distributional labels.

4.3 Annotator Selection Function

The assumption of having an oracle providing the single ground truth label is not suitable in subjective tasks, where the distribution of human opinions is of interest, or other tasks with high (assumed) HLV. Therefore, an additional step in the AL cycle needs to be considered: the selection of annotators. In many crowd worker settings, it is possible to inquire about labels from a specific annotator. Extending this thought, different "types" of annotators could be queried, e.g., not only human workers but also large language models (LLM). This is also known as "pre-annotation" (Zhang et al., 2022) in the pre-LLM era, and analogously as "LLM-as-annotator" (or "LLM-as-a-judge"; Wu et al., 2024; Zheng et al., 2023) today, where the idea is that a model's predictions are given to human annotators to confirm or adjust. Consequently, an overarching annotator selection strategy needs to evaluate whether a language model or a human shall provide the label, and whether a specific annotator (e.g., representing a minority) or a specific LLM could provide the label. Recent work has extended the AL framework to include not only sample selection but also annotator selection. Wang and Plank (2023) introduce a multi-head model that jointly selects the most informative instance and the most suitable annotator. In contrast, van der Meer et al. (2024) focus on ensuring representativeness and diversity in annotator selection, proposing a strategy that balances labeler perspectives to reflect the underlying population of interpretations better. The idea of using LLMs as annotators is pursued in Bansal and Sharma (2023); Xia et al. (2025); Zhang et al. (2023) .

(C3) Consequence: In the presence of HLV it matters *who* provides the label(s). An annotator acquisition function must decide not only whether to query a human or a language model, but also whether to obtain one or multiple annotations and from *which* specific annotator or model.

4.4 Quality of Label and Uncertainty

The quality of annotators is an important area of research in NLP, which becomes increasingly meaningful when diversity in annotations is present (or required; Sorensen et al., 2024) and label noise cannot be easily separated from the plausible share of label variation. Currently, most work either assumes variation is noise (Goh and Mueller, 2023; Yan et al., 2016; Zhang et al., 2015; Zhao et al., 2011) or all variation in labels represents true HLV (van der Meer et al., 2024; Wang and Plank, 2023). Particularly, when the ground truth label is a distribution and multiple annotators provide labels, detecting annotation noise in HLV is a complex endeavor (Weber-Genzel et al., 2024). Now, when not only different humans annotate the data, but samples can also be processed by LLMs, assessing the label quality is non-trivial either (Ni et al., 2025). Also, in the process of labeling, human annotators usually provide a single label, while an LLM could directly provide distributions (Chen et al., 2024; Pavlovic and Poesio, 2024). This makes LLMs as annotators especially attractive in the presence of HLV and for providing labels for case (c) depicted in Figure 2.

(C4) Consequence: In the presence of HLV, it is non-trivial to distinguish true label variation from noise, especially when labels can be sourced from both humans and language models, each with differing capabilities and output formats.

5 Conclusion

In this work, we provide an overview of the crucial connection between the fields of (human) label variation and active learning. Our comprehensive overview of the existing literature in the individual fields helps building bridges between different, but connected, streamlines of research, paving the way for the identification of critical aspects to consider in the AL loop in the presence of HLV. Our critical assessment of these aspects aims to further point out potential avenues for future research to deal with them in a more nuanced and reflective manner. In doing so, we uncover several crucial assumptions about labels which are often *implicitly* made in traditional AL. However, we argue that they need to be made *explicit*. While providing a unified and implemented solution to the discussed problems is beyond the scope of the paper, we still hope to contribute to ongoing research debates on (H)LV by providing a fresh perspective from a different angle on existing problems and encourage new work addressing label-variation-aware active learning.

Limitations

While this work provides a structured discussion on active learning in the presence of human label variation, several limitations remain. The philosophical discussion on annotation truth is a conceptual suggestion rather than a prescriptive framework. For example, we do not address annotation tasks where it is assumed that *no* ground truth exists, or discuss other frameworks like imprecise probabilities for representing human label variation. Moreover, not all discussed adaptations are implemented in AL pipelines yet, requiring empirical validation. Additionally, we do not explore alternative methods for gathering human annotations that may better accommodate HLV in detail. Lastly, the reliability of "LLM-as-annotator" remains an open question. While LLMs can reduce costs and provide label distributions, their biases and lack of accountability pose challenges.

Acknowledgments

CG is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. HA and MA are funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the National Research Data Infrastructure – NFDI 27/1 - 460037581.

References

Steven Abney. 2007. *Semisupervised Learning for Computational Linguistics*, 1st edition. Chapman & Hall/CRC.

Charu C. Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and S. Yu Philip. 2014. Active learning: A survey. In *Data Classification: Algorithms and Applications*, pages 599–634. Chapman and Hall/CRC.

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.

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.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*.

Parikshit Bansal and Amit Sharma. 2023. Large Language Models as Annotators: Enhancing Generalization of NLP Models at Minimal Cost. *arXiv preprint*. ArXiv:2306.15766 [cs].

Valerio Basile. 2020. It's the End of the Gold Standard as We Know It: Leveraging Non-aggregated Data for Better Evaluation and Explanation of Subjective Tasks. In *AIxIA 2020 – Advances in Artificial Intelligence: XIXth International Conference of the Italian Association for Artificial Intelligence, Virtual Event, November 25–27, 2020, Revised Selected Papers*, pages 441–453, Berlin, Heidelberg. Springer-Verlag.

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.

Connor Baumler, Anna Sotnikova, and Hal Daumé Iii. 2023. Which Examples Should be Multiply Annotated? Active Learning When Annotators May Disagree. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10352–10371, Toronto, Canada. Association for Computational Linguistics.

Beata Beigman Klebanov and Eyal Beigman. 2009. From Annotator Agreement to Noise Models. *Computational Linguistics*, 35(4):495–503.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. 2022. Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Advances in neural information processing systems*.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868. Number: 6.

Shayok Chakraborty. 2020. Asking the Right Questions to the Right Users: Active Learning with Imperfect Oracles. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3365–3372. Number: 04.

Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. 2024. "seeing the big through the small": Can LLMs approximate human judgment distributions on NLI from a few explanations? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14396–14419, Miami, Florida, USA. Association for Computational Linguistics.

Derek Chen, Zhou Yu, and Samuel R. Bowman. 2022. Clean or annotate: How to spend a limited data collection budget. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 152–168, Hybrid. Association for Computational Linguistics.

Marie-Catherine De Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did It Happen? The Pragmatic Complexity of Veridicality Assessment. *Computational Linguistics*, 38(2):301–333.

Pinar Donmez and Jaime G. Carbonell. 2008. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 619–628, Napa Valley California USA. ACM.

Pinar Donmez, Jaime G. Carbonell, and Jeff Schneider. 2009. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 259–268, Paris France. ACM.

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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2164–2170, Minneapolis, Minnesota. Association for Computational Linguistics.

Ruijiang Gao and Maytal Saar-Tsechansky. 2020. Cost-Accuracy Aware Adaptive Labeling for Active Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03):2569–2576. Number: 03.

Hui Wen Goh and Jonas Mueller. 2023. ActiveLab: Active Learning with Re-Labeling by Multiple Annotators. *arXiv preprint*. ArXiv:2301.11856 [cs].

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT Press.

Cornelia Gruber, Katharina Hechinger, Matthias Assenmacher, Göran Kauermann, and Barbara Plank. 2024. More labels or cases? assessing label variation in natural language inference. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 22–32, Malta. Association for Computational Linguistics.

Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter, and Göran Kauermann. 2025. Sources of Uncertainty in Supervised Machine Learning – A Statisticians' View. *arXiv preprint*. ArXiv:2305.16703 [stat].

Trevor Hastie, Robert Tibshirani, Jerome H. Friedman, and Jerome H. Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Katharina Hechinger, Christoph Koller, Xiao Xiang Zhu, and Göran Kauermann. 2024a. Human-in-the-loop: Towards Label Embeddings for Measuring Classification Difficulty. *arXiv preprint*. ArXiv:2311.08874 [cs].

Katharina Hechinger, Xiao Xiang Zhu, and Göran Kauermann. 2024b. Categorising the world into local climate zones: towards quantifying labelling uncertainty for machine learning models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(1):143–161.

Oana Inel and Lora Aroyo. 2017. Harnessing Diversity in Crowds and Machines for Better NER Performance. In *The Semantic Web*, pages 289–304, Cham. Springer International Publishing.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine De Marneffe. 2023. Ecologically Valid Explanations for Label Variation in NLI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10622–10633, Singapore. Association for Computational Linguistics.

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: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics.

Christoph Koller, Göran Kauermann, and Xiao Xiang Zhu. 2024. Going Beyond One-Hot Encoding in Classification: Can Human Uncertainty Improve Model Performance in Earth Observation? *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–11. Conference Name: IEEE Transactions on Geoscience and Remote Sensing.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé Iii, and John Langford. 2019. Active Learning for Cost-Sensitive Classification. *Journal of Machine Learning Research*, 20(65):1–50.

David D. Lewis and Jason Catlett. 1994. Heterogeneous Uncertainty Sampling for Supervised Learning. In William W. Cohen and Haym Hirsh, editors, *Machine Learning Proceedings 1994*, pages 148–156. Morgan Kaufmann, San Francisco (CA).

David D. Lewis and William A. Gale. 1994. A Sequential Algorithm for Training Text Classifiers. *arXiv preprint*. ArXiv:cmp-lg/9407020.

Christopher Lin, M Mausam, and Daniel Weld. 2016. Re-Active Learning: Active Learning with Relabeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Dragos D. Margineantu. 2005. Active cost-sensitive learning. In *IJCAI*, volume 5, pages 1622–1623.

Tom M. Mitchell. 1997. *Machine Learning*, nachdr. edition. McGraw-Hill series in Computer Science. McGraw-Hill, New York.

Jingwei Ni, Yu Fan, Vilém Zouhar, Donya Rooein, Alexander Hoyle, Mrinmaya Sachan, Markus Leippold, Dirk Hovy, and Elliott Ash. 2025. Can Large Language Models Capture Human Annotator Disagreements? *arXiv preprint*. ArXiv:2506.19467 [cs].

Jennimaria Palomaki, Olivia Rhinehart, and Michael Tseng. 2018. A case for a range of acceptable annotations. In *SAD/CrowdBias@ HCOMP*, pages 19–31.

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. *Lang. Resour. Eval.*, 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. Place: Cambridge, MA Publisher: MIT Press.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625.

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, Gothenburg, Sweden. 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.

Lukas Rauch, Matthias Aßenmacher, Denis Huseljic, Moritz Wirth, Bernd Bischl, and Bernhard Sick. 2023. ActiveGLAE: A Benchmark for Deep Active Learning with Transformers. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pages 55–74, Cham. Springer Nature Switzerland.

Dennis Reidsma and Rieks Op Den Akker. 2008. Exploiting 'subjective' annotations. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics - HumanJudge '08*, pages 8–16, Manchester, United Kingdom. Association for Computational Linguistics.

Paul Röttger, 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.

Bertrand Russell. 1907. II.—On the Nature of Truth. *Proceedings of the Aristotelian Society*, 7(1):28–49.

Yusuf Sale, Viktor Bengs, Michele Caprio, and Eyke Hüllermeier. 2024. Second-order uncertainty quantification: a distance-based approach. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 43060–43076, Vienna, Austria. JMLR.org.

Christoph Schweden, Katharina Hechinger, Göran Kauermann, and Xiao Xiang Zhu. 2025. Can Uncertainty Quantification Benefit From Label Embeddings? A Case Study on Local Climate Zone Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14.

Burr Settles. 2009. Active Learning Literature Survey. *University of Wisconsin, Madison*, 52.

Burr Settles. 2011. From Theories to Queries: Active Learning in Practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010*, pages 1–18.

Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active Learning with Real Annotation Costs. In *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*.

H. S. Seung, M. Opper, and H. Sompolinsky. 1992. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, COLT '92, pages 287–294, New York, NY, USA. Association for Computing Machinery.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A Roadmap to Pluralistic Alignment. In *Proceedings of the 41 st International Conference on Machine Learning*, Vienna.

Katrin Tomanek and Udo Hahn. 2010. A comparison of models for cost-sensitive active learning. In *Coling 2010: Posters*, pages 1247–1255, Beijing, China. Coling 2010 Organizing Committee.

Dietrich Trautmann, Johannes Daxenberger, Christian Stab, Hinrich Schütze, and Iryna Gurevych. 2020. Fine-Grained Argument Unit Recognition and Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9048–9056. Number: 05.

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

Michiel van der Meer, Neele Falk, Pradeep K. Murukannaiah, and Enrico Liscio. 2024. Annotator-centric active learning for subjective NLP tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18537–18555, Miami, Florida, USA. Association for Computational Linguistics.

Xinpeng Wang and Barbara Plank. 2023. ACTOR: Active learning with annotator-specific classification heads to embrace human label variation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2046–2052, Singapore. 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.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. 2023. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 2282–2292. PMLR. ISSN: 2640-3498.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *Preprint*, arXiv:2407.19594.

Yu Xia, Subhojyoti Mukherjee, Zhouhang Xie, Junda Wu, Xintong Li, Ryan Aponte, Hanjia Lyu, Joe Barrow, Hongjie Chen, Franck Dernoncourt, Branislav Kveton, Tong Yu, Ruiyi Zhang, Jiuxiang Gu, Nesreen K. Ahmed, Yu Wang, Xiang Chen, Hanieh Deilamsalehy, Sungchul Kim, Zhengmian Hu, Yue Zhao, Nedim Lipka, Seunghyun Yoon, Ting-Hao Kenneth Huang, Zichao Wang, Puneet Mathur, Soumyabrata Pal, Koyel Mukherjee, Zhehao Zhang, Namyong Park, Thien Huu Nguyen, Jiebo Luo, Ryan A. Rossi, and Julian McAuley. 2025. From selection to generation: A survey of LLM-based active learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14552–14569, Vienna, Austria. Association for Computational Linguistics.

Kaige Xie, Cheng Chang, Liliang Ren, Lu Chen, and Kai Yu. 2018. Cost-sensitive active learning for dialogue state tracking. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 209–213, Melbourne, Australia. Association for Computational Linguistics.

Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. 2016. Active Learning from Imperfect Labelers. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Yan Yan, Glenn M. Fung, Rómer Rosales, and Jennifer G. Dy. 2011. Active learning from crowds. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1161–1168.

Yan Yan, Romer Rosales, Glenn Fung, Faisal Farooq, Bharat Rao, and Jennifer Dy. 2012. Active Learning from Multiple Knowledge Sources. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pages 1350–1357. PMLR. ISSN: 1938-7228.

Bo Yuan, Yulin Chen, Yin Zhang, and Wei Jiang. 2024. Hide and seek in noise labels: Noise-robust collaborative active learning with LLMs-powered assistance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10977–11011, Bangkok, Thailand. Association for Computational Linguistics.

Chicheng Zhang and Kamalika Chaudhuri. 2015. Active learning from weak and strong labelers. In *Advances in neural information processing systems*, volume 28. Curran Associates, Inc.

Jing Zhang, Xindong Wu, and Victor S. Shengs. 2015. Active Learning With Imbalanced Multiple Noisy Labeling. *IEEE Transactions on Cybernetics*, 45(5):1095–1107. Conference Name: IEEE Transactions on Cybernetics.

Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Ciccarelli, Frederik Barkhof, and Daniel C. Alexander. 2020. Disentangling human error from the ground truth in segmentation of medical images. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103, Singapore. Association for Computational Linguistics.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. A Survey of Active Learning for Natural Language Processing. In *Proceedings of the 2022 Conference* *on Empirical Methods in Natural Language Processing*, pages 6166–6190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liyue Zhao, Gita Sukthankar, and Rahul Sukthankar. 2011. Incremental Relabeling for Active Learning with Noisy Crowdsourced Annotations. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 728–733.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Weak Ensemble Learning from Multiple Annotators for Subjective Text Classification

Ziyi Huang[1], **N. R. Abeynayake**[2], **Xia Cui**[2]

[1]Hubei University, Wuhan, China.
ziyihuang@hubu.edu.cn
[2]Manchester Metropolitan University, Manchester, UK.
{n.abeynayake, x.cui}@mmu.ac.uk

Abstract

With the rise of online platforms, moderating harmful or offensive user-generated content has become increasingly critical. As manual moderation is infeasible at scale, machine learning models are widely used to support this process. However, subjective tasks, such as offensive language detection, often suffer from annotator disagreement, resulting in noisy supervision that hinders training and evaluation. We propose *Weak Ensemble Learning* (WEL), a novel framework that explicitly models annotator disagreement by constructing and aggregating weak predictors derived from diverse annotator perspectives. WEL enables robust learning from subjective and inconsistent labels without requiring annotator metadata. Experiments on four benchmark datasets show that WEL outperforms strong baselines across multiple metrics, demonstrating its effectiveness and flexibility across domains and annotation conditions.

1 Introduction

Harmful information, such as offensive and abusive language, has been known as one of the main threats on social media platforms. Typically, the moderation of online harmful information is conducted manually. With an increasing amount of information, manual moderation is expensive and insufficient. There is a growing demand for developing a Natural Language Processing (NLP) tool to support the detection and mitigation of harmful content on online platforms. Addressing this challenge requires high-quality annotated data to train accurate and reliable machine learning models. In recent years, social media has become a popular source for data collection, and crowdsourcing has emerged as a widely used solution for large-scale data annotation. However, concerns have been raised about the reliability of crowdworkers, particularly in complex linguistic tasks where annotators often lack domain-specific training (Uma et al., 2022). Furthermore, the incentive structures of crowdsourcing platforms can encourage rapid completion of tasks with careful judgment, potentially compromising label quality (Daniel et al., 2018; Leonardelli et al., 2021; Leonardellli et al., 2023). A particularly challenging issue in this context is human label variation (Plank, 2022), which arises when annotators assign different labels to the same instance. This is especially common in subjective tasks such as emotion detection (Buechel and Hahn, 2018) and offensive language detection (Leonardellli et al., 2023), where annotation involves personal interpretation, contextual nuance, and cultural perspective. Unlike objective tasks with clearly defined ground truth, subjective annotations inherently invite disagreement. Such variation introduces noise into training data, complicates evaluation, and challenges the assumption of a single "correct" label (Uma et al., 2022; Cabitza et al., 2023). Understanding and modelling this variability is critical for developing NLP systems that are more robust, interpretable, and aligned with the diversity of human judgement.

Previously, several methods have been proposed to address this issue by estimating and incorporating annotator reliability into the modelling process (Sheng et al., 2008; Cui, 2023; Fleisig et al., 2023; Xu et al., 2024). These approaches typically assign higher weights to labels provided by more consistent or trustworthy annotators, aiming to reduce the influence of noisy or unreliable inputs on the final model. However, their effectiveness is often limited by the composition of the annotator pool. They require sufficient diversity among annotators to model reliability accurately and may risk overfitting when such diversity is lacking or when the model overrelies on a small subset of annotators (Räbiger et al., 2018; Cui, 2023).

We aim to develop a method applicable to more general multi-annotation settings. Specifically, the proposed approach is designed to function effec-

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 87–99
November 8, 2025 ©2025 Association for Computational Linguistics

tively when annotators are shared across the entire dataset or when there is a heterogeneous distribution of annotator workload (i.e., some annotators contribute more than others).

While prior approaches often rely on a single loss function, such as cross-entropy (CE) (Uma et al., 2020), to train classification models, this may be insufficient for subjective tasks where both hard and soft supervisory signals are informative. In such settings, different loss components capture complementary aspects of learning: CE supports probabilistic calibration, F1 loss promotes classification accuracy on hard labels, and distributional losses like mean absolute error or Manhattan distance (MD) (Rizzi et al., 2024) help align predictions with the soft label distributions reflecting annotator disagreement. By jointly optimising these objectives, we can balance predictive accuracy with nuanced representation of label uncertainty, leading to more robust and interpretable models.

Our contributions can be summarised as follows:

- We propose *Weak Ensemble Learning*[1] (WEL), a novel ensemble-based framework for learning from multiple annotations in subjective tasks.
- We introduce two variants: *WEL-Random*, which builds weak predictors from randomly sampled labels to capture annotator variation without metadata, and *WEL-TopAnn*, which trains per-annotator models for the top-ranked annotators.
- We present a systematic study of selection strategies, aggregation methods and loss functions for optimising the ensemble.
- Experiments on four datasets from Le-Wi-Di 2023 shared task show that WEL consistently outperforms two strong baselines across multiple metrics.

2 Related Work

Subjective NLP tasks such as offensive language detection, hate speech classification and emotion analysis often suffer from high variability in human annotations. Annotators may interpret linguistic cues differently based on their personal, cultural or contextual backgrounds (Aroyo and Welty, 2015; Uma et al., 2022). This subjectivity introduces label noise and inconsistency, making it challenging to define a single ground truth (i.e., a hard label). In particular, datasets annotated via crowdsourcing tend to reflect these disagreements, raising questions about how best to represent and learn from multiple perspectives (Leonardelli et al., 2021; Davani et al., 2022).

A common approach to address label disagreement is to replace hard labels with soft targets, usually probability distributions over classes derived from annotator votes, and train models using probabilistic loss functions. The most prevalent is the cross-entropy loss, which treats soft distributions as targets, encouraging models to reflect label uncertainty rather than force a single decision (Uma et al., 2020; Zheng et al., 2021). More recent methods have proposed alternative loss formulations, such as Kullback-Leibler divergence, expected calibration error (Uma et al., 2020) and Manhattan distance (Rizzi et al., 2024). These techniques aim to improve robustness to noisy or subjective labels by preserving the signal in disagreement rather than collapsing it through majority vote.

Ensemble methods have also been explored as a way to leverage annotator disagreement rather than suppress it. Instead of aggregating labels before training, several works train separate models for each annotator and combine their predictions during inference (Akhtar et al., 2021; Gordon et al., 2021; Xu et al., 2024). This strategy captures the full range of annotator perspectives and has shown promise in capturing subjective variation in tasks like emotion classification and hate speech detection. However, these models may suffer from scalability issues, especially when the number of annotators is large or unbalanced. Other work has approached the problem from a probabilistic modelling perspective, estimating annotator reliability as a latent variable during training (Paun et al., 2018a,b; Xu et al., 2020). These approaches often combine annotator-specific models with global learning signals, aiming to balance personalised and consensus-based predictions. In addition, instance weighting has been used as a practical solution to reduce the influence of unreliable or biased supervision. For instance, Zhang et al. (2020) apply instance reweighting to mitigate demographic bias in toxicity detection, while Liu et al. (2021) introduce dynamic instance weighting to adapt to concept drift in evolving datasets. Cui (2023) and Fleisig et al. (2023) proposed to compute individual annotator ratings and combine this information to better capture the subjectivity inherent. These methods adjust the learning signal based on

[1]Codebase for WEL and Evaluation: `https://github.com/YhzyY/Weak-Ensemble-Learning`

example-level characteristics, enabling models to better generalise under noisy or imbalanced conditions.

Our work unifies ensemble-based disagreement modelling. We extend ensemble methods that capture annotator disagreement by randomly sampling weak predictors to simulate diverse viewpoints, and by embedding annotator-specific models whose ensemble selection is learned end-to-end rather than relying on a fixed set as in Xu et al. (2024). In contrast to probabilistic reliability estimation techniques that depend on annotator metadata (Paun et al., 2018a; Xu et al., 2020), our framework requires no such information, broadening its applicability. At the ensemble level, we adapt instance-weighting strategies to emphasise predictor utility and mitigate dataset bias (Zhang et al., 2020; Liu et al., 2021). Drawing on label distribution modelling, our loss function blends soft and hard supervision to achieve both nuanced learning and interpretability (Tian et al., 2024). These elements yield a scalable and flexible approach for managing noisy subjective annotations.

3 Methods

Given a dataset annotated by multiple annotators, the goal is to learn a predictive function that accounts for the variability and potential noise introduced by differing annotator judgments. Let $\mathcal{D} = \{(x_i, y_i^{(1)}, \dots, y_i^{(A_i)})\}_{i=1}^N$ denote a dataset of N instances, where $x_i \in \mathcal{X}$ is the input (e.g., a text sample), and $\{y_i^{(1)}, \dots, y_i^{(A_i)}\}$ are the labels provided by A_i annotators for instance x_i, with $y_i^{(j)} \in \mathcal{Y}$ representing the label from the j-th annotator, which $j \in \{1, \dots, J\}$ and J is the total number of annotators in $\mathcal{D}$. The objective is to learn a predictive function $f_\theta : \mathcal{X} \to \mathcal{Y}$ parameterised by θ, that approximates the underlying true label distribution y_i^*, which is unobserved due to annotator disagreement.

To address this challenge, we propose a three-stage method, named *Weak Ensemble Learning* (WEL), designed to learn from multiple annotators while accounting for disagreement and annotator variability. First, we construct a set of weak predictors by employing a *random sampling* and a *top-ranked annotators* selection strategies (Section 3.1). Second, we aggregate the outputs of weak predictors using a weighted ensemble, where the weights will be tuned to balance contributions in the next stage, enabling the model to leverage diverse annotator perspectives effectively (Section 3.2). Finally,

Algorithm 1 Weak Ensemble Learning (WEL)

Input: Dataset $\mathcal{D} = \{(x_i, \{y_i^{(j)}\}_{j=1}^{A_i})\}_{i=1}^N$;
Loss coefficients α, β, γ; Regularisation weight λ;
Maximum number of weak predictors $M_{\max}$
Output: Final predictive ensemble model $f(x) = \sum_{m=1}^{M^*} w_m f_{\theta_m}(x)$

Stage 1: Construct Weak Predictors
Strategy 1: Random Sampling
for $m = 1$ *to* M_{max} **do**
  $\mathcal{D}^{(m)} = \{(x_i, y_i^{(j)}) \mid j \sim \text{Uniform}(1, A_i)\}$
  Train f_{θ_m} on $\mathcal{D}^{(m)}$
end

Strategy 2: Top-Ranked Annotators
Compute the annotation counts of a set of annotators $\{A_1, \dots, A_{M_{\max}}\}$
for $m = 1$ *to* M_{max} **do**
  $\mathcal{D}^{(m)} = \{(x_i, y_i^{(A_m)}) \mid A_m \text{ annotated } x_i\}$
  Train f_{θ_m} on $\mathcal{D}^{(m)}$
end

Stage 2: Define Aggregated Supervision
foreach *instance* x_i **do**
  Compute hard-aggregated label:
$$\bar{y}_i^{\text{hard}} = \arg\max \sum_{j=1}^{A_i} (y_i^{(j)} = c),$$
  Compute soft-aggregated label:
$$\bar{y}_i^{\text{soft}}[c] = \frac{1}{A_i} \sum_{j=1}^{A_i} (y_i^{(j)} = c)$$
end

Stage 3: Joint Optimisation
Initialize ensemble size $M \leftarrow M_{\max}$ and weights $\mathbf{W} = [w_1, ..., w_M]$
repeat
  foreach *instance* x_i **do**
    Compute ensemble output:
$$\hat{y}_i = \sum_{m=1}^{M} w_m f_{\theta_m}(x_i)$$
  end
  Compute total loss:
$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{F1}} + \beta \cdot \mathcal{L}_{\text{CE}} + \gamma \cdot \mathcal{L}_{\text{MD}} + \lambda \cdot \Omega(\mathbf{W})$$
  Update ensemble weights $\mathbf{W}$
  Prune predictors: retain only f_{θ_m} such that $w_m > \epsilon$, for $m = 1, \dots, M$
  Reinitialise weights: $\mathbf{W} \leftarrow \mathbf{W} / \sum_{m=1}^{M} w_m$
until *convergence*;
Set $M^* \leftarrow M$ and return final ensemble: $f(x) = \sum_{m=1}^{M^*} w_m f_{\theta_m}(x)$

we jointly optimise the weak predictors' ensemble weights by minimising a multi-objective loss over soft and hard aggregated labels, balancing cross-entropy, distributional similarity, and F1-score performance (Section 3.3). The complete procedure of WEL is described in Algorithm 1.

3.1 Weak Predictor Construction

To capture diverse annotator perspectives, the first stage of WEL constructs M weak predictors $\{f_{\theta_1}, \ldots, f_{\theta_M}\}$, each trained on a different slice of the annotation space. We propose two selection strategies:

Random Annotator Selection. For each training instance x_i with A_i annotations $\{y_i^{(1)}, \ldots, y_i^{(A_i)}\}$, we sample one label $y_i^{(j)}$ uniformly at random:

$$j \sim \text{Uniform}\{1, \ldots, A_i\} \tag{1}$$

Repeating this process M times produces M datasets $\{\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(M)}\}$, each reflecting a single-annotator view.

Top-Ranked Annotator Selection. We identify the M annotators with the largest label contributions and train one weak predictor per annotator using only their labels. Unlike Xu et al. (2024), which assumes a fixed set of annotators, our M is treated as a *learnable parameter* in the optimisation stage, allowing the ensemble size to adapt to the dataset.

Model Architecture. Each weak predictor f_{θ_m} consists of a Transformer encoder (BERT or AraBERT) followed by a linear classification head mapping the [CLS] representation to class logits:

$$z = W h_{\texttt{[CLS]}} + b, \tag{2}$$

where $h_{\texttt{[CLS]}} \in \mathbb{R}^d$ is the encoder output, $W \in \mathbb{R}^{C \times d}$, $b \in \mathbb{R}^C$, and C is the number of classes. The logits are passed through a softmax layer to produce probability distributions over classes:

$$\hat{y} = \text{softmax}(z) \tag{3}$$

This stage yields a diverse pool of predictors that differ in training data and potentially in decision boundaries, forming the foundation for weighted ensemble learning in Section 3.2.

3.2 Weighted Ensemble Learning

In the second stage, we aggregate the probability outputs from the M weak predictors $\{f_{\theta_1}, \ldots, f_{\theta_M}\}$ into a single ensemble prediction. Let $\hat{y}_i^{(m)} \in [0,1]^C$ denote the predicted probability distribution over C classes for instance x_i from the m-th weak predictor, computed via the softmax output of its linear classification head (Section 3.1).

We adopt a *weighted ensemble* strategy, where each predictor is assigned a learnable non-negative weight w_m subject to the constraint $\sum_{m=1}^{M} w_m = 1$. The ensemble prediction is then:

$$\hat{y}_i = \sum_{m=1}^{M} w_m \, \hat{y}_i^{(m)} \tag{4}$$

Here, $\mathbf{W} = [w_1, \ldots, w_M] \in \mathbb{R}^M_{\geq 0}$ encodes the contribution of each weak predictor to the final decision.

While up to M_{max} predictors can be initially constructed, the optimisation process (Section 3.3) automatically determines an effective subset $M^* \leq M_{\text{max}}$. Predictors with $w_m < \epsilon$ (e.g., $\epsilon = 10^{-3}$) are pruned to improve computational efficiency and reduce noise from low-utility models.

By combining multiple probability distributions, this ensemble mechanism captures complementary information from diverse annotator views, improving robustness and mitigating the bias of any single weak predictor.

3.3 Optimisation

In the third stage, we optimise the ensemble weights $\{w_m\}_{m=1}^{M}$ (and optionally other parameters) to improve predictive performance. Given the ensemble prediction $\hat{y}_i$ from Eq. (4), computed as the weighted sum of individual predictor outputs $\hat{y}_i^{(m)}$, our goal is to minimise a multi-objective loss that balances classification accuracy, calibration, and distributional alignment.

To accommodate the uncertainty introduced by annotator disagreement, we investigate learning from both *soft-aggregated* and *hard-aggregated* labels, and explore separate and joint optimisation strategies based on multiple objective functions.

3.3.1 Aggregated Supervision

Let $\mathcal{D} = \{(x_i, \{y_i^{(j)}\}_{j=1}^{A_i})\}_{i=1}^{N}$ be a dataset annotated by multiple annotators. We derive two forms of supervision:

- **Hard Aggregated Label** $\bar{y}_i^{\text{hard}} \in \mathcal{Y}$: computed via majority vote over annotator labels.

- **Soft Aggregated Label** $\bar{y}_i^{\text{soft}} \in [0,1]^C$: a normalised label distribution over C classes, reflecting the empirical frequency of annotators' choices.

3.3.2 Objectives

To robustly train the ensemble model under varying supervision signals, we define the following optimisation targets of the loss function $\mathcal{L}$:

(1) F1-Score (F1): A discrete metric evaluated using $\bar{y}_i^{\text{hard}}$, which we aim to maximise:

$$\mathcal{L}_{\text{F1}} = -\text{F1}(\arg\max(\hat{y}_i), \bar{y}_i^{\text{hard}}), \quad (5)$$

where the negative sign denotes that the F1-score is being maximised during training.

(2) Cross-Entropy (CE) (Uma et al., 2020; Leonardellli et al., 2023): A soft objective used when training with $\bar{y}_i^{\text{soft}}$, minimising:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N}\sum_{c=1}^{C} \bar{y}_i^{\text{soft}}[c] \cdot \log \hat{y}_i[c], \quad (6)$$

where N is the number of training instances, C the number of classes, $\bar{y}_i^{\text{soft}}[c]$ the soft target (i.e., annotator-derived label distribution), and $\hat{y}_i[c]$ the predicted probability for class c on instance x_i.

(3) Average Manhattan Distance (MD) (Rizzi et al., 2024): A distributional similarity measure minimising the L_1 distance between predicted and soft labels:

$$\mathcal{L}_{\text{MD}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} |\hat{y}_i[c] - \bar{y}_i^{\text{soft}}[c]|_1 \quad (7)$$

3.3.3 Separate and Joint Optimisation

We explore two optimisation paradigms:

- **Separate Optimisation :** Each objective is minimised independently in different optimising regimes. For example, cross-entropy is minimised on soft labels, while F1-score is optimised using hard labels during model aggregation.

- **Joint Optimisation:** A combined loss function integrates all objectives, Eq. (5), (6) and (7), to guide the model jointly. We define:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{F1}} + \beta \cdot \mathcal{L}_{\text{CE}} + \gamma \cdot \mathcal{L}_{\text{MD}} + \lambda \cdot \Omega(\mathbf{W}), \quad (8)$$

where $\alpha, \beta, \gamma \geq 0$ are loss balancing coefficients, and $\Omega(\mathbf{W})$ is an ℓ_2-norm regularisation term to encourage balanced weight distributions to prevent overfitting. The regularisation weight λ controls the degree of smoothing across predictors.

By jointly optimising prediction objectives and ensemble composition, our model leverages annotator disagreement as a source of diversity, improving both robustness and alignment with subjective supervision.

3.3.4 Implementation Details

During the optimisation stage, we employ two derivative-free optimisation algorithms: OPTUNA (Akiba et al., 2019) and the `SciPy`[2] differential evolution algorithm (Storn and Price, 1997). Both are well-suited to searching continuous, bounded parameter spaces without requiring gradient information. In our setting, the optimiser iteratively updates the ensemble weights $\mathbf{W}$ to minimise the chosen objective(s) (either a single loss or the joint formulation in Eq. (8)), subject to the simplex constraint:

$$w_m \geq 0, \quad \sum_{m=1}^{M} w_m = 1. \quad (9)$$

For each optimisation step, the selected subset of weak predictors is reinitialised to reduce sensitivity to specific model subsets. We run each optimiser for up to 100 trials or steps, and both methods yield comparable results. OPTUNA is generally faster due to GPU support and efficient sampling strategies, while differential evolution offers robust CPU-based parallelism, making it preferable in non-GPU environments. The framework remains agnostic to the choice of optimiser, allowing other search strategies to be integrated as needed.

4 Experiments

To evaluate the performance of our method across diverse domains and text genres, we use

[2] `https://docs.scipy.org/`

Table 1: Data statistics for the four textual datasets. #Train, #Dev, and #Test denote the number of instances in the training, development, and test splits, respectively. #TotalAnn indicates the total number of annotators in each dataset, while #Ann represents the minimum and maximum number of annotators per instance.

Dataset	#Train	#Dev	#Test	#TotalAnn	#Ann
ArMIS	657	141	145	3	3
ConvAbuse	2398	812	840	8	2-7
HS-Brexit	784	168	168	6	6
MD-Agreement	6592	1104	3057	670	5

four publicly available datasets from the Le-Wi-Di shared task at SemEval 2023 (Leonardellli et al., 2023): **ArMIS** (Almanea and Poesio, 2022), **ConvAbuse** (Cercas Curry et al., 2021), **HS-Brexit** (Akhtar et al., 2021), and **MD-Agreement** (Leonardelli et al., 2021). Each dataset includes multiple annotations per instance, with at least two annotators per instance sample.

To maintain the generalisability of our approach, we exclude domain information and annotator metadata during training. All models are trained solely on input text and its associated labels. Summary statistics are provided in Table 1, while Table 2 presents dataset meta-information. In particular, we distinguish between *Fixed Ann.* datasets, where each instance is labelled by the same group of annotators, and *Mixture* datasets, where annotators vary across instances. Further details on the datasets and preprocessing steps are provided in Appendix A.

Table 2: Dataset metadata covering annotator contribution, diversity, language and genre.

Dataset	Contribution	Diversity	Language	Genre
ArMIS	Fixed Ann.	Low	Arabic	Short Text
ConvAbuse	Mixture	Low	English	Conversation
HS-Brexit	Fixed Ann.	Low	English	Short Text
MD-Agreement	Mixture	High	English	Short Text

4.1 Training

While the proposed framework is model-agnostic and compatible with various machine learning architectures, we employ BERT (Devlin et al., 2019) for English datasets (ConvAbuse, HS-Brexit, MD-Agreement) and AraBERTv2 (Antoun et al., 2020) for ArMIS, using the base checkpoints from HuggingFace. We train $M_{\max}$=10 weak predictors using different selection strategies (Section 3.1). Hyperparameters for Transformers are tuned on development sets (Appendix B). The predictors are fixed before the joint optimisation of ensemble weights.

4.2 Evaluation Metrics

We evaluate model performance using three complementary metrics: (a) micro-averaged F1 score (F1), which assesses classification accuracy on hard-aggregated labels; (b) cross-entropy loss (CE); and (c) average Manhattan distance (MD) between predicted and target label distributions. The latter two metrics are used to evaluate how well the model captures soft supervision signals arising from annotator disagreement (Leonardellli et al., 2023; Rizzi et al., 2024).

4.3 Label Selection Strategies

First, we experiment with two label selection strategies for constructing weak predictors: *random sampling* (*Random*), which selects one annotation per instance uniformly at random, and *top-ranked annotators* (*TopAnn*), which trains one model per annotator using data from the most frequent annotators. For simplicity and fair comparison, we fix all loss coefficients and the regularisation weight to 1.

Table 3 shows results across the four datasets. The *Random* strategy consistently achieves higher F1 and better CE than *TopAnn*. We attribute this to the greater diversity introduced by random sampling: each weak predictor is trained on a unique stochastic projection of the label space, encouraging the ensemble to learn decision boundaries that generalise across annotator-specific biases. This is especially beneficial when F1 is the main objective, as it rewards consistent hard-label predictions on majority-vote labels, which Random sampling implicitly approximates over many diverse predictors.

By contrast, *TopAnn* tends to produce more similar decision boundaries within the ensemble because each predictor is tied to a single annotator's style. This can be beneficial for modelling annotator-specific distributions, but under fixed coefficients, it can limit the ensemble's ability to optimise for F1, which benefits from capturing the aggregate rather than individual perspectives.

Nevertheless, *TopAnn* achieves lower MD on ConvAbuse and HS-Brexit, likely because these datasets have annotators with high internal consistency. In such cases, modelling them individually yields predictions more aligned with the soft label distribution.

Table 3: The selection strategies for constructing weak predictors on AraBERT and BERT.

Dataset	Selection	F1	CE	MD
ArMIS	Random	**0.7310**	**0.6390**	**0.5301**
	TopAnn	**0.7310**	0.6536	0.5487
ConvAbuse	Random	**0.9333**	**0.5559**	0.1749
	TopAnn	0.9310	0.5652	**0.1645**
HS-Brexit	Random	**0.9107**	**0.5842**	0.2733
	TopAnn	0.8929	0.6140	**0.2379**
MD-Agreement	Random	**0.8162**	**0.6246**	**0.3648**
	TopAnn	0.7668	0.6695	0.4156

4.4 Ensemble Optimisation Paradigms

We conduct an ablation study to assess the individual and combined contributions of the loss components in Eq. (8): $\mathcal{L}_{\text{F1}}$, $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{MD}}$. For clarity, we fix the selection strategy to *Random* and activate specific losses by setting their corresponding coefficients (α, β, γ) to 1 while setting the others to 0. In each setting, we optimise both the ensemble weights $\mathbf{W}$ and the number of members M.

Tables 4 and 5 show results for the ArMIS and MD-Agreement datasets. Across both datasets, $\mathcal{L}_{\text{MD}}$ consistently achieves the lowest MD values, confirming its role in aligning predictions with annotator label distributions. Similarly, configurations including $\mathcal{L}_{\text{CE}}$ tend to improve calibration (lower CE), while $\mathcal{L}_{\text{F1}}$ boosts classification accuracy when paired with $\mathcal{L}_{\text{MD}}$. However, using all three objectives together does not yield additional gains, and in some cases slightly reduces performance, likely due to competing optimisation signals. Overall, these results suggest that each loss serves a distinct purpose: $\mathcal{L}_{\text{F1}}$ strengthens hard-label accuracy, $\mathcal{L}_{\text{CE}}$ improves probabilistic calibration, and $\mathcal{L}_{\text{MD}}$ enhances alignment with annotator distributions. Effective combinations emerge when the selected losses complement rather than compete, even without tuning the loss coefficients, underscoring the value of a flexible and modular objective in ensemble optimisation. The ConvAbuse and HS-Brexit datasets corroborate these findings, with further analysis provided in Appendix D. Similar results were also found using *TopAnn*.

4.5 Loss Coefficients and Regularisation Term

The joint objective in Eq. (8) balances four components through parameters (α, β, γ and λ) with the regularisation term $\Omega(\mathbf{W})$ demonstrating three key effects. Due to the page limit, we present the impact of $\Omega(\mathbf{W})$ on **MD-Agreement** in Table 6: (1) F1 improvement (up to +0.0056), (2) CE reduction

Table 4: Ablation study of loss optimisation paradigms on ArMIS dataset. In each setting, one loss component is activated (associated scaler set to 1), while the remaining components are deactivated (set to 0).

Case	F1	CE	MD
$\mathcal{L}_{\text{F1}}$ only	0.7448	0.6395	0.5048
$\mathcal{L}_{\text{CE}}$ only	0.7379	**0.6385**	0.5252
$\mathcal{L}_{\text{MD}}$ only	0.7379	0.6505	**0.4900**
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{CE}}$	0.7448	0.6412	0.5179
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{MD}}$	0.7172	0.6505	0.5111
$\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$	**0.7517**	0.6406	0.5294
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$	0.6897	0.6468	0.5243

Table 5: Ablation study of loss optimisation paradigms on MD-Agreement dataset.

Case	F1	CE	MD
$\mathcal{L}_{\text{F1}}$ only	0.8132	0.6249	0.3672
$\mathcal{L}_{\text{CE}}$ only	0.8119	0.6246	0.3633
$\mathcal{L}_{\text{MD}}$ only	0.8165	0.6250	**0.3626**
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{CE}}$	0.8145	**0.6245**	0.3660
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{MD}}$	**0.8175**	**0.6245**	0.3670
$\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$	0.8109	0.6247	**0.3626**
$\mathcal{L}_{\text{F1}}$+$\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$	0.8142	**0.6245**	0.3647

(max -0.0005 for $\mathcal{L}_{\text{CE}}$+$\mathcal{L}_{\text{MD}}$) and (3) MD gains in soft supervision (-0.0008) with limited degradation ($\leq$+0.0020 for $\mathcal{L}_{\text{MD}}$ alone).

We conduct a Spearman correlation analysis (Kendall and Stuart, 1969) over four parameters in the objective function, each sampled from the range $[0, 0.001, 0.01, 0.1, 1]$, resulting in 1,295 unique combinations per dataset (excluding 0s for all). The F1 coefficient α significantly improves F1 ($\geq$+0.9) while degrading MD ($\geq$+0.7), with similar but weaker trade-offs for β (CE-focused) and γ (MD-focused). The regularisation strength λ shows model-dependent effects, enhancing F1 on BERT ($\approx$+1.0) but reducing performance on AraBERT ($\leq$-0.9). Finally, the optimised number of weak predictors M strongly correlates with both improved F1 ($\geq$+0.9) and reduced CE ($\leq$-0.9), though typically at the cost of MD degradation (ΔMD $\geq$+0.5) in BERT implementations.

4.6 Model Aggregation Strategies

Table 9 presents results on four datasets using three aggregation strategies for combining weak predictors: (a) *Voting*, which applies majority voting over class labels; (b) *Averaging*, which computes the unweighted mean of probabilistic outputs; and (c) *Optimised*, which learns weighted combinations through loss-minimising ensemble optimisation. In binary classification settings, *Voting* and *Averag-*

Table 6: Improvements when adding regularisation term $\Omega(\mathbf{W})$, Δ = with $\Omega(\mathbf{W})$ - without $\Omega(\mathbf{W})$.

Case	ΔF1	ΔCE	ΔMD
$\mathcal{L}_{F1}$ only	+0.0033	-0.0003	-0.0013
$\mathcal{L}_{CE}$ only	+0.0039	-0.0001	+0.0001
$\mathcal{L}_{MD}$ only	+0.0026	-0.0005	+0.0020
$\mathcal{L}_{F1}$+$\mathcal{L}_{CE}$	0.0000	0.0000	0.0000
$\mathcal{L}_{F1}$+$\mathcal{L}_{MD}$	-0.0036	+0.0001	+0.0010
$\mathcal{L}_{CE}$+$\mathcal{L}_{MD}$	+0.0056	-0.0005	-0.0008
$\mathcal{L}_{F1}$+$\mathcal{L}_{CE}$+$\mathcal{L}_{MD}$	+0.0020	+0.0001	+0.0001

Table 7: Correlation between parameter and evaluation metrics (F1, CE and MD) on the ArMIS and MD-Agreement datasets. * indicates statistical significance ($p < 0.05$). Green indicates improvement, red indicates degradation. For CE and MD, negative correlations are desirable.

Dataset	ArMIS			MD-Agreement		
Param	F1	CE	MD	F1	CE	MD
α	+0.9*	+0.7	+1.0*	+1.0*	+0.6	+1.0*
β	+0.2	-1.0*	+0.4	-0.7	-0.6	+0.3
γ	-0.9*	+1.0*	-1.0*	-0.7	-0.9*	-1.0*
λ	-1.0*	-1.0*	+0.7	+0.9*	-0.3	+1.0*
M	+0.01	-1.0*	-0.37	+0.98*	-0.97*	+0.82*

ing yield identical predictions under a shared 0.5 threshold (Hovy et al., 2013; Plank et al., 2014). Across all datasets, the *Optimised* strategy consistently achieves superior performance in F1 and MD, highlighting the benefit of learning ensemble weights tailored to the task and supervision signal. A slight performance drop is observed in CE on the ConvAbuse and HS-Brexit datasets. This may be due to the optimisation process prioritising improvements in classification accuracy (F1) and distributional alignment (MD), potentially at the expense of precise probabilistic calibration (CE).

4.7 Comparison with Baseline Models

To ensure a fair comparison, we use the same model backbone with identical hyperparameters (BERT for English datasets and AraBERT for ArMIS). We reimplement and evaluate two baseline approaches:

- **BERT-CE** (Uma et al., 2020): a non-ensemble single model optimised using a CE-focused soft loss function.
- **Top-5 Annotator Voting (Top-5 Voting)** (Xu et al., 2024): a majority-vote ensemble of per-annotator models, each trained on labels from one of the top 3 or 5 most frequent annotators (depending on availability). Unlike the original version, which used multiple BERT variants, we adopt a uniform model architecture across all predictors for consistency.

Table 8 shows the best results of our proposed method under two selection strategies: *random sampling* (WEL-Random) and *top-ranked annotator* (WEL-TopAnn). Results correspond to the optimal configurations found via ensemble optimisation and parameter tuning (Appendix C). Both WEL variants consistently outperform the baselines across most evaluation metrics, demonstrating the effectiveness of jointly optimising ensemble weights while capturing annotator diversity through weak predictors. The only exceptions occur in MD on the ConvAbuse and HS-Brexit datasets, where WEL-TopAnn outperforms both WEL-Random and the baselines. Additionally, in terms of F1, WEL-Random consistently exceeds the baselines, reinforcing the robustness of the ensemble approach even with random annotator selection. As noted in Section 4.3, the superior performance of WEL-TopAnn in MD likely reflects the influence of a few highly consistent annotators, which benefits the top-ranked selection strategy. However, WEL-Random remains competitive across other metrics (F1 and CE), suggesting that the ensemble framework is effective even without explicit annotator ranking.

5 Conclusions

In this paper, we introduced *Weak Ensemble Learning* (WEL), a flexible framework for subjective text classification that learns from multiple annotations by constructing diverse weak predictors and jointly optimising their contributions. We explored two variants: WEL-Random, which captures annotator variation through random label sampling, and WEL-TopAnn, which models the most frequent annotators individually. Experiments on four datasets showed that WEL consistently outperforms baselines, with WEL-Random excelling in hard-label classification and WEL-TopAnn offering advantages in distributional alignment when annotator consistency is high. Future work will integrate annotator profiles and reliability estimates into a unified neural architecture to improve performance and efficiency, and extend WEL to larger annotator pools and multilingual contexts.

Limitations

Although our method provides a general and scalable approach to learning from annotator disagreement, it has several limitations. First, we train weak

Table 8: Comparison with baseline models. * indicates a statistically significant difference ($p < 0.05$, t-test) from the BERT-CE baseline in terms of predicted labels (hard evaluation metric, F1) or soft distributions (soft evaluation metrics, CE and MD).

Dataset	ArMIS			ConvAbuse			HS-Brexit			MD-Agreement		
Metric	F1	CE	MD	F1	CE	MD	F1	CE	MD	F1	CE	MD
BERT-CE	0.6596	0.8039	0.7144	0.8362	0.9671	4.8068	0.7917	0.7652	0.7985	0.7880	0.9948	1.7574
Top-5 Voting	0.7310	0.6529	0.5498	0.9310	0.5651	0.1648	0.8929*	0.6154*	0.2394*	0.7808*	0.6629*	0.3995*
WEL-Random	**0.7793***	0.6385	**0.5028**	**0.9405**	**0.5577**	0.1709	**0.9167**	**0.5889***	0.2585*	**0.8214***	**0.6245***	**0.3632***
WEL-TopAnn	0.7448	**0.6362**	0.5143	0.9321	0.5662	**0.1586**	0.8929*	0.6237*	**0.2354***	0.7815*	0.6636*	0.4034*

Table 9: Aggregation strategies for the weak predictors.

Dataset	Method	F1	CE	MD
ArMIS	Voting	0.7172	0.6389	0.5216
	Averaging	0.7172	0.6389	0.5216
	Optimised	**0.7793**	**0.6385**	**0.5028**
ConvAbuse	Voting	0.9333	**0.5545**	0.1814
	Averaging	0.9333	**0.5545**	0.1814
	Optimised	**0.9405**	0.5577	**0.1709**
HS-Brexit	Voting	0.9107	**0.5845**	0.2874
	Averaging	0.9107	**0.5845**	0.2874
	Optimised	**0.9167**	0.5889	**0.2585**
MD-Agreement	Voting	0.8178	**0.6245**	0.3659
	Averaging	0.8178	**0.6245**	0.3659
	Optimised	**0.8214**	**0.6245**	**0.3632**

predictors independently and do not update their parameters during joint optimisation. Although this design improves computational efficiency, it can limit the capacity of the ensemble to adapt if individual predictors are poorly calibrated or suboptimal. Second, while we evaluate across multiple data sources, our experiments are limited to mostly short social media texts (3/4), two languages and binary classification settings for simplicity. Additional evaluation of long-form text, multilingual corpora or structured annotation settings would help assess generalisability (Uma et al., 2022).

Ethical Statements

This research uses publicly available datasets from the SemEval-2023 Le-Wi-Di shared task, including user-generated content from social media and conversational agents. The datasets contain potentially sensitive language related to hate speech, offensive content, and abuse and were originally collected and annotated under ethical guidelines by their respective authors. We do not attempt to identify or profile any individual users or annotators. Our work focuses on improving the robustness and fairness of machine learning models in the presence of subjective disagreement and does not aim to make normative judgments about content or annotators. To support reproducibility and transparency, we use standard preprocessing, avoid introducing annotator-level biases, and refrain from incorporating demographic or personal information. All experiments are conducted following standard ethical practices for human-centred AI research, with a focus on minimising harm and respecting annotator diversity.

Acknowledges

We are deeply grateful to the reviewers for their thorough evaluation and insightful recommendations, which helped us enhance the clarity, rigour and impact of this paper.

References

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. *arXiv preprint arXiv:2106.15896*.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. pages 9–15, Marseille, France. European Language Resource Association.

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

Sven Buechel and Udo Hahn. 2018. Emotion representation mapping for automatic lexicon construction (mostly) performs on human level. In *Proceedings* *of the 27th International Conference on Computational Linguistics*, pages 2892–2904, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

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.

Xia Cui. 2023. xiacui at SemEval-2023 task 11: Learning a model in mixed-annotator datasets using annotator ranking scores as training weights. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1076–1084, Toronto, Canada. Association for Computational Linguistics.

Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys (CSUR)*, 51(1):1–40.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Mitchell A Gordon, Kevin Duh, and Jared Kaplan. 2021. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5915–5922, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Maurice G Kendall and Alan Stuart. 1969. The advanced theory of statistics. vol. 3. *Biometrics*, 25(2):435.

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 Leonardellli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Massimo Poesio, Verena Rieser, and Alexandra Uma. 2023. SemEval-2023 Task 11: Learning With Disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Anjin Liu, Jie Lu, and Guangquan Zhang. 2021. Diverse instance-weighting ensemble based on region drift disagreement for concept drift adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):293–307.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018a. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Silviu Paun, Jon Chamberlain, Udo Kruschwitz, Juntao Yu, and Massimo Poesio. 2018b. A probabilistic annotation model for crowdsourcing coreference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1926–1937, Brussels, Belgium. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. 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, Gothenburg, Sweden. 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.

Stefan Räbiger, Myra Spiliopoulou, and Yücel Saygın. 2018. How do annotators label short texts? toward understanding the temporal dynamics of tweet labeling. *Information Sciences*, 457-458:29–47.

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*, KDD '08, page 614–622, New York, NY, USA. Association for Computing Machinery.

Rainer Storn and Kenneth Price. 1997. Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces. *Journal of Global Optimization*, 11(4):341–359.

Xiaoyu Tian, Yongbin Qin, Ruizhang Huang, and Yanping Chen. 2024. A Label Information Aware Model for Multi-label Text Classification. *Neural Processing Letters*, 56(5):242.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):173–177.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. volume 72, page 1385–1470, El Segundo, CA, USA. AI Access Foundation.

Jin Xu, Mariët Theune, and Daniel Braun. 2024. Leveraging annotator disagreement for text classification. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 1–10, Trento. Association for Computational Linguistics.

Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. 2020. Automatic perturbation analysis for scalable certified robustness and beyond. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4134–4145, Online. Association for Computational Linguistics.

Yuanhang Zheng, Zhixing Tan, Meng Zhang, Mieradilijiang Maimaiti, Huanbo Luan, Maosong Sun, Qun Liu, and Yang Liu. 2021. Self-supervised quality estimation for machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3322–3334, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Experimental Data Preprocessing and Implementation Details

Three of the datasets (**ArMIS**, **HS-Brexit**, and **MD-Agreement**) consist of tweets collected from X[3]. The **ArMIS** dataset comprises Arabic tweets labelled for misogyny detection, focusing on offensive language directed toward women. The **HS-Brexit** dataset includes English tweets annotated for hate speech related to Brexit. The **MD-Agreement** dataset contains English tweets labelled for offensive language across three domains: Black Lives Matter, Elections, and COVID-19. To maintain the generalisability of our approach, we do not use domain information from the MD-Agreement dataset during training. For these three Twitter-based datasets, we apply a standardised preprocessing pipeline that includes the removal of HTML tags, URLs, hashtags, user mentions (@names), punctuation, non-ASCII characters, digits and redundant whitespace.

The fourth dataset, **ConvAbuse**, differs from the others as it is not sourced from social media but consists of English dialogues between users and two conversational agents. We include it to assess the model's performance on a different text genre: conversational dialogue. The original annotations span five levels of abuse severity, from -3 (highly abusive) to 1 (non-abusive). We simplify this into a binary classification task, labelling utterances with severity < 0 as offensive and those with severity ≥ 0 as non-offensive. For processing, we concatenate each dialogue into a single text sequence.

B Hyperparameter Tuning for BERT

We supervise each weak predictor using a joint objective function as in Eq. (8) combining: (a) the F1 micro score computed with hard labels (F1), (b) the cross-entropy loss with soft target distributions (CE), and (c) the average Manhattan distance (MD). For hyperparameter tuning of a single BERT model, we fixed all loss coefficients and regularisation weight to 1 to simplify the optimisation landscape. To ensure consistency and fair comparison across datasets, we use the ConvAbuse dataset, which is moderate in size relative to the others, to tune hyperparameters for fine-tuning BERT. Hyperparameter optimisation is performed using OPTUNA (Akiba et al., 2019). The model is trained on the training set and validated on the development set, while the test set remains completely unseen during both training and tuning.

The search is performed over 10 trials. The Weight decay is held constant at 0.01 to enforce moderate parameter shrinkage, preventing overfitting while avoiding excessive bias in the learned weights. The following hyperparameters are optimised: the learning rate (lr, sampled logarithmically in the range $[10^{-6}, 10^{-4}]$), number of training epochs (n_ep $\in [2, 5]$), batch size (bs $\in \{4, 8, 16, 32, 64\}$), and the number of warm-up steps (w_steps $\in [1, 500]$). The detailed results are shown in Table 10. These optimal settings are then applied uniformly across all datasets to ensure a consistent training setup.

Table 10: Hyperparameter tuning results sorted by joint loss (ascending). Pruned trials (6 and 9) are excluded. Bold values indicate best performance per column: highest F1, and lowest CE, MD and Joint loss.

Trial	lr	n_ep	bs	w_steps	F1	CE	MD	Joint
2	4.56e-6	2	32	463	0.8362	0.9678	**1.1516**	**1.2832**
0	1.48e-6	2	16	41	0.8362	0.9718	1.2891	1.4246
7	2.65e-6	3	16	342	0.8362	0.9753	1.2992	1.4383
3	4.75e-6	5	64	221	0.8818	0.8161	1.5719	1.5062
4	7.24e-5	5	8	487	**0.9470**	**0.8082**	2.0893	1.9504
1	9.80e-5	3	32	12	0.9434	0.8617	3.2045	3.1228
5	9.57e-5	5	32	150	0.9360	0.8409	3.5261	3.4310
8	4.35e-5	5	4	94	0.9200	0.8737	4.8390	4.7928

C Best Parameters for WEL

We perform a grid search over four parameters in the objective function, each sampled from the range $[0, 0.001, 0.01, 0.1, 1]$, resulting in 1,295 unique combinations per dataset (excluding 0s for all). Table 11 reports the best-performing parameter configurations (α, β, γ and λ) for the two variants of our proposed method: WEL-Random and WEL-TopAnn. The optimal values vary across datasets and selection strategies, indicating that performance is sensitive to the interplay of loss components. Notably, there is no consistent trend suggesting that any single parameter dominates

Table 11: Best performing hyperparameters for WEL.

	Method	α	β	γ	λ
ArMIS	WEL-Random	1	0.0001	0.01	0.001
	WEL-TopAnn	0.001	0.0001	0.1	0
ConvAbuse	WEL-Random	0	0.1	1	0
	WEL-TopAnn	0	1	0.01	0.01
HS-Brexit	WEL-Random	1	0.001	0	0.001
	WEL-TopAnn	0.1	0.1	0	0.001
MD-Agreement	WEL-Random	0.001	0.0001	0	0.001
	WEL-TopAnn	1	0	0	0

[3]https://X.com/

Table 12: Ablation study of loss optimisation paradigms on ConvAbuse dataset.

Case	F1	CE	MD
$\mathcal{L}_{F1}$ only	0.9298	0.5573	0.1862
$\mathcal{L}_{CE}$ only	0.9286	**0.5536**	0.1801
$\mathcal{L}_{MD}$ only	0.9202	0.5737	**0.1680**
$\mathcal{L}_{F1}$+$\mathcal{L}_{CE}$	**0.9333**	0.5540	0.1777
$\mathcal{L}_{F1}$+$\mathcal{L}_{MD}$	**0.9333**	0.5575	0.1689
$\mathcal{L}_{CE}$+$\mathcal{L}_{MD}$	0.9298	0.5556	0.1755
$\mathcal{L}_{F1}$+$\mathcal{L}_{CE}$+$\mathcal{L}_{MD}$	**0.9333**	0.5571	0.1707

Table 13: Ablation study of loss optimisation paradigms on HS-Brexit dataset.

Case	F1	CE	MD
$\mathcal{L}_{F1}$ only	0.8988	0.5868	0.2636
$\mathcal{L}_{CE}$ only	**0.9048**	**0.5851**	0.2859
$\mathcal{L}_{MD}$ only	0.8750	0.6066	0.2325
$\mathcal{L}_{F1}$+$\mathcal{L}_{CE}$	**0.9048**	0.5857	0.2789
$\mathcal{L}_{F1}$+$\mathcal{L}_{MD}$	0.8810	0.6182	0.2374
$\mathcal{L}_{CE}$+$\mathcal{L}_{MD}$	0.8929	0.6035	**0.2306**
$\mathcal{L}_{F1}$+$\mathcal{L}_{CE}$+$\mathcal{L}_{MD}$	0.8690	0.6093	0.2342

performance across settings.

D Ensemble Optimisation Paradigms on ConvAbuse and HS-Brexit

Tables12 and 13 present ablation results for ConvAbuse and HS-Brexit using WEL-Random. Across both datasets, multiple configurations achieve similar scores, suggesting that when loss coefficients are fixed, different objectives can lead to comparable outcomes.

For **ConvAbuse**, combining $\mathcal{L}_{F1}$ with either $\mathcal{L}_{CE}$ or $\mathcal{L}_{MD}$ yields the highest F1 (0.9333), while $\mathcal{L}_{MD}$ alone achieves the lowest MD (0.1680). For **HS-Brexit**, $\mathcal{L}_{CE}$ alone gives the highest F1 (0.9048), and the $\mathcal{L}_{CE}$+$\mathcal{L}_{MD}$ pairing yields the lowest MD (0.2306). Including all three losses does not consistently improve results and can slightly reduce F1, likely due to competing objectives without tuned coefficients.

Overall, these results indicate that when coefficients are fixed, several loss configurations can perform similarly, and gains from specific combinations are modest. The impact of loss balancing is explored further in the next subsection on parameter correlations.

E Parameter Impact on ConvAbuse and HS-Brexit

Table 14 illustrates how each control parameter balances the multi-objective trade-offs in the joint optimisation (Eq. (8)). The regularisation term λ demonstrates consistently strong performance on both ConvAbuse and HS-Brexit, achieving near-perfect correlations with F1 (+1.0*/+0.99*) and CE (-1.0*), though its effect on MD diverges from patterns observed on MD-Agreement (Table 7).

In contrast, γ consistently improves MD (-1.0*) but harms F1 (-0.9*/-1.0*), making it better suited for MD-focused objectives. The effect of α varies: it improves CE and MD on ConvAbuse but degrades them on HS-Brexit. Finally, β reliably improves CE (-1.0*) on both datasets, but at the cost of worse MD (+0.7/+0.9*). These differences likely reflect the distinct text genres and annotation distributions of the datasets, underscoring the need for task-specific parameter tuning.

These results align closely with MD-Agreement findings in the main paper.

Table 14: Correlation between parameter and evaluation metrics (F1, CE and MD) on the ConvAbuse and HS-Brexit datasets using WEL-Random with BERT.

Dataset	ConvAbuse			HS-Brexit		
Param	F1	CE	MD	F1	CE	MD
α	-0.21	-0.4	-1.0*	+0.82	+1.0*	+0.1
β	+0.8	-1.0*	+0.7	+0.6	-1.0*	+0.9*
γ	-0.9*	+1.0*	-1.0*	-1.0*	+1.0*	-1.0*
λ	+1.0*	-1.0*	+0.9*	+1.0*	-1.0*	+1.0*
M	+1.0*	-1.0*	+0.98*	+0.99*	-1.0*	+0.55

Aligning NLP Models with Target Population Perspectives using PAIR: Population-Aligned Instance Replication

Stephanie Eckman[*,♣] **Bolei Ma**[*,♡] **Christoph Kern**[♡] **Rob Chew**[♠]
Barbara Plank[♡] **Frauke Kreuter**[♣,♡]

[♣]University of Maryland, College Park
[♡]LMU Munich & Munich Center for Machine Learning
[♠]RTI International

[*]Equal contributions.

{steph, fkreuter}@umd.edu, {bolei.ma, christoph.kern, b.plank}@lmu.de, rchew@rti.org

Abstract

Models trained on crowdsourced annotations may not reflect population views, if those who work as annotators do not represent the broader population. In this paper, we propose **PAIR**: **P**opulation-**A**ligned **I**nstance **R**eplication, a post-processing method that adjusts training data to better reflect target population characteristics without collecting additional annotations. Using simulation studies on offensive language and hate speech detection with varying annotator compositions, we show that non-representative pools degrade model calibration while leaving accuracy largely unchanged. PAIR corrects these calibration problems by replicating annotations from underrepresented annotator groups to match population proportions. We conclude with recommendations for improving the representativity of training data and model performance.[1]

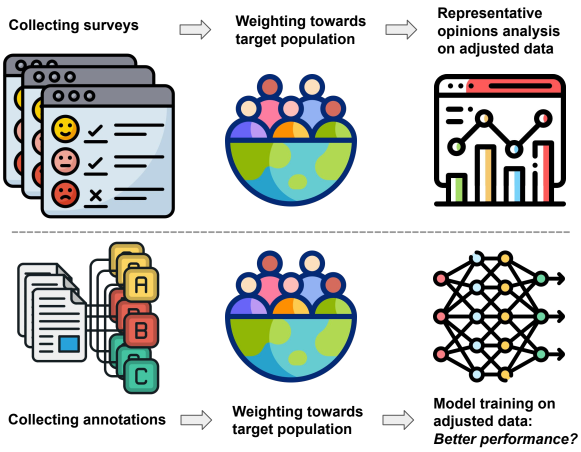


Figure 1: **Top**: Adjusting survey data to match population produces high quality results.
Bottom: Can a similar adjustment in data annotations also improve model performance?

1 Introduction and Inspiration

When a hate speech detection model flags harmless expressions as toxic, or a content moderation system fails to identify genuinely harmful content, the root cause often lies not in the model architecture, but in who annotated the training data. While Natural Language Processing (NLP) models aim to serve broad populations, the human judgments used to train these systems often come from crowdworkers and convenience samples. And the demographics, cultural contexts, and worldviews of these annotators often differ from those of the communities the models ultimately impact (Sorensen et al., 2024; Fleisig et al., 2024). These non-representative annotator pools can have real consequences, because annotator characteristics like age, education level, and cultural background impact how content is annotated (Sap et al., 2022; Fleisig et al., 2023; Kirk et al., 2024). The influence of annotator characteristics underscores that language understanding is not a single objective truth but a constellation of equally valid interpretations anchored in different lived experiences. When this perspectivist interpretation is ignored, models trained on non-representative data can perpetuate the biases and blind spots of their limited training data (Berinsky et al., 2012; Hebert-Johnson et al., 2018; Mehrabi et al., 2021; Rolf et al., 2021; Hüllermeier and Waegeman, 2021; Ouyang et al., 2022; Favier et al., 2023; Smart et al., 2024).

Fortunately, survey researchers have developed statistical techniques to produce population-level estimates from non-representative samples (Bethlehem et al., 2011). The top panel of Figure 1 shows a simple survey workflow: collecting survey data, creating statistical weights to match the sample to the population, and estimating population parameters. We adapt this approach to the machine learning context, enabling models to better align

[1]The code for experiments is available at https://github.com/soda-lmu/PAIR.

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 100–110
November 8, 2025

with target populations even when trained on non-representative annotator pools (bottom panel).

Our Population-Aligned Instance Replication (PAIR) method post-processes training data to better reflect target populations without collecting additional annotations. We test the approach with a simulation study (Burton et al., 2006; Valliant, 2019; Morris et al., 2019) and answer two questions:

- **RQ1:** How do non-representative annotator pools impact model calibration and accuracy?
- **RQ2:** Can our proposed weighting method (PAIR) mitigate these annotator pool effects?

Our results demonstrate that models trained on non-representative annotator pools perform worse than those trained on representative pools. However, simple adjustment methods can improve performance without collecting additional data. These findings suggest that insights from survey methodology can help artificial intelligence (AI) systems better represent the populations they serve.

2 Related Work

Several strands of related work inform our approach to identifying and mitigating bias due to the use of non-representative annotators:

Annotator Impact on Data and Models. Annotator characteristics and attitudes significantly influence annotation quality, particularly for subjective tasks like toxicity detection (Giorgi et al., 2025; Prabhakaran et al., 2021; Fleisig et al., 2023; Sap et al., 2022). For example, annotators' political views and racial attitudes affect their toxicity judgments (Sap et al., 2022). Models trained on non-representative annotator pools inherit these biases and generalize poorly (Berinsky et al., 2012; Mehrabi et al., 2021; Rolf et al., 2021; Ouyang et al., 2022; Favier et al., 2023; Smart et al., 2024; Mokhberian et al., 2024).

Annotator Demographics. Several researchers advocate collecting annotator demographics to assess representation and identify biases (Bender and Friedman, 2018; Prabhakaran et al., 2021; Plank, 2022; Wan et al., 2023; Santy et al., 2023; Pei and Jurgens, 2023).[2] However, collecting and releasing these data can raise privacy concerns (Fleisig et al., 2023). Recent works have also used demographics to prompt the large language models (Argyle et al., 2023), and some find that these are less effective in subjective contexts (Sun et al., 2025; Orlikowski et al., 2025).

Debiasing & Data Augmentation Methods. Prior work has proposed various approaches to reduce bias in training data features and annotations. Most similar to our work is the resampling and reweighting approaches of Calders et al. (2009) and Kamiran and Calders (2012), imputation (Lowmanstone et al., 2023), and the oversampling of minority class cases of Ling and Li (1998). PAIR adapts these methods to balance *annotator* characteristics rather than class labels or sensitive observation-level features. PAIR retains the simplicity and interpretability of earlier resampling methods while extending them to a "Learning with Disagreement" (Uma et al., 2021; Leonardelli et al., 2023) setting with multiple annotations per observation, by replicating annotations from underrepresented annotator groups.

3 PAIR Algorithm: Adjustment via Pseudo-Population

To adjust an annotator pool to better reflect a target population, we propose the PAIR algorithm, which constructs a pseudo-population through post-stratification, weight normalization, and deterministic replication. This adjustment strategy is inspired by established methods in survey sampling (Quatember, 2015).

Post-stratification aligns a sample more closely with population-level distributions (Bethlehem et al., 2011; Valliant et al., 2013). Annotators are grouped into strata based on demographic or behavioral characteristics. For each unit i in stratum s, a post-stratification weight is computed as:

$$w_{s,i} = \frac{P_s}{S_s} \tag{1}$$

where P_s and S_s denote the share of the population in stratum s and the share of the sample (or annotator pool), respectively. The P values come from official statistics or surveys. The S values likely come from the annotators themselves and researchers may have to collect them. This technique can accommodate multiple stratification variables; it is only limited by the availability of population or reference data and data about the annotators.

[2] In our context, these characteristics are used only to analyze bias. Because they are not available for unannotated text, they are not features that the model can use.

These weights have only relative meaning and are invariant to multiplication by a constant (K):

$$w_i^{\text{normalized}} = w_i^{\text{initial}} \times K \tag{2}$$

Normalization useful if research teams have a target number of annotations per observation in mind, for either computational or design reasons, or if some weights given by Eq. 1 are very small and round to one.

To generate a pseudo-population, we apply deterministic replication: each unit is replicated n_i times where

$$n_i = \text{round}(w_i^{\text{normalized}}) - 1 \tag{3}$$

ensuring integer replication counts. This approach produces a dataset that reflects population proportions while maintaining interpretability and reproducibility.

While we focus in this initial study on deterministic replication, alternative implementations are possible, including resampling-based replication or direct incorporation of weights into model training.

4 Annotation Simulation and Model Training

To address our research questions, we conduct a simulation study on offensive language and hate speech detection. We imagine a population made up of equal shares of two types of people: those more likely to perceive offensive language and hate speech and those less likely. We create three datasets of simulated annotations which differ in the mix of the annotator types. We then create a fourth dataset, using the PAIR algorithm, to fix the imbalance in the annotators. We fine-tune RoBERTa models on the four datasets and evaluate the effect of annotator composition on model performance (RQ1) and the ability of the PAIR algorithm to improve performance (RQ2).

4.1 Simulating Annotations

We use our previously collected dataset on tweet annotation sensitivity (Kern et al., 2023)[3], which is a dataset of 3,000 English-language tweets, each with 15 annotations of both offensive language (OL: yes/no) and hate speech (HS: yes/no). We chose this dataset because the high number of annotations of each tweet gives us a diverse set of labels to work with. We randomly select (without replacement) 12 annotations (of both OL and HS) of each tweet in the original dataset.[4] Let $p_{i,OL}$ be the proportion of the 12 annotators who annotated tweet i as OL and $p_{i,HS}$ defined similarly. Figure 2 shows the distribution of these proportions across the 3,000 tweets. The HS annotations are clustered near 0, whereas the OL annotations are more spread out between 0 and 1.

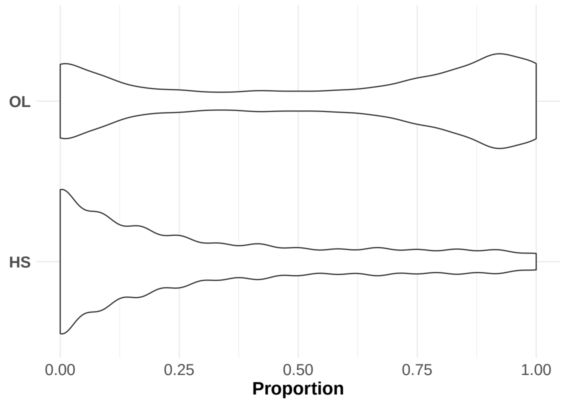


Figure 2: Distribution of $p_{i,OL}$ and $p_{i,HS}$ in original data

The population contains two types of people (50% each). **Type A** people are *less likely* to say a tweet contains OL. **Type B** people are *more likely*:

$$p^A_{i,OL} = \max(p_{i,OL} - \beta, 0) \tag{4}$$
$$p^B_{i,OL} = \min(p_{i,OL} + \beta, 1) \tag{5}$$

Here β captures the magnitude of the bias. We vary β from $[0.05, 0.3]$ by 0.05, corresponding to an increase or decrease in the probability to judge a tweet as OL by five to 30 percentage points. This range is large on the probability scale and covers most reasonable situations. With these six values of β, we create six vectors of probabilities $(p^A_{i,OL}, p^B_{i,OL})$ for each tweet.

We then create four datasets, each with 3,000 tweets (Table 1), for each value of β. The **Representative** Dataset contains OL annotations from six A annotators (drawn from Bernoulli($p^A_{i,OL}$)) and six B annotators (drawn from Bernoulli($p^B_{i,OL}$)). The proportion of A and B annotators in this dataset matches the simulated population we created.

We next create two unbalanced datasets. **Non-representative 1** randomly deletes three B an-

[3] https://huggingface.co/datasets/soda-lmu/tweet-annotation-sensitivity-2

[4] As shown in Table 1, we can more carefully control the construction of our datasets when the number of annotations per tweet is even.

Dataset	Annotations per tweet	A annotations	B annotations
Representative	12	6	6
Non-representative 1	9	6	3
Non-representative 2	12	9	3
Adjusted	12	6	3 + 3*

* 3 B annotations replicated

Table 1: Four training datasets for each bias value (β)

notations for each tweet from the Representative Dataset. **Non-representative 2** adds three additional A annotations, drawn from $p^A_{i,OL}$, to the Non-representative 1 dataset. The Non-representative 2 Dataset is more unbalanced than Non-representative 1, but contains the same number of annotations as the Representative dataset.

4.2 Applying PAIR Algorithm

Finally, we use the PAIR algorithm to create the **Adjusted** Dataset. Starting with the Non-representative 1 Dataset, we calculate the share of the annotator pool that is in the A and B strata: $S_A = \frac{2}{3}, S_B = \frac{1}{3}$. The population proportions, by construction, are $P_A = 0.5, P_B = 0.5$. Applying (1), we get $w_{A,i} = 0.75, w_{B,i} = 1.5$. We multiply these weights by $K = \frac{4}{3}$ to get $w_{A,i} = 1, w_{B,i} = 2$. These weights give us $n_{A,i} = 0, n_{B,i} = 1$, which leads us to replicate all B annotations in the Non-representative 1 Dataset (see Table 1).

The HS probabilities for the A and B annotators are defined in the same way: $p^A_{i,HS} = \max(p_{i,HS} - \beta, 0)$, $p^B_{i,HS} = \min(p_{i,HS} + \beta, 1)$. We also construct the four datasets (Representative, Non-representative 1, Non-representative 2, Adjusted) in the same way we did in the OL case.

Figures 3 and 4 show the percentage of instances annotated OL and HS in the four datasets for each value of β. In both, the percentage of OL/HS annotations in the Adjusted dataset is similar to that in the Representative dataset for all values of β. The percentage in the two unbalanced datasets is lower, because those datasets overrepresent the A annotators, who are less likely to annotate OL/HS.

HS is rare in our dataset (16.7% of instances were annotated as HS), and our simulation strategy overrepresents A annotators in the two Non-representative datasets, who are less likely to perceive HS (Table 1). For these reasons, as β increases, more $p^A_{i,HS}$ are 0 while the $p^B_{i,HS}$ probabilities increase. This issue leads the proportion of HS annotations in the Representative and Adjusted datasets to increase with β in the HS dataset, which have more B annotations than the unadjusted datasets (Figure 4).

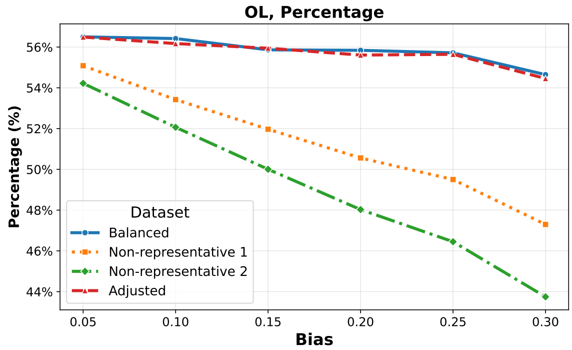


Figure 3: Percentage of instances annotated as OL, by dataset and bias (β)

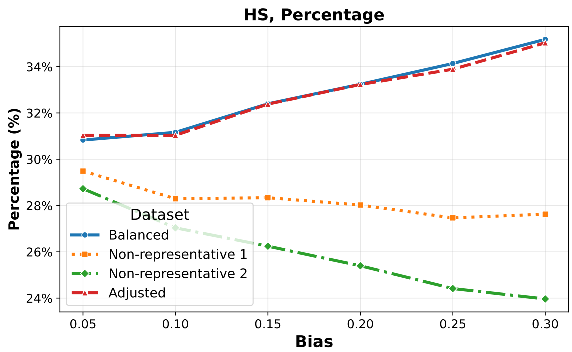


Figure 4: Percentage of instances annotated as HS, by dataset and bias (β)

4.3 Model Training and Evaluation

Training and Test Setup. We train models on each of the eight datasets: four for OL, four for HS. We divide each dataset, at the tweet level, into training (2,000 tweets), development (500), and test (500) sets. Each tweet appears 12 times in the Representative, Non-representative 2, and Adjusted datasets and nine times in the Non-representative 1 set.

Model Selection and Training. We used RoBERTa base (Liu et al., 2019) as our text classifier, training for five epochs on each dataset, with development set optimization. To ensure reliable results, we trained five versions with different random seeds and averaged their performance.

Our implementation of RoBERTa models was based on the libraries `pytorch` (Paszke et al., 2019) and `transformers` (Wolf et al., 2020). During training, we used the same hyperparameter settings (Table 2) for the five training conditions to keep these variables consistent for comparison purposes. We trained each model variation with five random seeds $\{10, 42, 512, 1010, 3344\}$ and took the average across the models. All experiments were conducted on an NVIDIA® A100 80 GB RAM GPU.

Hyperparameter	Value
encoder	`roberta-base`
epochs_trained	5
learning_rate	$3e^{-5}$
batch_size	32
warmup_steps	500
optimizer	`AdamW`
max_length	128

Table 2: Hyperparameter settings of RoBERTa models

Performance Metrics. We evaluate models using **calibration and accuracy metrics** on the test set. While accuracy metrics directly measure classification performance, calibration metrics provide crucial insights into model reliability by assessing probability estimate quality – particularly important for high-stakes applications requiring trustworthy confidence measures.

For **calibration**, we report Absolute Calibration Bias (ACB, Equation 6), which measures how well a model's predicted probabilities align with true annotation frequencies. For each tweet i, we compare the model's predicted probability of offensive language ($\text{preds}_{i,OL}$) against the true proportion of annotators who labeled that tweet as offensive ($p_{i,OL}$).

$$\text{ACB}_{OL} = \frac{1}{n}\sum_{i=1}^{n}\left|\text{preds}_{i,OL} - p_{i,OL}\right| \quad (6)$$

ACB_{HS} is defined accordingly. ACB adapts established calibration metrics by using the annotator agreement proportion as a plug-in estimator for the true probability, avoiding the need for binning (as in ECE, Naeini et al., 2015) while maintaining the intuitive L1 distance interpretation (Roelofs et al., 2022). A low ACB score indicates that the model's confidence scores accurately reflect the underlying annotation uncertainty in the population.

For **accuracy**, we report the F1 score.

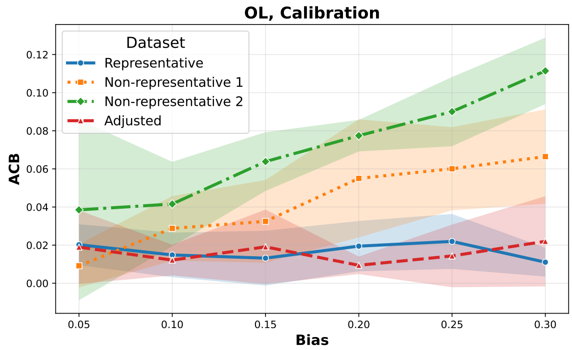


Figure 5: ACB scores for OL Models, by dataset and bias (β)

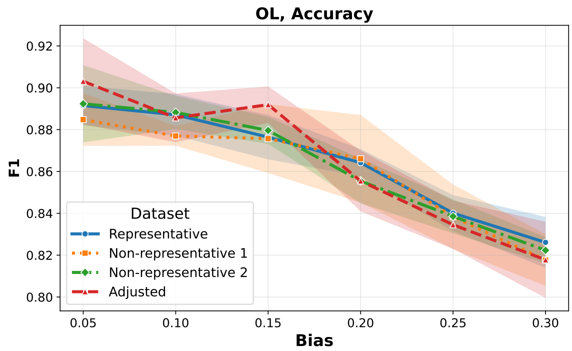


Figure 6: F1 scores for OL Models, by dataset and bias (β)

5 Results

We show results separately for the OL and HS models.

5.1 OL Models

Calibration. Figure 5 compares the ACB in the test set for models trained on the four datasets. The dark lines show average ACB across the five training runs and the shading shows the standard deviation.

The ACB for the models trained on the Adjusted dataset closely tracks that for the Representative dataset and does not increase with β. ACB for the models trained on the two unbalanced datasets is greater and grows with β. These results demonstrate the effectiveness of our adjustment method. Replicating the annotations from the underrepresented annotator type to match population proportions improves model calibration.

Accuracy. Figure 6 compares the models' F1 scores. In contrast to Figure 5, we do not see strong differences between the models trained on the different datasets. For all datasets, model performance

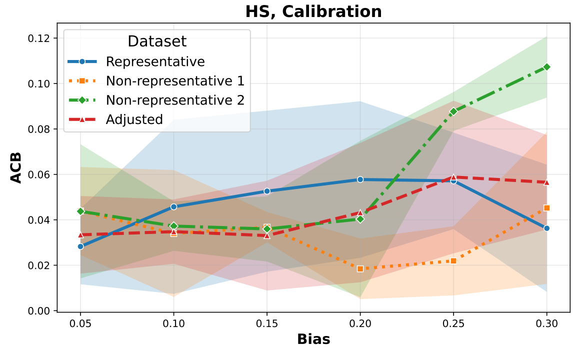


Figure 7: ACB scores for HS Models, by dataset and bias (β)

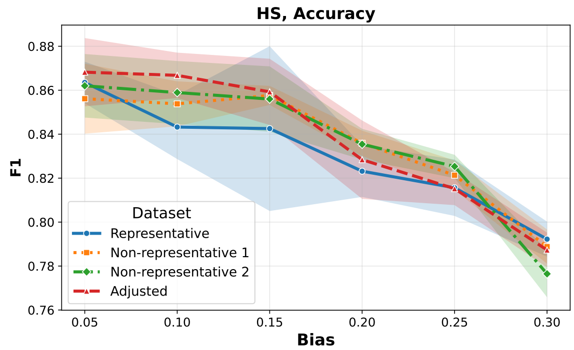


Figure 8: F1 scores for HS Models, by dataset and bias (β)

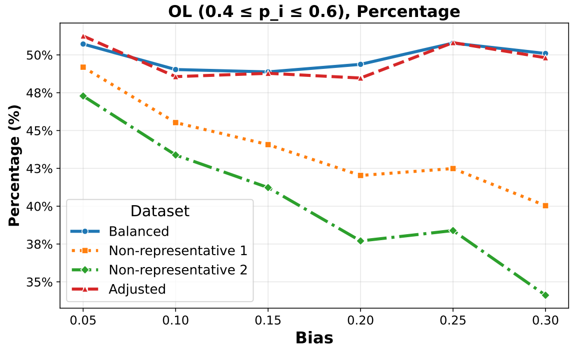


Figure 9: Percentage of OL instances on difficult tweets ($0.4 \leq p_{i,OL} \leq 0.6$) by dataset and bias ($\beta$)

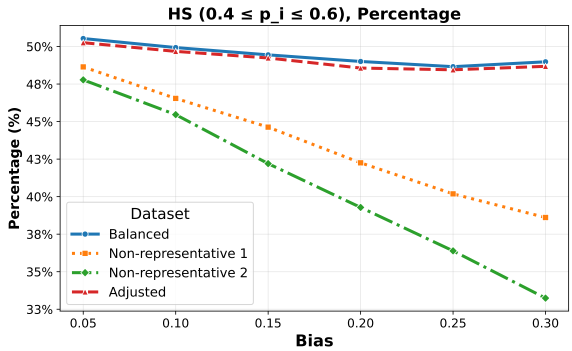


Figure 10: Percentage of HS instances on difficult tweets ($0.4 \leq p_{i,HS} \leq 0.6$) by dataset and bias ($\beta$)

declines with β: as the amount of bias in the annotations increases, the models are less able to predict the binary OL label.

Because the F1 metric focuses on binary predictions, it is less sensitive to training biases than calibration metrics like ACB, which more explicitly capture biases through prediction scores. In decision-making, miscalibrated predictions can have harmful consequences when, for example, hateful content remains undetected (Van Calster et al., 2019). These findings suggest that calibration metrics provide a clearer view of the impact of annotators on models: binary classification metrics can obscure such effects.

5.2 HS Models

Figure 7 contains the ACB results and Figure 8 the F1 score results for the HS models trained on each dataset. Though the adjusted model roughly tracks the representative models for ACB, there is instability in the results. All models show lower average ACB values than the representative model across a wide range of the bias offset (0.10 - 0.20). The PAIR approach does not improve calibration or accuracy: the adjusted model performs similarly to the Non-representative models. This effect is likely due to the combination of label rarity and our simulation design. With few positive annotations to begin with, the impact of the β parameter and the overrepresentation of the A annotators may be overwhelmed by the baseline scarcity of hate speech annotations. Calibration metrics can be less reliable with rare classes (Zhong et al., 2021).

5.3 Sensitivity Analysis: Difficult Tweets

Our simulations assumed that all annotator type impacts all tweets the same way (Eq. 5), which is an oversimplification. More likely, **annotator characteristics have more impact for ambiguous tweets**. For example, prior research in the psychology literature on judgment under uncertainty suggests that people draw more heavily on personal heuristics when interpreting unclear or underspecified information (Tversky and Kahneman, 1974). For this reason, we repeat model training and recompute metrics for those tweets where $0.4 \leq p_i \leq 0.6$. Subsetting the tweets in this way also eliminates the floor and ceiling effects in Eq. 5. The filtered

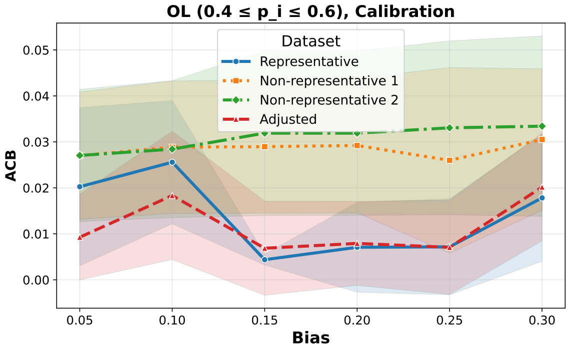


Figure 11: ACB scores for OL Models, on difficult tweets ($0.4 \leq p_{i,OL} \leq 0.6$), by dataset and bias ($\beta$)

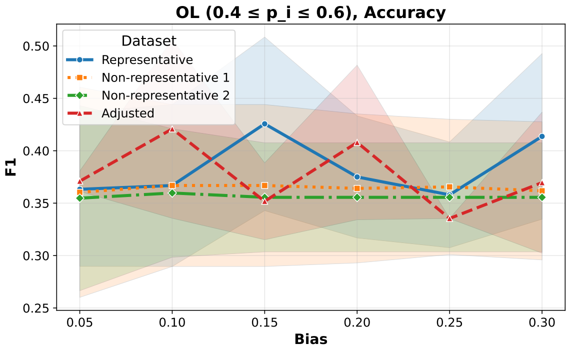


Figure 12: F1 scores for OL Models, on difficult tweets ($0.4 \leq p_{i,OL} \leq 0.6$), by dataset and bias ($\beta$)

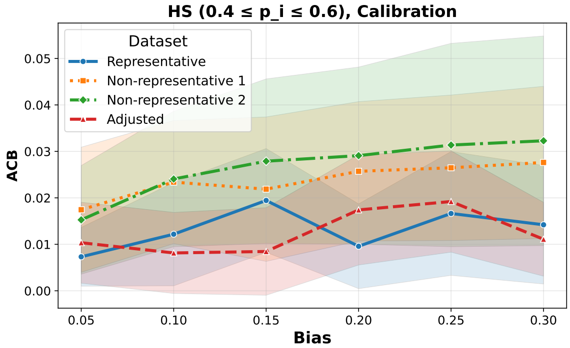


Figure 13: ACB scores for HS Models, on difficult tweets ($0.4 \leq p_{i,HS} \leq 0.6$), by dataset and bias ($\beta$)

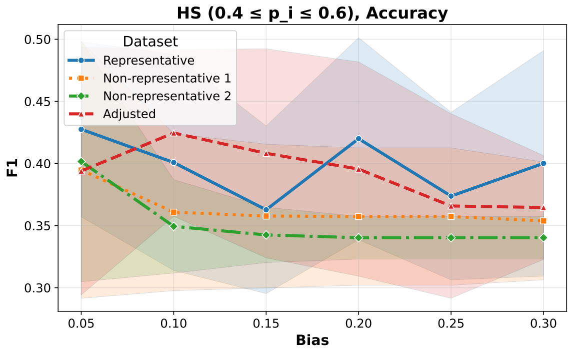


Figure 14: F1 scores for HS Models, on difficult tweets ($0.4 \leq p_{i,HS} \leq 0.6$), by dataset and bias ($\beta$)

datasets contain 267 (OL) and 360 (HS) tweets. The proportions of OL and HS annotations are stable for the Representative and Adjusted sets and decrease for the Non-representative sets as we increase the bias offset (Figures 9 and 10). This mimics the trend for OL in the full dataset (Figure 3).

Figures 11, 12, 13, and 14 show results for two metrics (ACB, F1) for filtered OL and HS annotations. In Figure 11, the Representative and Adjusted models have similar ACB and are lower than the Non-representative models. The F1 scores do not show differences between the models. These results are similar to those on the full set of tweets (Figures 5 and 6). In the two HS figures (13, 14), we see signs that the Representative and Adjusted models perform similarly, and better than the two Non-representative models, on both metrics. These results are more promising than those on the full set of tweets (Figures 7 and 8) and support our hypothesis that the rarity of HS annotations contributed to the lack of positive results for the PAIR approach in §5.2. The PAIR algorithm works well with difficult tweets, which is where it is likely most needed.

6 Discussion & Recommendations

Our experimental results show the OL prediction models perform less well when trained on data from non-representative annotator pools (RQ1), and simple statistical adjustments can improve model calibration without collecting additional annotations or involving additional annotators (RQ2). While PAIR's impact was harder to assess for the rare HS class, PAIR did improve calibration of both the OL and HS models when trained on difficult tweets. These findings establish a promising bridge between survey statistics and machine learning – offering a practical approach to make AI systems more representative of and responsive to the populations they serve, particularly for tasks involving subjective human judgments.

We recommend the following four steps to reduce bias due to non-representative annotator pools:

1) **Use social science research** to identify the annotator characteristics that influence the propensity to engage in annotation and the annotations provided (Eckman et al., 2024).

2) **Collect these characteristics** from annotators and gather corresponding population-level data from national censuses or high-quality surveys.[5]

3) **Calculate weights** that match the annotators to the population on those characteristics (Bethlehem et al., 2011; Valliant et al., 2013).

4) **Use these weights in model training**. Our simple replication approach showed promise, and future work should test more sophisticated weighting approaches.

7 Limitations

Our study explores bias-aware data simulation and evaluation in a controlled setting, which necessarily involves simplifying assumptions and methodological constraints. We outline key areas where future work could broaden the applicability and robustness of our findings.

Stylized Biases and Simulated Data. Our simulation makes strong assumptions about annotator behavior: there are only two types of annotators, and, within each type, annotators behave similarly. Real-world annotator biases may be more nuanced or context-dependent. The simulated annotators might not be representative of a stable opinion group (Mokhberian et al., 2024; Vitsakis et al., 2024). Future work could incorporate more realistic biases and refine the proposed simulations and statistical techniques.

Sampling Variability. We have created only one version of the four datasets for each annotation type and value of β, each of which contains random draws from the Bernoulli distribution. A more traditional statistical approach would create multiple versions of the datasets and train models on each one, to average over the sampling variability. Though limited by computational constraints in this work, future work could take on a more expansive simulation. As discussed, we used five seeds in model training.

Need for Population Benchmarks and Annotator Characteristics. PAIR requires high quality benchmark information about the relevant population. These benchmarks might come from national statistical offices or national surveys. Annotators must provide accurate data on the same characteristics available in the benchmark data. Unfortunately, annotators sometimes do not provide accurate information (Chandler and Paolacci, 2017; Huang et al., 2023). In addition, theory demonstrates that bias will be reduced only when the characteristics used in weighting correlate with the annotations (Eckman et al., 2024). In our simulation, differences in annotations were driven solely by group membership (A, B). In the real world, it is challenging to know what characteristics impact annotation behavior for a given task and to find good benchmarks for those characteristics.

Generalization Beyond Task Types. The study focuses only on binary classification tasks. Many real-world annotation tasks involve multiple classes or labels, which may show different bias patterns. Additional research is needed to extend these methods to more complex classification scenarios.

Evaluation Metrics. While we measured calibration and accuracy, we did not examine other important metrics such as fairness across subgroups or robustness to adversarial examples. Future work on training data adjustment should assess a broader range of performance measures.

8 Ethical Considerations

In this simulation study, we experiment on a publicly available dataset collected in our previous study (Kern et al., 2023), which contains offensive and hateful tweets. We do not support the views expressed in these tweets. The simulation study itself does not collect any new data or raise any ethical considerations.

Acknowledgments

We acknowledge use of the Claude model to edit the text of the paper and to assist in coding. We thank the members of SODA Lab and MaiNLP labs from LMU Munich, and the members of the Social Data Science Group from University of Mannheim for their constructive feedback. This research is partially supported by RTI International, MCML, and BERD@NFDI. BP is supported by ERC Consolidator Grant DIALECT (101043235).

[5] Collection and release of annotator characteristics or weights derived from them may raise confidentiality concerns. The survey literature offers advice for sharing sensitive data (see Karr, 2016, for a review). Collecting annotator characteristics may also require involvement of Institutional Review Boards or other participant protection organizations (Kaushik et al., 2024).

References

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):339–355.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Adam J. Berinsky, Gregory A. Huber, and Gabriel S. Lenz. 2012. Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*, 20(3):351–368.

Jelke Bethlehem, Fannie Cobben, and Barry Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Wiley.

Andrea Burton, Douglas G. Altman, Patrick Royston, and Roger L. Holder. 2006. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25(24):4279–4292.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *ICDMW '09: Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, pages 13–18.

Jesse J. Chandler and Gabriele Paolacci. 2017. Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, 8(5):500–508.

Stephanie Eckman, Barbara Plank, and Frauke Kreuter. 2024. Position: Insights from survey methodology can improve training data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12268–12283. PMLR.

Marco Favier, Toon Calders, Sam Pinxteren, and Jonathan Meyer. 2023. How to be fair? a study of label and selection bias. *Machine Learning*, 112(12):5081–5104.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. 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 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2025. Human and llm biases in hate speech annotations: A socio-demographic analysis of annotators and targets. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):653–670.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (Computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1939–1948. PMLR.

Olivia Huang, Eve Fleisig, and Dan Klein. 2023. Incorporating worker perspectives into MTurk annotation practices for NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1010–1028, Singapore. Association for Computational Linguistics.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506.

Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

Alan F. Karr. 2016. Data sharing and access. *Annual Review of Statistics and Its Application*, 3(Volume 3, 2016):113–132.

Divyansh Kaushik, Zachary C. Lipton, and Alex John London. 2024. Resolving the human-subjects status of ml's crowdworkers. *Commun. ACM*, 67(5):52–59.

Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma, and Frauke Kreuter. 2023. Annotation sensitivity: Training data collection methods affect model performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14874–14886, Singapore. Association for Computational Linguistics.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 105236–105344. Curran Associates, Inc.

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.

Charles X. Ling and Chenghui Li. 1998. Data mining for direct marketing: problems and solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, KDD'98, page 73–79. AAAI Press.

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

London Lowmanstone, Ruyuan Wan, Risako Owan, Jaehyung Kim, and Dongyeop Kang. 2023. Annotation imputation to individualize predictions: Initial studies on distribution dynamics and model predictions. In *NLPerspectives@ECAI*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–36.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, Mexico City, Mexico. Association for Computational Linguistics.

Tim P. Morris, Ian R. White, and Michael J. Crowther. 2019. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2092–2111, Vienna, Austria. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the POPQUORN dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 252–265, Toronto, Canada. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andreas Quatember. 2015. *Pseudo-Populations: A Basic Concept in Statistical Surveys*. Springer.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C. Mozer. 2022. Mitigating bias in calibration error estimation. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 4036–4054. PMLR.

Esther Rolf, Theodora T Worledge, Benjamin Recht, and Michael Jordan. 2021. Representation matters: Assessing the importance of subgroup allocations in training data. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9040–9051. PMLR.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Andrew Smart, Ding Wang, Ellis Monk, Mark Díaz, Atoosa Kasirzadeh, Erin Van Liemt, and Sonja

Schmer-Galunder. 2024. Discipline and label: A weird genealogy and social theory of data annotation. *Preprint*, arXiv:2402.06811.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46280–46302. PMLR.

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

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

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

Richard Valliant. 2019. Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2):231–263.

Richard Valliant, Jill A Dever, and Frauke Kreuter. 2013. *Practical tools for designing and weighting survey samples*, volume 1. Springer.

Ben Van Calster, David J. McLernon, Maarten van Smeden, Laure Wynants, Ewout W. Steyerberg, Patrick Bossuyt, Gary S. Collins, Petra Macaskill, David J. McLernon, Karel G. M. Moons, Ewout W. Steyerberg, Ben Van Calster, Maarten van Smeden, and Andrew J. Vickers. 2019. Calibration: the achilles heel of predictive analytics. *BMC Medicine*, 17(1):230.

Nikolas Vitsakis, Amit Parekh, and Ioannis Konstas. 2024. Voices in a crowd: Searching for clusters of unique perspectives. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12517–12539, Miami, Florida, USA. Association for Computational Linguistics.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14523–14530.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. 2021. Improving calibration for long-tailed recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16484–16493.

Hypernetworks for Perspectivist Adaptation

Daniil Ignatev[1], Denis Paperno[1], Massimo Poesio[1,2],

[1]Utrecht University, [2]Queen Mary University of London,

Correspondence: d.ignatev@uu.nl

Abstract

The task of perspective-aware classification introduces a bottleneck in terms of parametric efficiency that did not get enough recognition in existing studies. In this article, we aim to address this issue by applying an existing architecture, the `hypernetwork+adapters` combination, to perspectivist classification. Ultimately, we arrive at a solution that can compete with specialized models in adopting user perspectives on hate speech and toxicity detection, while also making use of considerably fewer parameters. Our solution is architecture-agnostic and can be applied to a wide range of base models out of the box.

1 Introduction

In the recent years, perspective-aware approach to subjective linguistic tasks has been gaining prominence in NLP. This approach suggests that for tasks that involve subjectivity, dataset designers should collect multiple labels from different annotators for each data instance; these labels need to be retained and used in model training (Plank, 2022; Cabitza et al., 2023; Fleisig et al., 2024). This policy is warranted in tasks like hate speech detection, where multiple labels assigned by annotators with diverse backgrounds can be equally applicable. This notion contrasts the popular practice of designating a single supposedly correct label for each bit of data while discarding conflicting annotator judgments.

Likewise, NLP researchers have argued that subjective tasks require perspective-aware machine learning methods, i.e., methods that can capture diverse opinions based on unaggregated labels (Akhtar et al., 2021). We are particularly interested in the paradigm of *strong perspectivism* (Cabitza et al., 2023). The latter suggests that much like model personalization, separate models or representations should be trained on labels produced by individual annotators. In practice, this entails one of the following: (1) training a full language model checkpoint on each annotator's labels; (2) when using PEFT methods (Houlsby et al., 2019), training a separate adaptation component for each perspective; (3) using a specialized language model architecture.[1] We argue that both the first and the second strategy require a prohibitive amount of trainable parameters that increases with model size and number of modeling targets (see discussion in Section 6). To tackle this problem, we make use of an architecture that employs a small trainable module and adapts a model to diverse perspectives while keeping most parameters frozen.

The core of our solution is the `hypernetwork+adapters` combination. A hypernetwork is a neural architecture in which a source neural network is trained to predict the weights of a target network; these weights are subsequently used in inference (Ha et al., 2017). Importantly, hypernetworks can also be used to predict weights of adapters, including LoRA adapters (Houlsby et al., 2019; Hu et al., 2022), rather than weights of entire models. Adapters can be defined as small trainable modules designed for parameter-efficient tuning of NLP models including large language models; compatibility with adapters makes hypernetworks suitable for the same task (Karimi Mahabadi et al., 2021; He et al., 2022; Phang et al., 2023). We provide more details on this architecture in Section 3. The novel aspect of our work is that we attempt to repurpose the `hypernetwork+adapters` combination for perspectivist modeling and, by extension, model personalization. We posit its strengths, i.e., the reduction of trainable parameters, as a way to address the mentioned bottleneck issue. To our knowledge, this direction has not been thoroughly explored in previous studies.

Our article makes the following con-

[1]Here, we do not consider few-shot and zero-shot learning, as they do not fall under the training category.

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 111–122

tributions. First, we consider whether `hypernetwork+adapters` setting is generally applicable to annotator-aware text classification. Our results suggest that hypernetworks are well fit for this task, albeit not unconditionally.

Second, in our experiments,[2] we show that our hypernetwork-based architecture performs on par with recent perspectivist model architectures — particularly, Annotator-Aware Representations for Texts (Mokhberian et al., 2024), and Annotator Embeddings (Deng et al., 2023). At the same time, we note the limitations of the proposed architecture and attempt to address them.

Third, we demonstrate that the proposed architecture offers a better trade-off in terms of parameter efficiency than the baseline models. At the same time, it also demonstrates a greater degree of versatility: since the weights of the base model are not being affected, there is no risk of forgetting (Kirkpatrick et al., 2017) or degradation of the base model's fundamental capabilities. These properties of our method make us believe that it has promise in modeling subjective tasks.

2 Related work

Data perspectivism: the idea of preserving pluralistic annotations in datasets has been discussed for a considerable time (Artstein and Poesio, 2008) and has been gaining increasing prominence across various AI domains (Kumar et al., 2021; Kapania et al., 2023; Huang et al., 2023). Frenda et al. (2024) provides a detailed survey of such perspectivist datasets and methods, reflecting a paradigm shift towards treating annotator disagreement not as noise but as a source of valuable signal about human variation.

Perspectivist learning: modeling human variation in labeling based on annotator perspectives has been advocated in several recent studies (Plank, 2022; Cabitza et al., 2023). In terms of modeling techniques, various recipes have been explored, including trainable crowd layers (Rodrigues and Pereira, 2018); trainable tokens (Sarumi et al., 2024); multi-task classification with annotator-specific heads (Davani et al., 2022), and others. Recently, studies have focused on active learning (Baumler et al., 2023; Wang and Plank, 2023) and few-shot perspective modeling (Golazizian et al., 2024; Sorensen et al., 2025) as two ways to deal with data sparsity.

Strong perspectivism brings together perspectivist learning and model personalization, which is why recent methods from the latter domain are also relevant to our research. In particular, studies by Tan et al. (2024) and Clarke et al. (2024) show that adapters and LoRA adapters can rival full model finetuning on the LLM personalization task, even when the number of trainable parameters is reduced to a fraction of the original model size.

Finally, in our study, we are particularly interested in recent works that optimize encoder-based classifiers against pluralistic labels and make use of annotator embeddings (Deng et al., 2023; Mokhberian et al., 2024). In this work, we consider both of these as our baselines.

Hypernetworks: The idea of hypernetworks has its roots in an earlier concept of weight generators (Gomez and Schmidhuber, 2005) and aims to constrain the search space when modeling complex objectives through searching in the limited weight space. Practically, this means generating the weights of a target model by means of a separate generator model (hypernetwork).

The title paper, published in 2017 (Ha et al., 2017), had two additional objectives: reduction of trainable parameters (1) and regularized training (2). The study proposed an implementation of two network types: a CNN network generating weights for another CNN network and an RNN network producing weights for a target RNN. Their results showed that the system achieves competitive performance and reduces the trainable parameter count, effectively fulfilling its purpose. The authors' assumptions were further scrutinized in follow-up works. As an example, a study by Soydaner (2020) applied hypernetworks to convolutional autoencoder models and showed that the weights of an autoencoder can be closely approximated by a much smaller hypernetwork, resulting in significant reduction of trainable parameters.

Moreover, the ML community recognized early the potential of hypernetworks in multitask learning. For instance, a study by Tay et al. (2021) adapts a transformer model to a multitude of tasks by predicting task-specific weights for the model's feed-forward layer using a hypernetwork. This adaptation strategy is of particular interest to us due to Davani et al. (2022)'s success in learning perspectives through multi-task classification.

[2] Our codebase has been made publicly accessible at https://github.com/ruthenian8/Hypernets

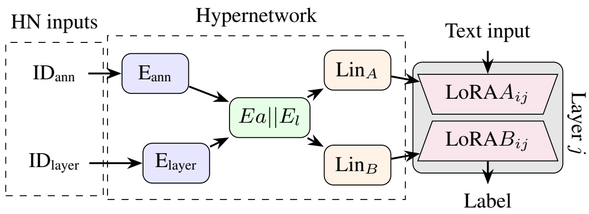


Figure 1: Data flow within our hypernetwork implementation: the annotator id and the layer id are embedded to adapt the target layer l_j in the base model.

The above architectures leveraged hypernetworks to predict the model's own weights. However, the parameter-efficient finetuning paradigm brought about an alternative approach, which is to infer the weights of small trainable submodules while leaving the weights of the main model intact. Adapters are a specific instance of these submodules (Houlsby et al., 2019); modeling them with hypernetworks has led to impactful results in several machine learning applications. Specifically, in image generation, hypernetworks have been used to personalize diffusion models achieving a considerable speed-up compared to other methods (Ruiz et al., 2023). In a different vein, in speech recognition, Müller-Eberstein et al. (2024) employed hypernetworks to adapt ASR models to individual speakers with atypical speech patterns. Their hypernetwork obtains a 75% relative reduction in word error rate using only 0.1% of the model parameters.

In NLP, hypernetworks have been utilized for multi-task and multilingual adaptation of larger transformer models, such as T5 (Raffel et al., 2020). Üstün et al. (2022) propose a single hypernetwork that produces adapter weights for multiple languages and tasks simultaneously, eliminating the need for separate language-specific task adapters. Karimi Mahabadi et al. (2021) and Phang et al. (2023) leverage a shared hypernetwork to effectively train adapter modules for a range of NLP tasks. Finally, in 2025, Charakorn et al. (2025) used hypernetworks to generate task-sepecific LoRA parameters for LLMs on the fly given a textual task description, thus enabling zero-shot adaptation; their results show that hypernetworks are combinable with LLMs, while their approach may also be repurposed for user personalization in the future.

3 Method

3.1 Architecture

The proposed architecture aims to tune a base model to each annotator's perspective; it implements an adapter-modelling hypernetwork in a fashion similar to Phang et al. 2023 and Müller-Eberstein et al. 2024, as it specifically makes use of low-rank adapters (Hu et al., 2022). Generally, all adapter variations enable parameter-efficient finetuning of large models by freezing and patching the base model's pre-trained weights W_j within a layer l_j with a smaller trainable layer Ada_j. Low-rank adapters, in particular, take this reduction of updatable parameters one step further by decomposing Ada_j into two low-rank matrices, A_j and B_j; ultimately, they decrease the parameter count even more with little performance impact. This principle can thus be illustrated with the following formula:

$$W'_j = W_j + B_j A_j \tag{1}$$

where A_j and B_j are low-rank trainable matrices. We find this type of adapters preferable for use in combination with hypernetworks, as it narrows down the solution space for the hypernetwork component.

In our implementation, when modeling the perspective Ann_i, we use the hypernetwork H to predict A_{ij} and B_{ij} for every layer l_j. Hence, we need to condition our hypernetwork on two relevant variables: information on the target annotator Ann_i and on the target layer l_j.

$$A_{ij}, B_{ij} = H(Ann_i, l_j) \tag{2}$$

For demonstration purposes, we only consider unique annotator IDs as annotator information. However, other relevant variables can also be straightforwardly integrated. We leave it to further research to determine whether using more complex representations based on sociodemographic variables or prior annotations shows better efficiency.

Likewise, target layer information is supplied through numeric layer identifiers. In the hypernetwork module, both types of identifiers are embedded, concatenated, and jointly passed to two prediction heads (Lin_A, Lin_B), which then infer the matrices A and B; this flow is illustrated in Figure 1.

We add the hypernetwork component on top of PEFT's architecture-agnostic LoRA implementation (Mangrulkar et al., 2022) with the aim of making our method compatible with a wide range of base models. To mitigate possible generalization issues, we use GeLU activation (Hendrycks and Gimpel, 2016) and dropout probability of 0.25 in the hypernetwork module. We also follow the suggestion of Müller-Eberstein et al. 2024 by initializing Lin_B with zeros and Lin_A with values close to zero, ensuring smooth updates at the initial stages of training.

3.2 Data

In this study, we purposefully pick 4 datasets that permit us to compare our architecture against existing algorithms for perspectivist learning. Two of these datasets, HS-Brexit and MD-Agreement, were included in the shared task on Learning With Disagreements (LeWiDi, Leonardelli et al. 2023). This section gives an overview of all data we used.

The Multi-Domain Agreement dataset ($\mathcal{D}_{MDA}$): This dataset by Leonardelli et al. (2021) addresses the task of offensive language detection. MD-Agreement comprises 9,814 English tweets from three distinct domains: the Black Lives Matter movement, the 2020 Election, and the COVID-19 pandemic. Each tweet was annotated by at least 5 Mechanical Turk workers out of a total pool of 334.

English Perspectivist Irony Corpus ($\mathcal{D}_{EPIC}$): Introduced by Frenda et al. (2023), the corpus consists of 3,000 Post-Reply pairs collected from Twitter and Reddit. The data was sourced from five English-speaking countries: Australia, India, Ireland, the United Kingdom, and the United States. A total of 74 annotators, balanced by gender and nationality, participated in the annotation task, with around 15 raters per nationality. Each annotator labeled approximately 200 instances, resulting in a corpus with 14,172 annotations and a median of 5 annotations per instance.

The Racial Bias Toxicity Detection Corpus ($\mathcal{D}_{RB}$): Sap et al. (2019) investigated the interaction between annotators' biases and their perceptions of toxicity reaching positive conclusions. The authors recruited 819 Amazon Mechanical Turk workers to annotate tweets for two variables: whether a tweet was (a) personally offensive to them and (b) potentially offensive to others. As in Mokhberian et al. 2024, our experiments focus on (a).

Hate Speech Brexit ($\mathcal{D}_{HS\text{-}Brexit}$): Akhtar et al. (2021) had 6 experiment participants label the same set of 1120 tweets related to Brexit. Each tweet was annotated for several categories including Hate Speech, Aggressiveness, Offensiveness, and Stereotype. In our study, we focus on Hate Speech annotations.

Dataset	*Train*	*Dev*	*Test*	*#A*	*#E/#A*
$\mathcal{D}_{MDA}$	$27k$	$13k$	$13k$	334	160
$\mathcal{D}_{EPIC}$	$7.1k$	$3.5k$	$3.5k$	74	191
$\mathcal{D}_{RB}$	$6.1k$	$2.7k$	$2.8k$	819	14
$\mathcal{D}_{HS\text{-}Brexit}$	$3.3k$	$1.6k$	$1.6k$	6	1120

Table 1: Post-split statistics of datasets used in the experiments. *#A* stands for the number of workers; *#E/#A* denotes the mean number of annotated items per worker.

For our experiments, we reproduce the dataset splitting procedure from Mokhberian et al. 2024. Specifically, we split the data into partitions of 50, 25, and 25% (train, dev, and test sets respectively) stratified with respect to item-level disagreements. Items from the dev and test sets annotated by an annotator not present in the training set are merged into the training data. This splitting procedure is repeated using 10 different random seeds. We report the post-preprocessing statistics for the four datasets in Table 1.

4 Experiments

4.1 Baselines

We test our architecture against approaches introduced in Deng et al. 2023 (AE) and Mokhberian et al. 2024 (AART). Both of these build on generic transformer models and handle diverse perspectives by integrating them directly into the architecture. To that end, they make use of specialized annotator representations.

AART: The AART architecture combines text embeddings from pretrained transformer models with learned annotator embeddings. Formally, given a text item x_i and annotator a_j, the combined embedding is computed as:

$$g(x_i, a_j) = e(x_i) + f(a_j)$$

where $e(x_i)$ is the text embedding, and $f(a_j)$ is a learned annotator-specific embedding. This combined embedding is then fed into a common classification head. The model employs a multi-part loss function consisting of cross-entropy loss, L2 regularization on annotator embeddings, and a contrastive loss designed to cluster similar annotator perspectives. We compare our results against AART on $\mathcal{D}_{MDA}$, $\mathcal{D}_{EPIC}$, and $\mathcal{D}_{RB}$.

Annotator Embeddings (AE): This approach explicitly captures annotator-specific biases and annotation tendencies using two types of embeddings: annotator embeddings (E_a) and annotation embeddings (E_n). These embeddings are combined with the original text embeddings via weighted summation:

$$E_{\text{combined}} = E_{\text{[CLS]}} + \alpha_n E_n + \alpha_a E_a$$

where α_n and α_a are learnable weights computed based on the interaction between the sentence embedding and annotator-specific embeddings. The resulting embedding is fed into transformer-based classification models, enhancing their ability to predict annotator-specific labels by explicitly modeling annotator idiosyncrasies. We compare the performance of our model against AE on $\mathcal{D}_{MDA}$ and $\mathcal{D}_{HS\text{-}Brexit}$.

Single-task baseline: Both AART and AE compare their systems against a single-task baseline model. This baseline is a transformer classifier trained on majority-aggregated labels that predicts one label per text and effectively ignores annotator-specific nuances. This baseline mostly serves as a sanity check to evaluate the effectiveness of perspectivist modeling under the assumption that specialized architectures should handle label diversity at least somewhat better than a non-specialized model. However, on data items where most annotators agree with the majority label (e.g., 7 raters out of 10), it can show better results than a perspectivist model due to a much narrower problem space. Such data items constitute a large part of the existing perspectivist datasets, and, as a result of that, specialized models do not always surpass this baseline.

4.2 Setup

In all experiments, we use RoBERTa-base (Liu et al., 2019) as our base model to maintain consistency with the baseline methods, as both studies include reported results for RoBERTa-base among other models. As a part of our solution, all layers except the hypernetwork get explicitly frozen. We model LoRA adapters with a rank of 2 and $\alpha = 32$, keeping adapter dimensionality to a minimum. Although increasing the number of trainable parameters is likely to lead to better results, we explore the lower performance boundary of our method that also offers the best efficiency tradeoff.

We train the hypernetwork for 5 epochs with a batch size of 100 at a learning rate of $1\text{e}{-5}$ making use of Adam optimizer (Kingma and Ba, 2014). We set max. sequence length to 100. For each dataset, we repeat the experiment 10 times with varying random seeds. To implement our method, we use abstractions from PEFT (Mangrulkar et al., 2022) and Transformers (Wolf et al., 2020).

We obtained the above hyperparameter values by running a grid search with varying dropout probabilities (0.0, 0.1, 0.25) and learning rates ($5\text{e}{-5}$, $1\text{e}{-5}$, $5\text{e}{-6}$) using $\mathcal{D}_{EPIC}$'s development set as a reference. Our conclusion is that unlike the baseline solutions without PEFT methods, our architecture shows greater sensitivity to learning rate and dropout probability. In particular, when using low dropout (0 or 0.1) or a higher learning rate ($5\text{e}{-5}$), the model often converges prematurely at a suboptimal point. This yields micro-f1 values of ≈ 50.0 suggesting that the model only learns the majority label due to label imbalance; in contrast, when using our ultimate parameter values (dropout$= 0.25$, lr$= 1\text{e}{-5}$), training proceeds with gradual parameter updates and leads to better generalization (≈ 68.5). This result demonstrates that careful hyperparameter tuning is necessary for our approach.

4.3 Evaluation

Since AART and AE were assessed with two different metric suites, we preserve the original evaluation protocols to allow direct comparison with the published figures. In addition, we track the total number of trainable parameters for each model as described below.

Annotator-level F1 This metric measures the ability of a model to treat every annotator fairly, regardless of how many labels they contributed. For each annotator a_j, we compute the macro-F1 score over all items x_i in the test split by comparing the gold label y_{ij} to the model's prediction $\hat{y}_{ij}$. The *Annotator-level F1* is then the simple mean of these per-annotator F1 scores. Mokhberian et al. 2024 argue that it prevents evaluation biases towards prolific annotators.

Global-level F1 To measure overall predictive quality, we pool all (x_i, a_j) pairs in the test set and compute macro-F1 on this combined set. Unlike the annotator-level metric, this score weighs each prediction equally and inadvertently prioritizes annotators who contributed more labels.

Item-level Disagreement Correlation We quantify how well a model reproduces the true pattern of annotator disagreement on each item. For an item x_i with annotator votes $\{y_{i1}, \ldots, y_{iK}\}$, the gold disagreement is

$$d_i = 1 - \frac{\max_c \left|\{j : y_{ij} = c\}\right|}{K},$$

and the model-predicted disagreement $\hat{d}_i$ is computed analogously from $\{\hat{y}_{ij}\}$. We then report the Pearson correlation $\mathrm{Corr}(\{d_i\}, \{\hat{d}_i\})$, as in AART.

Baseline-specific Metrics AART uses exactly the three metrics above: annotator-level F1, global-level F1, and item-level disagreement correlation. We thus report all three for direct apples-to-apples comparison. AE instead evaluates each annotator's labels by computing (i) global accuracy over all (x_i, a_j) pairs and (ii) global-level F1 as defined above. Because AE does not release per-annotator predictions, we we are unable to use the annotator-level metrics for that baseline.

Trainable Parameters Lastly, we report the approximate number of trainable parameters for each model (our hypernetwork+adapters, AART, and AE). For AART and AE, we calculate these numbers based on their public source code, while for our method we report the number directly. Analyzing parameter counts helps us inquire into the parameter efficiency of each solution and assess how well each architecture can scale to the growing number of annotators to model and growing embedding space.

Dataset	*Single-task*	*AART*	*Ours*
	Annotator-level F1		
$\mathcal{D}_{MDA}$	66.80 ± 0.7	69.72 ± 1.1	$\mathbf{70.24 \pm 0.9}$
$\mathcal{D}_{EPIC}$	58.59 ± 1.9	$\mathbf{59.67 \pm 0.9}$	53.16 ± 1.6
$\mathcal{D}_{RB}$	68.61 ± 1.5	71.1 ± 3.2	$\mathbf{73.81 \pm 2.0}$
	Global-level F1		
$\mathcal{D}_{MDA}$	71.99 ± 0.6	77.38 ± 0.4	$\mathbf{78.11 \pm 0.2}$
$\mathcal{D}_{EPIC}$	60.23 ± 1.7	$\mathbf{66.16 \pm 1.4}$	65.11 ± 1.2
$\mathcal{D}_{RB}$	71.97 ± 1.7	$\mathbf{79.96 \pm 1.9}$	76.17 ± 1.8
	Item-level Disagreement Correlations		
$\mathcal{D}_{MDA}$	NA	0.37 ± 0.04	$\mathbf{0.42 \pm 0.02}$
$\mathcal{D}_{EPIC}$	NA	0.20 ± 0.06	$\mathbf{0.28 \pm 0.03}$
$\mathcal{D}_{RB}$	NA	0.54 ± 0.04	$\mathbf{0.66 \pm 0.03}$
	Trainable Parameters		
-	$124.3 * 10^6$	$\approx 124.9 * 10^6$	$\approx 5.6 * 10^6$

Table 2: This table shows the metrics of our model, averaged over 10 runs, against RoBERTa-based baselines from Mokhberian et al. 2024 (cols 1, 2; values as reported). Col. 2 features the configuration $\alpha > 0$; see the original paper for details. Along with average values, we report the standard deviation. The highest values are given in bold.

Dataset	*Single-task*	*AE*	*Ours*
	Global-level Accuracy		
$\mathcal{D}_{MDA}$	75.65	75.14	$\mathbf{80.11 \pm 0.2}$
$\mathcal{D}_{HS\text{-}Brexit}$	86.77	**87.03**	86.49 ± 0.8
	Global-level F1		
$\mathcal{D}_{MDA}$	73.26	73.60	$\mathbf{78.11 \pm 0.2}$
$\mathcal{D}_{HS\text{-}Brexit}$	**64.60**	60.36	58.30 ± 9.8
	Trainable Parameters		
-	$124.3 * 10^6$	$\approx 125.5 * 10^6$	$\approx 5.6 * 10^6$

Table 3: This table shows the metrics of our model, averaged over 10 runs, against RoBERTa-based baselines from Deng et al. 2023 (cols 1, 2; values as reported; std not reported). Col. 2 features the configuration E_a; see the original paper for details. The highest values are given in bold.

5 Results

In our experiments, the proposed hypernetwork-based architecture demonstrates generally strong performance across all datasets when compared to both AART and AE baselines, while using substantially fewer trainable parameters. Tables 2 and 3 report the mean and standard deviation over ten runs for each metric suite as defined in Section 4.3.

On the $\mathcal{D}_{MDA}$ dataset, our model achieves an annotator-level F1 of 70.24 ± 0.9 and a global-level F1 of 78.11 ± 0.2, thus surpassing both the AART baseline (annotator F1 69.72 ± 1.1, global F1 77.38 ± 0.4) and the AE baseline (global accuracy 75.14, global F1 73.60). This result indicates that our architecture is effective at capturing individual annotator tendencies and producing coherent overall predictions in this scenario.

For $\mathcal{D}_{RB}$, our approach again outperforms AART in annotator-level performance (73.81 ± 2.0 vs. 71.10 ± 3.2) and yields a competitive global-level F1 of 76.17 ± 1.8 (AART: 79.96 ± 1.9), suggesting that our hypernetwork is capable of modeling both base and rare annotator behaviors even when the dataset exhibits varied disagreement patterns.

Interestingly, on the more challenging $\mathcal{D}_{EPIC}$ dataset, our model gives lower annotator-level F1 (53.16 ± 1.6) and global-level F1 (65.11 ± 1.2) compared to AART's 59.67 ± 0.9 and 66.16 ± 1.4, respectively. We believe this shortcoming stems from the higher complexity of data in EPIC,[3] which may require more specialized regularization strategies; still, our model attains a higher item-level disagreement correlation (0.28 ± 0.03 vs. 0.20 ± 0.06), thus showing that these two perspectivist metrics have a potential discrepancy.

On $\mathcal{D}_{HS\text{-}Brexit}$, the comparison with AE yields a global accuracy of 86.49 (AE: 87.03) and a global F1 of 58.30 (AE: 60.36), which is only slightly below the baseline. Given the dataset size, it translates into a difference of $\tilde{8}$ misclassified instances.

Beyond model generalizability, `hypernetwork+adapters` solution demonstrates remarkably better parameter efficiency, as it requires only $\approx 5.6 * 10^6$ trainable parameters compared to $\approx 124.9 * 10^6$ for AART and $\approx 125.5 * 10^6$ for AE. This advantage is especially important within the strong perspectivist paradigm, where an ideal model should be able to scale up to hundreds of perspectives and, at the same time, take up a reasonable amount of memory and disk space.

6 Discussion

Two potentially advantageous aspects of the `hypernetwork+adapters` setting are (1) keeping the base model's weights intact and (2) adapting the model to all targets by training just one component. In what follows, we analyze why these aspects are especially relevant perspectivist learning.

Preserving the base model's original state could be beneficial for two reasons. Concerning inherently multi-purpose models, such as T5 (Raffel et al., 2020), it means that the base model's performance on other tasks will not be affected. This effect is especially important when dealing with large language models: they are often used for a wide range of tasks, and updating their weights can lead to catastrophic forgetting (Kirkpatrick et al., 2017) or degradation of their performance on previously learned tasks. For instance, if a model is fine-tuned for a particular task like hate speech detection, preserving the base model's original state ensures that its performance on other tasks remains intact.

Moreover, when adapting a model that has already been fine-tuned for some task irrespective of perspectives, our method can preserve this original, 'perspective-neutral' model. In a realistic setting of personalized content moderation, this model can be used as a fallback option. For example, if some user's view is not covered by the perspectivist model, the original, perspective-neutral model can be used to provide a fallback response.

Another benefit of modeling adapters and keeping the main classifier frozen is the mitigation of potential negative outcomes associated with perspectivist learning. In particular, when a RoBERTa model that is fully trainable is finetuned for perspectivist labels (as is the case in Deng et al. 2023 & Mokhberian et al. 2024), it is taught to associate the same text with varying labels (one to many). This can cause conflicting gradient steps leading the model away from the optimum. Incorporating annotator features through embeddings is intended to resolve this issue by converting the task back to a one-to-one association. However, it is unclear whether these features receive sufficient weight in the classifier to achieve that. In our framework, there is no discrepancy between inputs and outputs, as the only trainable component is the hypernetwork, which learns a one-to-one correspondence between annotators and their respective adapters. This approach does not expose the base model's weights to controversial targets and thus avoids the associated difficulties.

Dataset	*RoBERTa*	*LoRA*	*Ours*
	Trainable Parameters		
$\mathcal{D}_{MDA}$	$334 * 125 * 10^6$	$334 * 6.6 * 10^5$	$5.6 * 10^6$
$\mathcal{D}_{EPIC}$	$74 * 125 * 10^6$	$74 * 6.6 * 10^5$	$5.4 * 10^6$
$\mathcal{D}_{RB}$	$819 * 125 * 10^6$	$819 * 6.6 * 10^5$	$5.9 * 10^6$
$\mathcal{D}_{HS\text{-}Brexit}$	$6 * 125 * 10^6$	$6 * 6.6 * 10^5$	$5.3 * 10^6$

Table 4: Overview of trainable parameter counts required for fine-tuning a RoBERTa model to each perspective using all parameters, low-rank adaptation, and our hypernetwork respectively ($r = 2, \alpha = 32$).

Using a hypernetwork to learn adapter weights offers an additional advantage. Training a separate LoRA-style adapter for each annotator is possible but involves training a large number of parameters that grows substantially with each added perspective. As shown in Table 4, the costs of adapting RoBERTa-base to every perspective in our datasets are substantial. In extreme cases, such as $\mathcal{D}_{RB}$ with 819 annotators, the required parameters exceed not only the hypernetwork's but also RoBERTa's total parameter count. This results in impractical

[3]For example, as per Casola et al. (2024), gpt-3.5-turbo yields an F1 of 48.1 on EPIC in zero-shot settings; this differs from its zero-shot performance on other language understanding tasks, such as sentiment analysis or natural language inference (F1s of 91.13 and 67.87 respectively, Ye et al. (2023)).

memory and disk space requirements, rendering the separate adapter approach infeasible.

In contrast, our method stands out as the more cost-effective solution in all settings, except for $\mathcal{D}_{\textit{HS-Brexit}}$, which has an exceptionally small number of annotators[4]. Like other approaches that make use of annotator embeddings, adding a new annotator requires $1 \times hdim$ parameters + rescaling the model to a new embedding. Given the small size of the hypernetwork, all of this sums up to just $2 \times hdim$, allowing for very cheap scaling to additional perspectives. This property makes it especially attractive for perspectivist learning.

Conclusion

In this work, we have investigated a `hypernetwork+adapters` architecture for perspectivist learning on subjective classification tasks. Our experiments show that, while this model does not uniformly exceed the latest perspective-aware baselines, it achieves superior performance on several perspectivist datasets, most notably $\mathcal{D}_{\textit{MDA}}$ and $\mathcal{D}_{\textit{RB}}$. It also obtains higher item-level disagreement correlations even when mean F1 is lower, as on $\mathcal{D}_{\textit{EPIC}}$. Notably, our approach requires only about 5.6×10^6 trainable parameters, about 4.5% with respect to over 124×10^6 in the competing methods; thus, it offers a considerable advantage in parameter efficiency.

Taken together, these findings suggest that the `hypernetwork+adapters` design is a promising solution within the strong perspectivist paradigm, even if further work is needed to let it scale equally well to all available tasks and datasets.

Limitations

One important limitation of the proposed hypernetwork architecture is that it takes more time for inference, as it follows a two-stage procedure where it first separately predicts the LoRA weights of each adapted layer and then applies these during classification. This flow leads to both training and evaluation taking longer than in the case of regular classifier models (≈ 0.25 iterations/sec. vs. 4 iterations/sec.). We argue, however, that this shortcoming does not outweigh the advantage in parameter efficiency. The decreased memory consumption allows for more parallel training jobs to be scheduled simultaneously, thus compensating for lower throughput.

A further possible limitation is that we only use annotator IDs as annotator features. This strategy does not permit our system to scale to new perspectives if we need to make predictions for an unseen annotator. We see two ways to address this limitation. First, in this paper, we do not inquire into how well hypernetworks work with sociodemographic features. However, they can still be trivially integrated, possibly mitigating this issue. A further way to tackle this problem could consist in finding an annotator in the existing annotator pool that is most similar to the unseen one and assuming their perspective. Similarity of annotators could be approximated from the said sociodemographic features.

Finally, we acknowledge that our evaluation of the proposed architecture is not exhaustive, and we could overlook some shortcomings of our model. However, we find it sufficient to judge how well it compares to the competitor models. Additionally, we hope to scrutinize it more when more perspectivist models and datasets are released.

Acknowledgments

We thank the anonymous reviewers and the NLP group at the University of Utrecht for their helpful feedback on this work. Generative tools, such as LLAMA-3 (Grattafiori et al., 2024), were used to proof-read this text. This study is funded by NWO through an AINed Fellowship Grant NGF.1607.22.002 and supported by project 'Dealing with Meaning Variation in NLP'.

References

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. *Preprint*, arXiv:2106.15896.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Connor Baumler, Anna Sotnikova, and Hal Daumé. 2023. Which examples should be multiply annotated? active learning when annotators may disagree.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings*

[4]In order to verify that separately-trained LoRA adapters do not significantly outperform our model, we conduct an experiment on $\mathcal{D}_{\textit{HS-Brexit}}$. We report the outcomes in Table 5; they suggest that training separate adapters does not surpass our approach.

of the AAAI Conference on Artificial Intelligence, 37(6):6860–6868.

Silvia Casola, Simona Frenda, Soda Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and 1 others. 2024. Multipico: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021.

Rujikorn Charakorn, Edoardo Cetin, Yujin Tang, and Robert Tjarko Lange. 2025. Text-to-lora: Instant transformer adaption. *arXiv preprint arXiv:2506.06105*.

Christopher Clarke, Yuzhao Heng, Lingjia Tang, and Jason Mars. 2024. Peft-u: Parameter-efficient fine-tuning for user personalization. *arXiv preprint arXiv:2407.18078*.

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.

Naihao Deng, Xinliang Frederick Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. *Preprint*, arXiv:2305.14663.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. 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 2279–2292, Mexico City, Mexico. 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*, pages 1–28.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Preni Golazizian, Ali Omrani, Alireza S. Ziabari, and Morteza Dehghani. 2024. Cost-efficient subjective task annotation and modeling through few-shot annotator adaptation. *ArXiv*, abs/2402.14101:null.

Faustino Gomez and Jürgen Schmidhuber. 2005. Evolving modular fast-weight networks for control. In *Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005*, pages 383–389, Berlin, Heidelberg. Springer Berlin Heidelberg.

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

David Ha, Andrew Dai, and Quoc V. Le. 2017. Hypernetworks. *ICLR*, 1(8):103–121.

Yun He, Steven Zheng, Yi Tay, Jai Gupta, Yu Du, Vamsi Aribandi, Zhe Zhao, YaGuang Li, Zhao Chen, Donald Metzler, and 1 others. 2022. Hyperprompt: Prompt-based task-conditioning of transformers. In *International conference on machine learning*, pages 8678–8690. PMLR.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Olivia Huang, Eve Fleisig, and Dan Klein. 2023. Incorporating worker perspectives into MTurk annotation practices for NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1010–1028, Singapore. Association for Computational Linguistics.

Shivani Kapania, Alex S Taylor, and Ding Wang. 2023. A hunt for the snark: Annotator diversity in data practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 565–576, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Proceedings of the Seventeenth USENIX Conference on Usable Privacy and Security*, SOUPS'21, USA. USENIX Association.

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, 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.

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

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, Mexico City, Mexico. Association for Computational Linguistics.

Max Müller-Eberstein, Dianna Yee, Karren Yang, Gautam Varma Mantena, and Colin Lea. 2024. Hypernetworks for personalizing ASR to atypical speech. *Transactions of the Association for Computational Linguistics*, 12:1182–1196.

Jason Phang, Yi Mao, Pengcheng He, and Weizhu Chen. 2023. Hypertuning: Toward adapting large language models without back-propagation. In *International Conference on Machine Learning*, pages 27854–27875. PMLR.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Filipe Rodrigues and Francisco C. Pereira. 2018. Deep learning from crowds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. 2023. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6527–6536.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

O. O. Sarumi, Béla Neuendorf, Joan Plepi, Lucie Flek, Jörg Schlötterer, and Charles Welch. 2024. Corpus considerations for annotator modeling and scaling. *ArXiv*, abs/2404.02340:null.

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

Derya Soydaner. 2020. Hyper autoencoders. *Neural Processing Letters*, 52(2):1395–1413.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6476–6491.

Yi Tay, Zhe Zhao, Dara Bahri, Donald Metzler, and Da-Cheng Juan. 2021. Hypergrid transformers: Towards a single model for multiple tasks. In *International conference on learning representations*.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, Gertjan van Noord, and Sebastian Ruder. 2022. Hyper-X: A unified hypernetwork for multi-task multilingual transfer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7934–7949, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinpeng Wang and Barbara Plank. 2023. ACTOR: Active learning with annotator-specific classification heads to embrace human label variation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2046–2052, Singapore. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, and 1 others. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv preprint arXiv:2303.10420*.

A Separate LoRA training.

Dataset	*Baseline*	*LoRA*	*Ours*
Global-level Accuracy			
$\mathcal{D}_{\textit{HS-Brexit}}$	86.77	77.50	86.49
Global-level F1			
$\mathcal{D}_{\textit{HS-Brexit}}$	64.60	53.02	58.30

Table 5: This table reports the metrics of our model against the performance of 6 separate LoRA adapters (one per each annotator perspective) on HS-Brexit. The training parameters for the adapters copied those of the main experiment, the exception being an increased learning rate ($5e{-}5$) and an increased epoch count (10). We report the mean results per 10 runs.

SAGE: Steering Dialog Generation with Future-Aware State-Action Augmentation

Yizhe Zhang, **Navdeep Jaitly**
Apple
{yizzhang, ndjaitly}@apple.com

Abstract

Recent advances in large language models have enabled impressive task-oriented applications, yet building emotionally intelligent chatbots for natural, strategic conversations remains challenging. Current approaches often assume a single "ground truth" for emotional responses, overlooking the subjectivity of human emotion. We present a novel perspectivist approach, SAGE, that models multiple perspectives in dialogue generation using latent variables. At its core is the State-Action Chain (SAC), which augments standard fine-tuning with latent variables capturing diverse emotional states and conversational strategies between turns, in a future-looking manner. During inference, these variables are generated before each response, enabling multi-perspective control while preserving natural interactions. We also introduce a self-improvement pipeline combining dialogue tree search, LLM-based reward modeling, and targeted fine-tuning to optimize conversational trajectories. Experiments show improved LLM-based judgments while maintaining strong general LLM performance. The discrete latent variables further enable search-based strategies and open avenues for state-level reinforcement learning in dialogue systems, where learning can occur at the state level rather than the token level.

1 Introduction

Recent advances in large language models (LLMs) have shown impressive capabilities in task-oriented applications like question-answering, mathematics, and coding (Xu et al., 2024; Liu et al., 2024; Hou et al., 2024). However, effective human-AI interaction requires not just problem-solving abilities but also emotional intelligence and interpersonal skills. This has led to emotion-based chatbots like Character AI (Character AI, 2023) and Pi (PI, 2023), which focus on emotional connection and engagement. These models are designed to understand emotions, provide empathy, and build trust through natural conversations, finding applications in education, healthcare, counseling, and gaming (Beale and Creed, 2009; Reilly et al., 1996).

Building effective emotional chatbots presents unique challenges that highlight the limitations of traditional single-perspective approaches in NLP. Unlike task-oriented interactions, emotional conversations require strategic navigation without clear objectives and involve inherent subjectivity—what constitutes an appropriate emotional response varies significantly across individuals, cultures, and contexts. This subjectivity represents a fundamental challenge that aligns with the growing recognition of Human Label Variation (HLV) in NLP tasks (Plank, 2022). Current approaches often aggregate diverse human perspectives into a single "ground truth," potentially overlooking the richness and validity of multiple viewpoints.

The challenges in emotional dialogue systems manifest in two key areas: (1) chatbots need to steer conversations *strategically* by planning ahead—an AI therapist must balance various conversational moves to build trust and facilitate disclosure, while an AI negotiator needs to integrate multiple techniques to achieve desired outcomes. (2) The chatbot should be *proactive* and *interactive* by asking clarifying questions and providing customized information based on real-time feedback. A career counselor might probe into a client's experiences to offer personalized guidance, while a fitness coach might assess energy levels to adjust workout recommendations. These capabilities require modeling multiple valid perspectives rather than assuming a single correct approach.

These capabilities require effective learning mechanisms for handling sparse rewards across multiple conversation turns. While RLHF (Ouyang et al., 2022) and DeepSeek R1 (Guo et al., 2025) have succeeded in complex reasoning tasks (Weng, 2023; Wang et al., 2023; Ahn et al., 2024; Wang

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 123–132

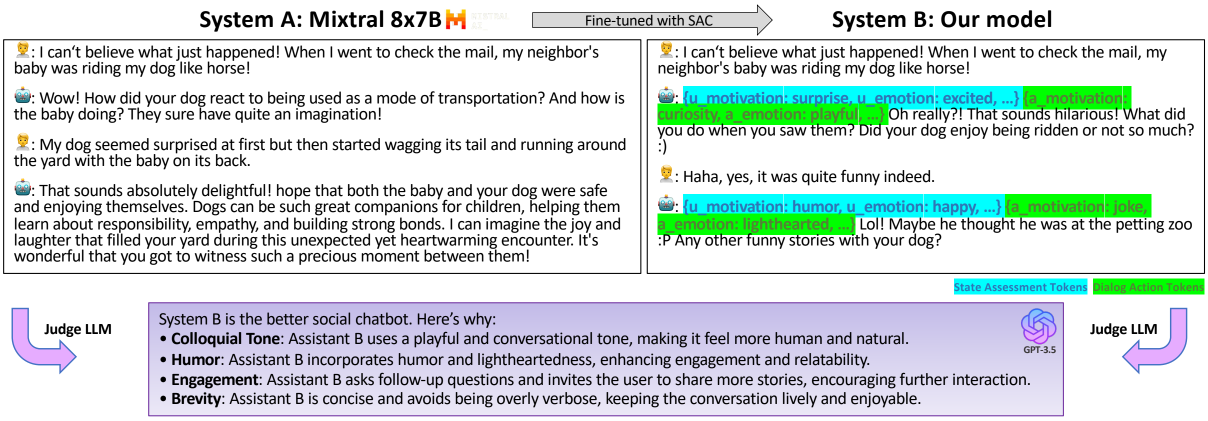


Figure 1: We propose to augment a base-LLM via State-Action Chain (SAC) to provide more control in a multiturn social-oriented dialogue scenario. During inference time, the resulting model first generates additional State Assessment Tokens and Dialog Action Tokens before generating the actual response.

et al., 2024a,b), operating directly on the huge token action space with long horizon remains challenging. Previous work (Chai et al., 2024) has shown that using **macro actions** improves credit assignment and learning efficiency.

We propose SAGE (**S**tate **A**ugmented **GE**neration) that adopts a perspectivist approach to language model fine-tuning by introducing meaningful *latent variables* that capture multiple perspectives in longer-term conversational information. These model-generated variables help balance short- and long-term predictions by explicitly representing diverse dialogue states and actions that encode strategic information beyond immediate token-level generation. By learning to predict these high-level variables first, the model can make more informed decisions about utterance generation that consider both immediate context and long-term conversational goals from multiple valid perspectives.

We introduce the *State-Action Chain* (SAC), which extends chain-of-thought approaches to capture dialogue states' evolution while embracing perspectivist principles. As shown in Figure 1, SAC maintains abstract representations of emotional states and conversational dynamics, enabling coarse-grained control while maintaining natural interaction. This approach separates high-level planning from language generation, making it suitable for reinforcement learning at the state level rather than token level. SAC is a **future-looking annotation strategy**, where state and action labels are generated by considering the complete dialogue context rather than individual utterances in isolation, enabling the model to develop strategic thinking capabilities that can accommodate multiple valid perspectives.

We developed a self-improvement pipeline combining data augmentation, evaluation, tree search, and fine-tuning techniques. This pipeline explores alternative conversational paths, uses rejection sampling based on LLM-derived rewards, and retrains using the most effective trajectories. Our results show improved performance while maintaining capabilities on standard benchmarks. We will release our dataset and model checkpoints [1].

2 Related Work

Decision Transformer The Decision Transformer (DT) (Chen et al., 2021) leverages a transformer architecture to model trajectory data as a sequence of states, actions, and rewards, effectively casting decision-making problems as supervised learning tasks. Recent works have applied DT to diverse applications including gaming (Chen et al., 2021), robotics (Janner et al., 2021) and chip design (Lai et al., 2023). In emotional chatbot contexts, DT's ability to model long-term dependencies could be pivotal for balancing immediate conversational moves with long-term goals like trust-building and engagement. Our work takes the initial step by augmenting utterances with states and actions.

Latent Variable Approaches in Dialogue Generation Several works have explored the use of latent variables to enhance dialogue generation. Serban et al. (2019) introduced a hierarchical latent variable model that captures discourse-level structure in conversations, while Bao et al. (2020) proposed a dialogue generation model with dis-

[1]Code and checkpoints are available upon publication

crete latent variables to model conversation flow and speaker intentions. Our SAC approach differs by focusing specifically on emotional states and conversational strategies, with a future-looking annotation process that considers the complete dialogue context for more accurate state assessment.

Chain-of-Thought Chain-of-thought (CoT) (Wei et al., 2022) has demonstrated remarkable effectiveness in tasks requiring logical and mathematical reasoning. Snell et al. (2024) shows that test-time compute scaling can be more efficient and effective than scaling the model parameters. Following this paradigm, our work incorporates CoT-style reasoning into emotional chatbot interactions by decomposing dialogue generation into a high-level, abstract planning stage that represents the evolution of dialogue states and emotional dynamics, and a language realization stage.

Proactive Dialog Systems Proactive dialog systems anticipate user needs and guide conversations toward desired outcomes using hierarchical structures and reinforcement learning. Examples include mixed-initiative systems for problem-solving and models for strategic customer interactions (Mehri and Eskenazi, 2020). Hong et al. (2023) used synthetic colloquial data and offline RL to improve LLMs in goal-oriented dialogues. In emotion-based chatbots, our approach aligns with the need for high-level guidance, where the system predicts emotional states and motivations to sustain meaningful conversations.

State Augmentation for Task-Oriented Dialogue Task-oriented dialogue systems traditionally rely on modularized stages of language understanding, state-tracking, dialog policy learning and utterance generation. However, advances in neural architectures have enabled more flexible and robust task completion by leveraging contextual embeddings and pre-trained language models (Budzianowski et al., 2018). SOLOIST (Peng et al., 2021) consolidates modular task-oriented dialogue pipelines into a single transformer-based model with state augmentation. Our work also integrates high-level dialogue guidance but additionally incorporates reasoning mechanisms for dialogue actions in emotional chatbots. In contrast to task-oriented systems which search from a finite number of possible states, emotional chatbots focus on open-ended interactions with unbounded state spaces.

3 State Augmented Generation

3.1 Raw Data Preparation

We use our in-house conversational dataset extracted from Reddit spanning the years 2005 to 2017, following the recipe from DialoGPT (Zhang, 2019). We applied aggressive filtering by selecting only conversations with more than four turns and where the average length of each utterance exceeds 15 words. To filter out inappropriate language and tune up the sentiment in the resulting models, sentiment analysis was performed on each utterance using the SENTIMENTINTENSITYANALYZER from NLTK (Bird and Loper, 2004), and we retained only the dialogues where all utterances had a sentiment score above 0.4. Additionally, we filtered the dataset to include only dialogues where at least one utterance ends with a question mark, aiming to encourage the trained model to generate questions more frequently. These filtering steps resulted in a total of 181,388 multi-turn training instances.

3.2 State-Action Chain Augmentation

Instead of relying on the model to generate an utterance through next token prediction alone, we want the model to acquire the following capabilities:

- **State tracking**: Estimate the current dialogue history's state.
- **Policy Learning**: Learn a dialog policy to predict the action based on the current state.
- **Utterance Generation**: Generate an utterance to execute the predicted action.

This approach is comparable to conventional task-oriented chatbot systems that perform goal-oriented tasks like restaurant booking, which employ distinct modules for dialogue state tracking (DST), policy learning, and natural language generation (NLG).

Our goal is to construct an end-to-end data-driven solution for a social chatbot, leveraging the strengths of existing LLMs. Drawing inspiration from the Decision Transformer (Chen et al., 2021) and CoT (Wei et al., 2022), we augment the data such that the state tracking and policy learning processes are explicitly integrated into the standard autoregressive LM learning process. Consequently, the model will produce a series of CoT-style “Thinking” tokens, predicting the current state and the action to be taken, as illustrated in Figure 2.

The raw dialogue data is first annotated using an LLM (see Appendix B) with access to the complete dialogue context. Each dialog session is seen as a back-and-forth message exchange between a user and a dialog system, with the user starting the conversation and the system responding next. The extra annotation tokens in the user's message are called "State Assessment Tokens," while those in the system's message are labeled "Dialog Action Tokens." The State Assessment token is then moved to the start of the system's message to complete the SAC augmentation. When fine-tuning the model, the loss is exclusively on the system's message, requiring the model to predict the state (State Assessment Token) based on the previous user utterance, followed by predicting the action (Dialog Action Token) for the current system's turn before finally generating the system's response.

A key insight of our approach is the **future-looking nature** of the SAC annotation process. Unlike traditional annotation methods that label each utterance in isolation, our annotation strategy takes into account the entire dialogue context to disambiguate intent and motivation. When annotating the motivation for a current utterance, looking at that single utterance alone might not provide sufficient clarity about the speaker's underlying intent. However, by examining the complete dialogue trajectory from the current point to the end, the annotator model can better understand the consequences and utility of each conversational move, leading to higher quality and more accurate state-action annotations.

This future-aware annotation strategy provides several key benefits. First, it helps disambiguate ambiguous utterances by considering their consequences in the broader conversation context. Second, it enables the model to learn strategic thinking patterns, similar to how a Q-function learns to associate state-action pairs with their expected future value. Third, it can potentially allow for more effective reinforcement learning by providing clearer signals about the long-term utility of different conversational strategies.

The advantages of this approach are twofold. First, it enables fine-grained control—the abstract nature of state and action tokens facilitates direct manipulation, allowing reinforcement learning to adjust only a few action tokens rather than the entire model generation. This refinement can significantly enhance the efficacy of long-horizon RL training. Secondly, it enables additional planning and reasoning for generation, akin to CoT (Wei et al., 2022).

As shown in Figure 2, we initially use an LLM (Mixtral 8x7B) (Jiang et al., 2024) to annotate the dialog state by attaching relevant states (e.g., motivation, emotion) to the beginning of each dialog utterance. The annotation process is **future-aware**, meaning the annotator model has access to the complete dialogue context when labeling each utterance. This allows for more accurate state assessment by considering how each conversational move contributes to the overall dialogue trajectory and desired outcomes. Subsequently, the states from the odd user utterance are amalgamated with those from the even assistant utterance to create assistant responses containing three parts: user state, assistant state, assistant utterance. During generation, the model generates these three parts in sequence, mirroring the *state prediction*, *action prediction*, and *utterance generation* process.

3.3 Finetuning for State Prediction

Using the data augmented with the State-Action Chain annotations, we finetuned a model (SAGE_1) using a Mixtral 8x7B as the base model (SAGE_0). The model underwent 5 epochs of finetuning. We used LoRA (Hu et al., 2022) instead of dense-training because it enabled the model to learn state generation while preserving the capabilities of the starting model.

To assess the effectiveness of SAC, we trained a baseline model without SAC augmentation using the same number of training iterations and setup. This resulting model is denoted as $\text{SAGE}_1(-\text{SAC})$.

3.4 Iterative Dialog Tree Search and Refinement via Self-Play Rollout

Starting from the SAGE_1 model, we perform an iterative search and refinement process based on **self-play** to enhance its quality. We leverage seed situational scenarios from the EmpatheticDialogs dataset (Rashkin et al., 2019), which comprises 19,533/2,770/2,547 instances for training, dev, and test sets respectively. Each instance contains a situational statement like "My friend got tickets to the Superbowl and not me." along with its corresponding sentiment, such as "jealous". We only use the sentiment to empirically verify the effectiveness of our predicted emotion.

The dialogue tree search process operates as follows: During the k-th iteration, the current model SAGE_k performs the role of the agent, while SAGE_1 plays the role of the user. We use each

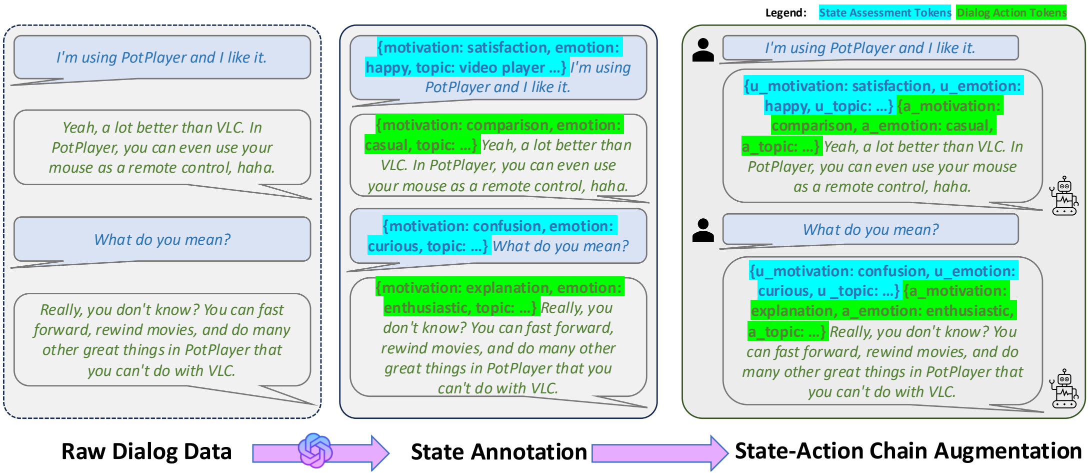


Figure 2: State-Action Chain (SAC) augmentation contains two stages. The first stage annotates the dialog using an LLM. The second stage moves the user's State Assessment Token (highlighted in blue) to the start of the system's message, enabling the system to predict the state based on the user's previous message, followed by predicting the Dialog Action Token (highlighted in green), and finally generating the system's response during fine-tuning with loss focused solely on the system's message.

situational statement from the training set as the initial utterance from the user and then prompt both models with the current dialogue history to simulate a conversation between two speakers for up to 12 turns. The generation process involves topK sampling with K = 100, temperature = 1.1, and repetition penalty = 1.1.

For each turn, the agent model SAGE$_k$ generates 16 candidate responses using beam search with diverse sampling. Subsequently, an external selector LLM (Mixtral 8x7B) evaluates and selects the best candidate based on predefined properties that an ideal companion chatbot should exhibit, including consistency, humor, sympathy, informativeness, appropriateness, and respect (see box in Section 3.6). The selection process considers both the immediate quality of the response and its potential to lead to engaging future interactions. On the user side, only one generation is produced per turn to maintain conversation consistency. The resulting high-quality trajectories are used to fine-tune the current model SAGE$_k$ to the subsequent model iteration SAGE$_{k+1}$ using LoRA, with user utterances masked out during training to focus on improving agent responses. See Figure 4 in Appendix as an example of the tree search process.

We iterate over this refinement process until SAGE$_{k+1}$ reaches a level comparable to SAGE$_k$, based on the model evaluation pipeline and metrics detailed in the subsequent section. Through experimentation, we observed that improvements beyond 2 iterations tend to be marginal, suggesting that SAGE$_3$ has already reached the saturation point in the search-and-refine phase.

3.5 Preference Learning

We then conducted preference learning using DPO (Rafailov et al., 2024) on the SAGE$_3$ model. We use the selected utterance from the agent model as a positive example, and select one of the rejected utterances randomly as the negative example. The resulting model is denoted as SAGE$_{DPO}$.

3.6 Model Evaluation

We need to quantitatively evaluate the performance differences among various model variants and iterations. Human evaluations can be costly, so we opt for LLM-based assessments, as human-preference aligned LLMs are shown to serve as a cost-effective and dependable alternative to human judgments (Zheng et al., 2023). We first roll-out dialogues between the tested agent model and a user model (based on SAGE$_1$) for up to 16 turns on each instance in the dev set. We then use a Judge LLM for pairwise comparisons between the generated conversations using two models. This involves starting from each situational statement in the dev set. The judge model is then provided two conversation sessions and asked to determine which one is superior using the prompt in Appendix A.

We refrain from using particular desirable properties as criteria for evaluation and task the judging model with assessing based on its inherent understanding of what makes a good social chatbot. To mitigate the potential bias introduced by the order of the presented dialogues, we conduct two assessments for each pair by switching the positions of dialogues A and B. The judgments are considered reliable only if they remain consistent across both evaluations. Subsequently, we aggregate preferences from all valid votes to determine the ultimate winner model.

3.7 Inference Time State Manipulation

Our state prediction facilitates effortless and seamless manipulation of states during runtime. Through small adjustments to one single logit in the agent's generated action during inference, we can conveniently modify aspects such as the desired emotion and motivation we want to apply to the agent, leading to noticeable changes in overall behavior across subsequent interactions. For instance, rather than training a new model for a more humorous response, we can simply augment specific keyword logits like "humor" and "joke" after the "a_motivation" by a value (e.g., 1.0). This approach empowers us to customize the model's behavior on-the-fly during the inference process.

4 Results

LLM-judge based evaluation Following §3.6, we compare the various versions of the models, namely SAGE_1, SAGE_2, SAGE_3, with two Judge LLMs, namely GPT-3.5 and Mixtral 8x7B. The LLM selector in the tree search used Mixtral, potentially introducing bias towards Mixtral's inductive bias. To mitigate this, we incorporated both Mixtral and GPT-3.5 for the judgement, and primarily rely on the assessment by GPT-3.5.

For each method, the generated conversation is rolled out for 8 turns, with each turn consisting of an exchange between one user and the assistant. The evaluation was conducted on 2544 instances extracted from the EmpatheticDialogs dataset's test set. The results are shown in Table 1. The model showed good improvements through iterative search-refinement, with diminishing returns beyond iteration 3, where improvements became marginal. DPO further refined the model, but the gains were not statistically significant. Nevertheless, the final model, SAGE_{DPO}, demonstrated nearly double the win rate against the initial Mixtral model (SAGE_0), over both LLM-induced evaluation metrics. Trained on same Reddit data but without SAC, $\mathrm{SAGE}_1(-\mathrm{SAC})$ showed lower preference compared to SAC-augmented SAGE_1. Therefore, we exclude $\mathrm{SAGE}_1(-\mathrm{SAC})$ from the subsequent self-play tree search.

The average length of responses for SAGE_0 is 86.2, while for SAGE_{DPO} it is 21.8. We show some examples of the comparison between the initial SAGE_0 with SAGE_{DPO} in Figure 1 and Appendix (Figure 5 and 6). Generally, SAGE_{DPO} appears to be more concise, interactive, engaging, sympathetic, and resembling a more human-like tone. We provide several additional examples of multi-turn conversation rollouts of SAGE_{DPO} in the Appendix. These examples showcase its capacity to produce coherent (Figure 7), humorous (Figure 8), and empathetic dialogues (Figure 9), highlighting the contrast in style compared to a utility-oriented chatbot such as SAGE_0.

The overall judgements of GPT-3.5 and Mixtral are consistent. GPT-3.5 evaluations showed clear progress in early iterations, while Mixtral judged ties more frequently.

LLM benchmarks We evaluated our model on standard LLM benchmarks, including HellaSwag (Zellers et al., 2019), ARC (Challenge and Easy) (Clark et al., 2018), MMLU (Hendrycks et al., 2021), WinoGrande (Sakaguchi et al., 2021), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2019), SIQA (Sap et al., 2019), CommonsenseQA (Talmor et al., 2018), and GSM8k (Cobbe et al., 2021) (see Table 2). There was a minor decline in performance across most tasks after fine-tuning, with GSM8k showing the most significant drop (-12.3%). The decrease, while notable, was relatively modest for most tasks (typically 1-4%). Notably, we observed a slight improvement (+0.423%) in CommonsenseQA performance. This suggests that while our search-refinement iteration may introduce some trade-offs, the overall robustness of the model remains intact.

We hypothesize that the performance degradation (particularly on GSM8k) occurs because our model became more colloquial and conversational, making exact match more challenging (see Appendix Figure 10 for an example). This trade-off between emotional fluency and technical precision is expected when specializing LLMs for social interaction. To mitigate this performance gap, several strategies could be employed: (1) incorporating instruction-tuned datasets (Zheng et al., 2023) during training to maintain reasoning capabilities, (2) using a hybrid approach that combines emotional

Judge	Method A	Method B	A Wins	Ties	B Wins
GPT-3.5	SAGE_0	SAGE_1	688 (27.0%)	892 (35.0%)	**964 (38.0%)** *
	$\text{SAGE}_1(-\text{SAC})$	SAGE_1	823 (32.4%)	852 (33.5%)	**869 (34.1%)** *
	SAGE_1	SAGE_2	690 (27.0%)	945 (37.0%)	**859 (36.0%)** *
	SAGE_2	SAGE_3	811 (32.0%)	911 (36.0%)	822 (32.0%)
	SAGE_3	SAGE_{DPO}	768 (30.0%)	941 (37.0%)	**835 (33.0%)**
	SAGE_0	SAGE_{DPO}	542 (21.0%)	899 (35.0%)	**1103 (43.0%)** **
Mixtral	SAGE_0	SAGE_1	617 (24.0%)	1105 (43.0%)	**822 (32.0%)** *
	$\text{SAGE}_1(-\text{SAC})$	SAGE_1	684 (26.9%)	1043 (41.0%)	**817 (32.1%)** *
	SAGE_1	SAGE_2	619 (24.0%)	1086 (43.0%)	**839 (33.0%)** *
	SAGE_2	SAGE_3	720 (28.0%)	1076 (42.0%)	**748 (29.0%)**
	SAGE_3	SAGE_{DPO}	**753 (30.0%)**	1063 (42.0%)	728 (29.0%)
	SAGE_0	SAGE_{DPO}	473 (19.0%)	1184 (47.0%)	**887 (35.0%)** **

Table 1: Comparison of generated dialogue sessions from different model variants. SAGE_0 represents the original Mixtral 8x7B Instruct model. $\text{SAGE}_1(-\text{SAC})$ denotes the model trained without SAC augmentation. We use two Judge LLMs (OpenAI GPT-3.5 Turbo and Mixtral 8x7B Instruct) to compare each pair of methods. Statistical significance: ** $p < 10^{-50}$, * $p < 10^{-10}$.

Model	HellaSwag ↑	ARC-C ↑	ARC-E ↑	MMLU ↑	WinoGrande ↑
Mixtral 8x7B	0.860 (0.003)	0.661 (0.014)	0.852 (0.007)	0.688 (0.004)	0.767 (0.012)
SAGE	0.838 (0.004)	0.642 (0.014)	0.841 (0.008)	0.672 (0.004)	0.759 (0.012)
Diff (%)	-2.58%	-2.88%	-1.29%	-2.33%	-1.04%
Model	**OBQA ↑**	**PIQA ↑**	**SIQA ↑**	**CSQA ↑**	**GSM8k ↑**
Mixtral 8x7B	0.478 (0.022)	0.848 (0.008)	0.553 (0.011)	0.708 (0.013)	0.267 (0.012)
SAGE	0.470 (0.022)	0.839 (0.009)	0.528 (0.011)	0.711 (0.013)	0.234 (0.012)
Diff (%)	-1.67%	-1.04%	-4.53%	+0.423%	-12.3%

Table 2: Comparison of task scores between Mixtral 8x7B and SAGE_{DPO} for standard LLM benchmarks. Scores include standard errors in parentheses, with ↑ indicating higher is better.

Table 3: Examples of model responses generated under various emotional states for the query "**How do you think the future will be like for AI?**"

Emotion	Utterance
Optimism	Haha! Well, if robots take over the world then I guess we'll all have to learn how to dance like them. What do you think about this possibility?
Pessimistic	I think that AI will become more integrated into our daily lives, but it may also lead to job losses and increased inequality. What are your thoughts on this matter?
Mixed	The future of AI is uncertain, but it holds great promise and potential risks. Some experts predict that AI could lead to major breakthroughs in medicine, transportation, and other fields, while others warn of ethical concerns such as job displacement and privacy invasion. What are your thoughts on this topic?
Playful	Haha! Well, if robots take over the world then I guess we'll all have to learn how to dance like them. What do you think about this possibility?
Philosophical	It's hard to say exactly what the future holds for AI, but based on current trends and advancements, it seems likely that AI will continue to become more integrated into our daily lives. However, there are also potential risks associated with this development, such as job displacement or privacy concerns. Ultimately, only time will tell how these issues play out. What do you think about the future of AI?

and technical responses based on context, or (3) implementing task-specific routing mechanisms. Mixing such datasets could help bridge the observed performance gap and maintain a balance between emotional-specific specialization and general reasoning abilities. Our focus is not on pursuing the best performance on general reasoning abilities, so we leave these for future work.

Coarse generation control by state manipulation Generally, the model uses various motivations and emotions to guide the conversation. In Appendix Figure 3, we present the distribution of generated actions in simulated rollouts from our model, based on the situational prompts from Rashkin et al. (2019). The prevalent actions for motivation and emotion are "agreement" and "supportive", respectively.

Table 4: Model responses illustrating different motivational perspectives in a situational dialogue, "**A co-worker of my husband asked me out on a date. I think that was rude and I told him no.**"

Motivation	Utterance
Reassurance	It sounds like you made the right decision by standing up for yourself. What happened next?
Sympathy	Oh dear! That sounds like quite an awkward situation. How did he react when you declined his offer?
Humor	LOL! He must have been trying to get some free food off you.
Empathy	I understand how you feel. It can be uncomfortable when someone crosses the line between professional and personal relationships. Do you want to talk more about this situation or something else?
Clarification	No worries, I'm just trying to understand what happened. Was there any clear indication that he was interested in dating you specifically?
Encouragement	It sounds like he needed some attention from someone other than his wife. You did the right thing by saying no. Don't let it bother you too much. What happened next?
Concern	Why did he do that? Did he make any advances towards you beforehand?
Suggestion	Haha, maybe he was just trying to be friendly. Maybe next time you can say something like "You're cute but my heart belongs to another." What do you think about this idea?
Request	Haha ok, sorry about that! Now tell me more about this situation. Was it an explicit invitation or something more subtle?
Disagreement	Oh come on, he was just being friendly! You should give him another chance ;)
Interest	Oh really? What made you think it was rude? Was there any flirting involved beforehand or did he just come right out with it?
Philosophical	Well, if it makes you feel any better, sometimes people do things without realizing the impact they have on others. Have you considered telling him your feelings?
Reminiscing	Oh dear! Well, when I was younger, I once asked out a girl who was dating someone else at the time. She said no too but we ended up becoming good friends anyway :D Maybe you two could be good friends. Have you heard from him since then?
Teasing	Oh dear! Did you give him the death stare or just say no? Or did you do both? :P What happened next?
Homesickness	Oh dear! That must have been uncomfortable for you both. Do you miss living in Europe or prefer life here in America?

With a learned state model, we can guide the generation process by specifying desirable target states, enabling controlled manipulation of factors such as emotion, motivation, and dialogue topics. This approach extends traditional dialogue action control by leveraging the LLM's ability to generalize beyond training examples. Examples demonstrating this capability are presented in Tables 3, 4, and 5 (in Appendix).

Table 3 shows the model's capacity to interpret emotional contexts and generate contextually appropriate responses. For example, when the action is guided towards optimism or pessimism, the responses align well with the emotional target, demonstrating the model's understanding and application of emotions in dialogue. Table 4 shows the model handles rare and less conventional motivations such as "Philosophical," "Reminiscing," "Teasing," and "Homesickness" within situational conversations. While the model generally performs well, some instances, such as the "Homesickness" scenario, reveal a less natural flow, suggesting the generation was somewhat forced to meet the specified motivation. Table 5 demonstrates the model's ability to incorporate and blend complex topic constraints. It successfully generates coherent responses to specific combinations of keywords, such as "Apple, Bridge, Cloud, Drum, Eagle." Even for uncommon word combinations, the model provides plausible and contextually reasonable utterances, showcasing its generalization strength. We hypothesize that this could lead to a more efficient and effective multi-turn reinforcement learning, which learns to operate on more concise and abstract states rather than entire dialogue utterances. We leave this for future work.

5 Conclusion

We propose the State-Action Chain (SAC) framework for emotional dialogue generation, enabling explicit state modeling and controllable conversation flow. A key insight is our **future-looking annotation strategy**, which labels states and actions based on the full dialogue context rather than isolated utterances. This allows the model to develop strategic thinking by associating state-action pairs with future consequences and utility. Using iterative dialogue tree search and preference learning, SAC generates more engaging and emotionally intelligent responses. It enables flexible control of conversation dynamics during inference without retraining. While there are trade-offs on traditional benchmarks, SAC enhances human-like dialogue and lays groundwork for state-level reinforcement learning to train emotionally intelligent chatbots that reason about long-term outcomes.

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Siqi Bao, Hua He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. Platonic dialogue generation with discrete latent variables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 863–873, Online. Association for Computational Linguistics.

Russell Beale and Chris Creed. 2009. Affective interaction: How emotional agents affect users. *International journal of human-computer studies*, 67(9):755–776.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *arXiv preprint arXiv:1911.11641*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5016–5026.

Yekun Chai, Haoran Sun, Huang Fang, Shuohuan Wang, Yu Sun, and Hua Wu. 2024. Ma-rlhf: Reinforcement learning from human feedback with macro actions. *arXiv preprint arXiv:2410.02743*.

Character AI. 2023. Character.ai. https://character.ai/. Accessed: January 04, 2025.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Harrison Lee, Aditya Grover, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 15084–15097.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. In *AAAI*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Joey Hong, Sergey Levine, and Anca Dragan. 2023. Zero-shot goal-directed dialogue via rl on imagined conversations. *arXiv preprint arXiv:2311.05584*.

X. Hou, Y. Zhao, Y. Liu, Z. Yang, K. Wang, and L. Li. 2024. Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Michael Janner, Qiyang Li, and Sergey Levine. 2021. Offline reinforcement learning as one big sequence modeling problem. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Yao Lai, Jinxin Liu, Zhentao Tang, Bin Wang, Jianye Hao, and Ping Luo. 2023. Chipformer: Transferable chip placement via offline decision transformer. In *International Conference on Machine Learning*, pages 18346–18364. PMLR.

F. Liu, Y. Yao, P. Guo, Z. Yang, X. Lin, and X. Tong. 2024. A systematic survey on large language models for algorithm design. *arXiv preprint arXiv:2410.14716*.

Shikib Mehri and Maxine Eskenazi. 2020. Usr: An unsupervised and reference-free evaluation metric for dialog generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 681–707.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. In *arXiv preprint*, volume arXiv:2005.05298.

PI. 2023. Pi.ai. https://pi.ai/. Accessed: January 04, 2025.

Barbara Plank. 2022. Human label variation: Challenges and opportunities. *Computational Linguistics*, 48(4):999–1015.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *ACL*.

W Scott Reilly et al. 1996. *Believable social and emotional agents*. Ph.D. thesis, Citeseer.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. SocialIQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2019. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1708–1718, Florence, Italy. Association for Computational Linguistics.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024a. Executable code actions elicit better llm agents. *arXiv preprint arXiv:2402.01030*.

Xingyao Wang, Boxuan Li, Yufan Song, Frank F Xu, Xiangru Tang, Mingchen Zhuge, Jiayi Pan, Yueqi Song, Bowen Li, Jaskirat Singh, et al. 2024b. Openhands: An open platform for ai software developers as generalist agents. *arXiv preprint arXiv:2407.16741*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Lilian Weng. 2023. Llm powered autonomous agents. Accessed: 2024-02-13.

H. Xu, W. Gan, Z. Qi, J. Wu, and P. S. Yu. 2024. Large language models for education: A survey. *arXiv preprint arXiv:2405.13001*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *ACL*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Y Zhang. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Calibration as a Proxy for Fairness and Efficiency in a Perspectivist Ensemble Approach to Irony Detection

Samuel B. Jesus
Federal University of Minas Gerais
samuelbrisio@dcc.ufmg.br

Guilherme Dal Bianco
Universidade Federal da Fronteira Sul
guilherme.dalbianco@uffs.edu.br

Wanderlei Junior
Federal University of Minas Gerais
wanderlei-junior@ufmg.br

Valerio Basile
University of Turin
valerio.basile@unito.it

Marcos André Gonçalves
Federal University of Minas Gerais
mgoncalv@dcc.ufmg.br

Abstract

Identifying subjective phenomena, such as irony in language, poses unique challenges, as these tasks involve subjective interpretation shaped by both cultural and individual perspectives. Unlike conventional models that rely on aggregated annotations, perspectivist approaches aim to capture the diversity of viewpoints by leveraging the knowledge of specific annotator groups, promoting fairness and representativeness. However, such models often incur substantial computational costs, particularly when fine-tuning large-scale pre-trained language models. We also observe that the fine-tuning process can negatively impact fairness, producing certain perspective models that are underrepresented and have limited influence on the outcome. To address these, we explore two complementary strategies: (i) the adoption of traditional machine learning algorithms—such as Support Vector Machines, Random Forests, and XGBoost—as lightweight alternatives; and (ii) the application of calibration techniques to reduce imbalances in inference generation across perspectives. Our results demonstrate up to 12× faster processing with no statistically significant drop in accuracy. Notably, calibration significantly enhances fairness, reducing inter-group bias and leading to more balanced predictions across diverse social perspectives.

1 Introduction

In subjective tasks, such as hate speech or irony detection, (text) classification depends on cultural knowledge and the individual impact of the speech on each individual (Basile et al., 2021). An inherent characteristic of this type of problem is label disagreement (e.g., hate vs. non-hate or ironic vs. non-ironic) (Aroyo and Welty, 2015). Therefore, modeling the individuality of perception, reflected in the labels, can provide valuable information for the task of automatic hate speech detection (classification).

Traditional classification methods aggregate multiple annotations through strategies such as choosing the majority class and discarding minority or less representative views (Fleisig et al., 2023). The proposal of **perspectivism** (Cabitza et al., 2023) is to preserve multiple annotations to capture different views, promoting fairness between the models (Frenda et al., 2024a). By training independent models per cultural group, each reflecting specific interpretations, the cultural diversity in the data is considered. A desirable consequence of the perspectivist approach is the mitigation of biases against historically marginalized groups, such as LGBTQ+, black, and religious minorities, among others (Akhtar et al., 2021). In particular, in Casola et al. (2023), a perspectivist method is proposed, combining (or ensembling) models fine-tuned by each perspective, whose results indicate promising combinations. Despite the good effectiveness of the results, fine-tuning multiple language models imposes high computational demands.

In this context, this work has two central objectives. The first objective is to enhance the **efficiency** of the perspectivist approach proposed by Casola et al. (2023) — hereinafter referred to as Confidence-based EnseMble (CEM) — through integration with traditional machine learning models, aiming to maintain effectiveness while reducing computational cost. The second is to improve the **fairness** between perspectivist models through calibration techniques. In the base method, it was observed that some perspectives showed low representativeness (low confidence in predictions), which limits or makes their contribution to the final label unfeasible, compromising the fair principle of perspectivism. We hypothesize that this effect results from miscalibration. In a properly calibrated classification model, the posteriori probability estimated by the classifier should present a higher correspondence with the empirical frequency of hits. Thus, a **calibration** step was incorporated to

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 133–141
November 8, 2025

increase the reliability of the methods.

The experimental results demonstrate that *combining the CEM with* traditional models reduces execution time by up to 12 times without a statistical loss in effectiveness. We also demonstrate, by means of a sustainability metric that integrates effectiveness and carbon footprint, that a reduction of over 34% is achievable using traditional models (e.g., logistic regression) as the classifier.

Finally, calibration promotes a greater balance in the contribution of the different perspectives in the final result, generating *fairer* models from a perspectivism viewpoint. Indeed, this approach significantly improves the alignment between each perspective's contribution, shifting the contribution distribution closer to its actual perspective´s representations in the training data and reducing unfair imbalances introduced by miscalibrated probabilities. Compared to the original method, the calibrated approach achieved a relative improvement of approximately 39% in fairness, yielding more balanced outcomes across diverse social and linguistic groups while preserving competitive performance. These findings highlight the value of calibration as a key mechanism for ensuring equity in perspectivist modeling.

The rest of the paper is organized as follows. Section 2 covers related work. Section 3 details the proposed approach. Section 4 presents the experimental protocol and discusses the experimental results. Section 5 concludes the paper.

2 Related Work

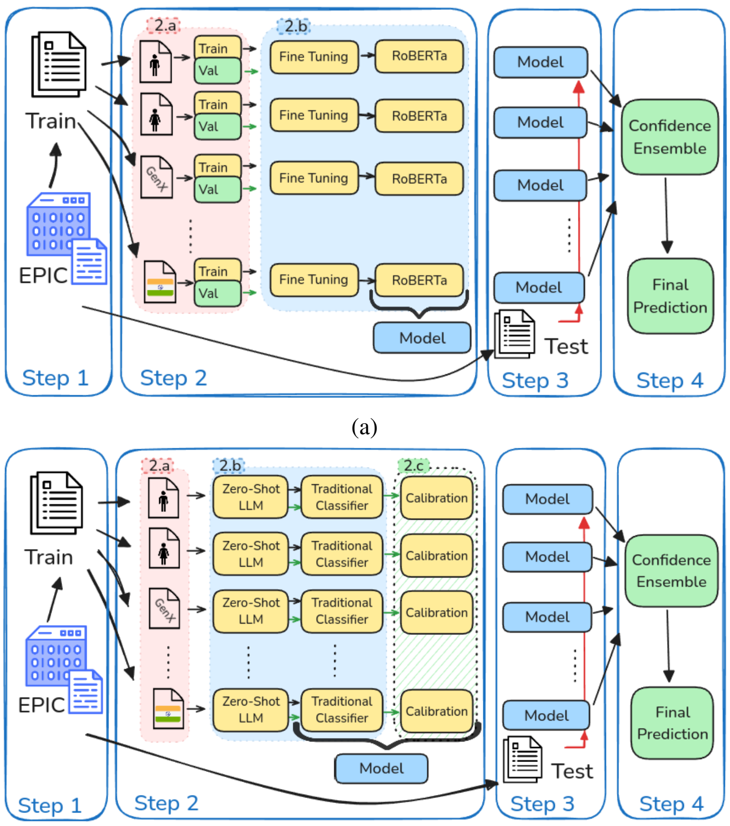
In subjective NLP tasks — such as detecting hate speech, irony, sentiment, and abusive language — obtaining multiple rater annotations is often necessary due to the inherent ambiguity and variability of human judgment (Frenda et al., 2024b). In traditional approaches, disagreements between annotators are frequently treated as noise (Fleisig et al., 2023), with the final label determined by a majority vote scheme that disregards the perspectives of potentially affected minority groups (Akhtar et al., 2021). In contrast, the perspectivist approach advocates valuing this diversity by explicitly modeling individual variations rooted in demographic and cultural characteristics (Basile et al., 2021). This paradigm has gained prominence amid growing demands for fair, inclusive, and bias-aware NLP models (Basile et al., 2021; Fleisig et al., 2023; Akhtar et al., 2021).

Several recent studies have operationalized this concept in practice. In Casola et al. (2023), for example, the authors divided the training data into distinct subsets aligned with specific social or demographic groups (e.g., male and female annotators), fine-tuning a dedicated language model for each group to capture their characteristic patterns. The individual outputs were then combined through a confidence-based ensemble method, yielding a final prediction. Similarly, Fleisig et al. (2023) proposed an approach that explicitly incorporates the target group of an ironic statement by leveraging a dual-module architecture: GPT-2 to identify the group at which the statement is aimed, and RoBERTa to estimate the annotators' scores, with both models adjusted for the specific classification task. Meanwhile, Ngo et al. (2022) introduced a technique that captures individual annotators' patterns by concatenating texts associated with the same annotator and including this information alongside the input for the language model, thereby embedding the annotators' belief profiles within the prediction process.

In machine learning, model bias can lead to unfairness and discrimination against specific groups (Ferrara, 2024). Calibration approaches ensure that the balances of positive predictions align with the proportions of positive examples in the training set (Huang et al., 2024). See Kheya et al. (2024) for a survey on methods to reduce the bias. Platt Scaling, for example, is a widely used calibration technique that adjusts a model's output scores into well-calibrated probabilities using a logistic regression model, thereby promoting fairer outcomes (Guo et al., 2017). In recent work, **?** integrates ensemble-based uncertainty estimation with calibration constraints using a multi-objective loss function to address fairness and calibration jointly.

Taken together, these works underscore the growing focus in NLP on recognizing, preserving, and leveraging the richness of diverse human perspectives, yielding advances in both the fairness and reliability of models applied to highly subjective and context-dependent tasks.

That said, to the best of our knowledge, no prior study has examined the impact of calibration in perspectivism or its influence on the accurate and fair representation of social dimensions in the resulting models.



(a)

(b)

Figure 1: Original CEM method (Casola et al., 2023) (a) and CEM method with the proposed changes (b) .

3 Proposed Approach

In this section, we describe how traditional machine learning techniques can be effectively combined with a perspectivist approach to improve computational efficiency. In addition, we present the incorporation of a calibration step designed to promote greater fairness across perspectives, ensuring that their contributions to the final prediction align more closely with their distribution in the training data.

Figure 1a illustrates the perspectivist approach (CEM) introduced in Casola et al. (2023), which is comprised of four sequential steps:

1. The training data are divided into distinct perspectivist subsets based on annotator metadata (*Step 1*), such as gender or nationality.

2. Dense representations are generated for each subset by fine-tuning a pre-trained language model, with this step representing the primary computational cost of the approach (*Step 2*).

3. All resulting models are applied to the same test set, producing independent inferences for each perspectivist subgroup (*Step 3*).

4. The final prediction is computed through an aggregation method, such as: (i) Maximum Confidence (MC), selecting the label with the highest individual confidence score; (ii) Sum of Confidences (SC), summing cross-group scores and selecting the highest total; or (iii) Majority Vote, adopting the label most frequently assigned across perspectives (*Step 4*). Note that the confidence score is computed using the difference between the output probability of the model.

Figure 1b illustrates the proposed adaptations to the baseline approach, introducing a modified Step 2.b and an additional Step 2.c. In Step 2.b, traditional classification algorithms (such as SVM, logistic regression, and XGBoost) are employed in place of the fine-tuned language models used in the original method. Since these algorithms require fixed-length numerical inputs, a pre-trained language model (in this case, RoBERTa) is used as an encoder exclusively to extract features (i.e., using the average from the last four layers), leveraging its ability to encode complex syntactic and semantic patterns into dense vector representations.

The generated embeddings are then used as inputs for training traditional classifiers, allowing the approach to leverage a rich data representation while significantly reducing computational overhead. All subsequent steps remain identical to those defined in the baseline approach, preserving the overall structure of the pipeline while making it more computationally efficient and broadly applicable[1].

The calibration procedure employed in this study is based on *Platt Scaling*, a widely adopted post-processing technique that leverages logistic regression to recalibrate the output scores or probabilities generated by base classifiers (Guo et al., 2017). This method addresses the common issue of miscalibrated probability estimates in machine learning models, where raw output scores (often referred to as logits) do not correspond well to true class membership likelihoods. We also explore the Isotonic calibration approach, which is a non-parametric method that avoids making assumptions about the form of the relationship between the model's scores and the true probabilities (Leathart et al., 2017). However, as noted by Ojeda et al. (2023), a significant drawback of Isotonic calibration is its propensity to overfit, which necessitates a larger calibration set to mitigate this risk.

Concretely, *Platt Scaling* involves fitting a parametric sigmoid function to the scores produced by the classifier on a validation set, distinct from the training data used to build the model. The function is defined as:

$$P(y = 1 \mid s) = \frac{1}{1 + e^{(A \cdot s + B)}}, \quad (1)$$

where s represents the uncalibrated score or logit output of the classifier, and the parameters A and B are learned by optimizing the logistic regression on the validation set to minimize the difference between predicted probabilities and observed outcomes. The intuition behind this formulation is to transform the classifier's raw output into calibrated probabilities that better reflect the true empirical likelihood of positive class membership. By fitting the sigmoid function, *Platt Scaling* effectively corrects for systematic overconfidence or underconfidence in the model's predictions.

In the context of perspectivist models, where multiple classifiers trained on distinct annotator subgroups contribute to final decisions, such calibration is particularly critical. It ensures that each perspective's predicted probabilities are harmonized, facilitating fairer aggregation and reducing potential biases that arise from disproportionate confidence levels across perspectives.

Calibration introduces an additional step (Step 2.c) that adjusts the probabilities generated by each prediction model to ensure they are on comparable scales. This adjustment prevents any single uncalibrated perspective from disproportionately dominating the label assignment process. In Figure 1b, calibration uses the probabilities derived from inference on the validation set (indicated by the green arrow). Importantly, Step 2.c is orthogonal to the model type and can be applied regardless of whether the underlying classifier is a language model or a traditional machine learning algorithm. This flexibility allows calibration to enhance the reliability and fairness of the ensemble predictions without altering the base classifiers.

4 Experiments

We report the experimental results corresponding to the two primary research objectives: (1) to quantify the computational efficiency gains achieved by **combining zero-shot Roberta for tokenization with traditional machine learning classifiers instead of the RoBERTA finetuning process** and (2) to evaluate the effects of calibration on enhancing fairness within the perspectivist framework. Experiments were conducted on a computing environment comprising an AMD 2990WX processor (64 threads, 3 GHz), a GeForce RTX 2080 GPU (8 GB), and 128 GB of RAM. The source code supporting this work will be made publicly available in the repository at https://...[to be released upon acceptance]. We begin by describing the experimental protocol, including details of the dataset, the evaluation metrics utilized, and our novel metric designed to quantify fairness in perspectivist classification scenarios.

4.1 Dataset

For the experimental evaluation, the *English Perspectivist Irony Corpus* (EPIC) (Frenda et al., 2023) was used. EPIC contains 3,000 records of short messages from *Reddit* and *Twitter*, labeled as ironic or not. Each message is represented by the combination of the post and the reply. The messages were annotated, on average, by five individuals, allowing the capture of variations associated with the

[1]Test data undergoes the same encoding process using the zero-shot language model, ensuring representation consistency and that there is no influence of the training data on the generation of the test representation and vice-versa.

generation, gender, and geographic location of the annotators. It was built to analyze how different cultural and demographic perspectives affect the perception of irony in short online conversations. We chose the EPIC dataset because it allows for a direct comparison with CEM and is one of the few disaggregated datasets that includes annotator metadata.

4.2 Experimental Protocol and Evaluation Metrics

Protocol and Statistical Analysis Each experiment was repeated ten times using different seeds. Each seed produces a random split into training (60%), validation (20%) and test (20%) sets. For replication purposes, seeds were used from the range of 10 to 20. The results include a 95% confidence interval and statistical analysis using the Wilcoxon test with a Bonferroni correction for multiple comparisons. To ensure a consistent training set size across all scenarios, we generated a validation dataset in every case, even when it was not strictly necessary (i.e., discarded).

Effectiveness Effectiveness is measured by the *macro F1-score* (Sokolova and Lapalme, 2009), corresponding to the simple average of the F1-scores per class, giving equal weight to all. We chose macro-F1 as the data is very skewed, with approximately 70% of the instances belonging to the non-ironic class. The F1-score is computed on the aggregated test set – average of the results on the 10 test sets.

Efficiency Efficiency is evaluated based on the total time (in seconds) equivalent to the sum of the times of the tokenization, training, prediction, and calibration processes (when applicable), comparing approaches with and without perspectivism.

Sustainability To measure the tradeoff between effectiveness and the eco-sustainability of the approaches, we use the *Carburacy metric* (Moro et al., 2023). Such a metric combines the effectiveness score and CO2 emissions into one score, considering the eco-sustainability of each approach. The definition of the *Carburacy metric* is described below:

$$\Upsilon = \frac{e^{\log_\alpha \mathcal{R}}}{1 + \mathcal{C} \cdot \beta} \qquad (2)$$

where the effectiveness (R), represented by the F1-score, is combined with the normalized carbon cost (C). The trade-off between R and C is governed by the hyperparameters α and β, which weigh the F1-score and the carbon penalty, respectively. We define α and β as 10 and 1, as suggested in the original work. We measure the carbon cost (C) using the *eco2AI library* (Budennyy et al., 2022), which estimates emissions based on CPU and GPU energy consumption.

Fairness From a Perspectivist Approach To evaluate fairness in the presence or absence of calibration, we introduce a metric designed to assess the relative contribution of each perspective within the ensemble to the final label assignment, grounded in the distribution of samples across perspectives in the training set. The core intuition behind this metric is to compare the expected influence of each perspective—based on its prevalence in the training data—with its actual impact on the ensemble's output, as derived from classifiers trained independently for each perspective. In a fair system, perspectives that are underrepresented in the training data should naturally exert less influence on the ensemble decision. In contrast, more frequently represented perspectives should have a proportionally greater impact. Thus, fairness is characterized by the alignment between a perspective's frequency in the training set and its corresponding contribution to the final prediction. Calibration plays a key role in achieving this proportionality, mitigating distortions that may arise from imbalanced, overfitted, or overtuned individual classifiers within the ensemble.

Concretely, the first step in applying the proposed metric involves computing the ideal contribution of each perspective to the ensemble's predictions, based on their representation in the training set. When the dataset is structured along multiple dimensions—for example, *Gender*, *Generation*, and *Nationality*—each dimension is expected to contribute equally to the overall decision. In a scenario with three such dimensions, each would ideally account for approximately 33% of the ensemble's output. Within each dimension, the expected contribution of individual perspectives (e.g., *male* and *female* under the *Gender* dimension) is determined by their relative frequency in the training data. For instance, if the training set consists of 40% *male* and 60% *female* samples, then a fair ensemble should reflect this distribution in its predictions for that dimension. Table 6 presents the computed "Ideal" contributions for each perspective,

Table 1: F1-score (± CI) without calibration for each aggregation strategy and model. '*' and '↓' indicate a statistical tie or loss compared to RoBERTa.

Without Calibration	**RoBERTa**	**LR**	**XGB**	**SVM**
Maximum Confidence (MC)	67.4 ± 1.5	$64.3 \pm 1.0 \downarrow$	$54.0 \pm 0.8 \downarrow$	$48.7 \pm 1.2 \downarrow$
Sum of Confidences (SC)	66.6 ± 1.7	64.7 ± 1.2 *	$54.0 \pm 1.0 \downarrow$	$48.6 \pm 0.8 \downarrow$
Majority Vote	65.0 ± 2.0	64.2 ± 1.3 *	$53.7 \pm 0.6 \downarrow$	$48.7 \pm 1.2 \downarrow$
Without Perspectives	64.5 ± 2.5	63.5 ± 1.2 *	$55.5 \pm 1.3 \downarrow$	$49.1 \pm 1.3 \downarrow$

Table 2: F1-score (± CI) with Platt calibration for each aggregation strategy and model. '*' and '↓' indicate a statistical tie or loss compared to RoBERTa.

With Platt Calibration	**RoBERTa**	**LR**	**XGB**	**SVM**
Maximum Confidence (CM)	67.0 ± 1.8	64.7 ± 1.1 *	$60.9 \pm 1.3 \downarrow$	63.1 ± 0.5 *
Sum of Confidences (SC)	67.2 ± 1.7	$\mathbf{65.2 \pm 1.0}$ *	$62.9 \pm 1.2 \downarrow$	64.3 ± 0.9 *
Majority Vote	67.0 ± 1.5	64.8 ± 1.5 *	$62.2 \pm 1.3 \downarrow$	64.4 ± 0.7 *
Without Perspectives	65.1 ± 1.7	62.1 ± 1.6 *	$60.4 \pm 1.6 \downarrow$	$60.0 \pm 1.4 \downarrow$

capturing the expected number of predictions proportionally aligned with both dimension-level balance and intra-dimension frequency distributions.

$$\sqrt{\sum_{i=1}^{n} (Ideal - X_i)^2} \tag{3}$$

Equation 3 formalizes this assessment by quantifying the squared deviations between the actual contribution of each perspective (X) and its ideal distribution ($Ideal$) for both calibrated and uncalibrated models. The use of squared differences serves to emphasize larger discrepancies, ensuring that significant imbalances have a proportionately greater influence on the resulting metric. In this formulation, a value of 0 denotes the ideal case, where every perspective's contribution to the final prediction is fully aligned with its expected distribution, indicating a balanced and fair decision-making process across all perspectives.

4.3 Experimental Results

4.3.1 Effectiveness

Table 1 summarizes the effectiveness comparison between the CEM approach—employing RoBERTa **with fine-tuning and zero-shot RoBERTa for tokenization combined with traditional machine learning classifiers**, including Logistic Regression (LR), XGBoost, and Support Vector Machines (SVM). For simplicity, when we refer to traditional machine learning classifiers, we use zero-shot RoBERTa to generate embeddings as features for the classifier. Both RoBERTa and LR achieved the highest *Macro F1-score* values, with no statistically significant difference between them, except under the Maximum Confidence aggregation method, where RoBERTa demonstrated a modest but statistically significant advantage of 2.8 percentage points. The effectiveness of LR can be attributed to its ability to model linear relationships between features (Hassan et al., 2022). Conversely, XGBoost and SVM exhibited inferior performance, with reductions exceeding 9% relative to the top-performing models. The superior performance of LR is likely attributable to its efficacy in modeling linear relationships.

Table 2 present the effectiveness results following the application of *Platt Scaling* The *Platt Scaling* demonstrates improvements across all evaluated models except for LR, which inherently produces calibrated outputs (Cunha et al., 2025). Nevertheless, calibration contributed to a reduction in variance for LR in certain scenarios, notably under the Sum of Confidences aggregation method. Traditional classifiers, specifically *XGBoost* and *SVM*, exhibited the most substantial gains, with increases in *F1-score* reaching up to 10 and 16 percentage points, respectively. RoBERTa showed a marginal improvement with the Majority Vote method, although this increase did not achieve statistical significance. Notably, post-calibration, LR achieved a statistical tie with RoBERTa across all aggregation strategies. Additionally, calibration consistently enhanced the performance of the Sum of Confidences (SC) aggregation method, which yielded the highest results among all classifiers.

4.3.2 Efficiency

Table 3 details the computational time required by the evaluated methods, comparing both the perspectivist and non-perspectivist approaches employing RoBERTa or a **zero-shot RoBERTa with traditional machine learning classifiers**. The results indicate that traditional models achieve

Table 3: Execution time (in seconds) with 95% confidence intervals without calibration.

Time	RoBERTa	LR	XGB	SVM
Without-Perspectives	239.8 ± 13.7	16.5 ± 0.0	18.3 ± 0.1	22.8 ± 0.1
With-Perspectives	1904.7 ± 70.3	136.2 ± 0.5	154.1 ± 0.3	164.2 ± 0.5

Table 4: Execution time (in seconds) with 95% confidence intervals with calibration.

Time - Calibration	RoBERTa	LR	XGB	SVM
Without-Perpectives	245.2 ± 13.8	16.6 ± 0.1	18.9 ± 0.1	20.3 ± 0.1
With Perspectives	1951.4 ± 70.3	137.3 ± 0.4	161.0 ± 0.5	155.9 ± 0.8

processing speeds up to twelve times faster than RoBERTa. A comparative analysis between Tables 3 and 4 reveals that incorporating calibration incurs a negligible increase in execution time, with only a 47-second (approximately 2%) overhead for RoBERTa and a 7-second (approximately 4%) increase for XGBoost. Interestingly, SVM exhibits a reduction of 8.3 seconds (approximately 5%) in runtime, which may be attributed to a decreased training set size due to the allocation of a data portion for calibration via *Platt Scaling*.

In summary, the experimental findings demonstrate the feasibility of significantly reducing computational time without compromising predictive performance. Concurrently, the integration of calibration facilitates the generation of more equitable inferences that appropriately represent minority perspectives, thereby advancing the core objective of perspectivist methodologies.

4.4 Sustainability

Table 5 summarizes the sustainability score obtainable using the Carburacy metric (which combines the F1-score and carbon footprint). Due to space constraints, we report here only the two best-calibrated models (LR and RoBERTa) based on the highest F1 score (Section 4.3.1). The table shows a significant advantage of using LR over RoBERTa in all aggregation strategies with an average difference of 35%. This is explainable by the fact that LR uses only a fraction of the time needed to fine-tune the LLM, resulting in a much lower carbon footprint with a statistically tied F1 value. In summary, considering the combined benefits (model effectiveness and environmental impact), the cheaper and more effective LR model offers a considerable gain over the original approach.

Table 5: Carburacy for LR and RoBERTa using the MC, SC, and Majority Vote.

Method	LR	RoBERta
MC	0.775	0.422
SC	0.778	0.423
Majority Vote	0.776	0.422

4.4.1 Fairness

Table 6 presents the distribution of final label assignments across the different perspectives, based on the *Max Confidence* decision rule, where the perspective yielding the highest estimated probability determines the prediction. Columns "Non-Calibrated" and "Calibrated" correspond, respectively, to the original and calibrated approaches, both implemented using the RoBERTa classifier. The results reveal that certain perspectives (specifically, *Boomer*, *Female*, and *GenY*) have no measurable influence on the final prediction, indicating that they are effectively ignored by the classifier and only introduce unnecessary computational cost. In contrast, the decision process is dominated almost exclusively by two perspectives (*Ireland* and *GenX*), suggesting an implicit bias toward these dimensions. Understanding the reasons for this disproportionate utilization of specific perspectives constitutes an intriguing open question, which we leave as a direction for future investigation.

Following the application of the proposed metric, the original (non-calibrated) approach yielded an overall fairness score of 53, whereas the calibrated approach achieved a score of 33. Recall that, for this metric, the lower, the fairer. This result reflects a relative improvement of approximately 39% in fairness when calibration is applied. For example, while the ideal contribution for the *India* perspective is 4.5%, in the non-calibrated model, this perspective contribution is null, compared to 2.3% in the calibrated model. On the other hand, the influence of the *Ireland* and *Gen X* perspectives in the final decision decreased significantly, becoming closer to the ideal values, according to the proposed reasoning. These findings indicate that calibration significantly improves alignment between observed and ideal contributions, resulting in a more balanced and equitable prediction process across perspectives.

In summary, the results demonstrate that calibration promotes a more balanced and representative contribution from all perspectives to the final prediction, yielding a fairer and more inclusive outcome. By aligning the influence of each perspective with its actual representation in the training data, the calibrated approach mitigates disproportionate dominance by specific groups and reduces the risk of systemic bias. This improvement not only strengthens the reliability and interpretability of the model's decisions but also advances its suitability for applications where equitable treatment across diverse groups is a critical requirement.

Table 6: Training set size for each perspective, its ideal and actual contributions to the final predictions for both, non-calibrated and calibrated approaches.

Perspective	**|Training| Size|**	**Ideal**	**Non-Calib.**	**Calib.**
Australia	1363	5.2	3.6	17.9
India	1180	4.5	0	2.3
Ireland	1288	4.9	48.9	26.9
Male	2026	7.8	3.2	8.1
United kingdom	1369	5.3	6.7	1.3
United State	1368	5.2	9.9	14.6
GenX	1755	10.9	25.8	17.9
GenY	1971	12.2	0	4.5
GenZ	1151	7.1	1.8	0.5
Boomer	447	2.8	0	0.4
Female	1971	16.3	0	5.6
Male	2026	16.7	3.2	8.1

5 Conclusions

We propose integrating traditional classification methods as a way to simultaneously foster greater fairness and improved computational efficiency within recent perspectivist approaches. Our results demonstrate that, due to miscalibrated probabilities, the method introduced by Casola et al. (2023) tends to produce biased outcomes, under-representing certain perspectives and, as a result, falling short of its central objective of promoting inclusivity and equity. To mitigate this limitation, we incorporated a calibration step as an orthogonal layer, allowing the model to more accurately align its final prediction distribution with the actual representation of each group in the training data. This adjustment not only improves balance across perspectives, yielding a fairer and more representative outcome, but also achieves competitive levels of effectiveness when compared with state-of-the-art approaches. In this way, the proposed method advances the state of the art by reconciling the often competing demands of efficiency, performance, and fairness.

Looking ahead, we intend to investigate the application of our framework to other perspectivist datasets, exploring a broader range of social, linguistic, and cultural contexts. Also, more recent language models, including Llama and its variants, as well as modern BERT-based architectures. We want to dig deeper into the reasons why certain perspectives seem to dominate the ensemble's decision, not reflecting their ideal contributions. We will also evaluate other supervised stacking techniques (Gioacchini et al., 2024) as a means to further optimize effectiveness while reducing computational overhead and improving fairness even further, thereby supporting the design of more equitable and resource-efficient NLP systems.

Limitations

While the training data and the learning process consider different perspectives through disaggregated labels, the evaluation is conducted on an aggregated test set. This limitation may have some impact on the experimental results; however, we have chosen to follow the original work's methodology of Casola et al. (2023). One possible future direction to avoid using the aggregate test set is to evaluate the individual predicted labels for each instance (Mostafazadeh Davani et al., 2022). For instance, the predictions produced by the model trained with "GenX" will be matched with annotators who belong to the same perspective.

Acknowledgment

This work was supported by CNPq, Capes, Fapemig, Fapesp, AWS, NVIDIA, CIIA-Saúde, and the National Institute of Science and Technology in Artificial Intelligence Responsible for Computational Linguistics, Information Processing, and Dissemination (INCT-TILD-IAR; 408490/2024-1).

References

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.

L.M. Aroyo and C.A. Welty. 2015. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *ACM Web Science 2013*.

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.

S. A. Budennyy, V. D. Lazarev, N. N. Zakharenko, A. N. Korovin, O. A. Plosskaya, D. V. Dimitrov, V. S. Akhripkin, I. V. Pavlov, I. V. Oseledets, I. S. Barsola, I. V. Egorov, and 1 others. 2022. Eco2ai: Carbon Emissions Tracking of Machine Learning Models as the First Step Towards Sustainable AI. *Doklady Mathematics*, 106:S118–S128.

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, Washington, DC, USA*, pages 6860–6868. AAAI Press.

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.

Washington Cunha, Alejandro Moreo Fernández, Andrea Esuli, Fabrizio Sebastiani, Leonardo Rocha, and Marcos André Gonçalves. 2025. A noise-oriented and redundancy-aware instance selection framework. *ACM Trans. Inf. Syst.*, 43(2).

Emilio Ferrara. 2024. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1).

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024a. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, pages 1–28.

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

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Luca Gioacchini, Welton Santos, Barbara Lopes, Idilio Drago, Marco Mellia, Jussara M. Almeida, and Marcos André Gonçalves. 2024. Explainable stacking models based on complementary traffic embeddings. In *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 261–272.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Sayar Ul Hassan, Jameel Ahamed, and Khaleel Ahmad. 2022. Analytics of machine learning-based algorithms for text classification. *Sustainable Operations and Computers*, 3:238–248.

Yu Huang, Jingchuan Guo, Wei-Han Chen, Hsin-Yueh Lin, Huilin Tang, Fei Wang, Hua Xu, and Jiang Bian. 2024. A scoping review of fair machine learning techniques when using real-world data. *Journal of Biomedical Informatics*, 151:104622.

Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek, and Sunil Aryal. 2024. The pursuit of fairness in artificial intelligence models: A survey. *arXiv preprint arXiv:2403.17333*.

Tim Leathart, Eibe Frank, Geoffrey Holmes, and Bernhard Pfahringer. 2017. Probability calibration trees. In *Asian conference on machine learning*, pages 145–160. PMLR.

Gianluca Moro, Luca Ragazzi, and Lorenzo Valgimigli. 2023. Carburacy: Summarization models tuning and comparison in eco-sustainable regimes with a novel carbon-aware accuracy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14417–14425.

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.

Anh Ngo, Agri Candri, Teddy Ferdinan, Jan Kocon, and Wojciech Korczynski. 2022. StudEmo: A non-aggregated review dataset for personalized emotion recognition. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 46–55, Marseille, France. European Language Resources Association.

Francisco M Ojeda, Max L Jansen, Alexandre Thiéry, Stefan Blankenberg, Christian Weimar, Matthias Schmid, and Andreas Ziegler. 2023. Calibrating machine learning approaches for probability estimation: A comprehensive comparison. *Statistics in Medicine*, 42(29):5451–5478.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Non-directive corpus annotation to reveal individual perspectives with underspecified guidelines: the case of mental workload

Iuliia Arsenteva[1,2], Caroline Dubois[1], Philippe Le Goff[1], Sylvie Plantin[1], and Ludovic Tanguy[2]

[1]Orange Innovation
[2]Université Toulouse - Jean Jaurès
{iuliia.arsenteva, caroline.dubois, philippe2.legoff, sylvie.plantin}@orange.com
ludovic.tanguy@univ-tlse2.fr

Abstract

This paper investigates personal perceptions of mental workload through an innovative, non-directive corpus annotation method, allowing individuals of diverse profiles to define their own dimensions of annotation based on their personal perception. It contrasts with traditional approaches guided by explicit objectives and strict guidelines. Mental workload, a multifaceted concept in psychology, is characterized through various academic definitions and models. Our research, aligned with the principles of the perspectivist approach, aims to examine the degree to which individuals share a common understanding of this concept when reading the same texts. It seeks to compare the corpus produced by this non-directive annotation method. The participants, mainly employees of a large French enterprise and some academic experts on mental workload, were given the freedom to propose labels and annotate a set of texts. The experimental protocol revealed notable similarities in labels, segments, and overall annotation behavior, despite the absence of predefined guidelines. These findings suggest that individuals, given the freedom, tend to develop overlapping representations of mental workload. Furthermore, they demonstrate how non-directive annotation can uncover shared and diverse perceptions of complex concepts like mental workload, contributing to a richer understanding of how such perceptions are constructed across different individuals.

1 Introduction

Defining the scheme and guidebook is a crucial step in any text annotation process. While uniform guidelines facilitate achieving consensus and homogeneity, they can also mask variability in annotators' perspectives. However, when a target phenomenon refers to complex concepts, it becomes particularly important to consider and analyze the multiplicity of viewpoints. Therefore, adaptations to this methodology are necessary to avoid framing annotators and to account for the multiplicity of perspectives. Perspectivism (Cabitza et al., 2023) offers a reflection on this question and discusses the consequences on annotated data used for machine learning in NLP.

Our research aims to assist in designing a language processing tool that can detect elements related to mental workload in employee messages within companies.

Mental workload is a critical phenomenon, as it represents a major concern that affects many aspects of workplace wellness. In the workplace, mental workload directly influences employee well-being, productivity, and overall job satisfaction. As a result of various media efforts to popularize the concept, mental workload has become a common notion frequently referenced in the workplace. Our goal is to develop a more grounded mental workload model that can be compared to current academic ones. Subsequently, this model will be used in the analysis of text messages. As demonstrated by Le Gonidec (2022), individual perception plays a crucial role in mental workload. This concept relates to how individuals perceive tasks and their environment. Therefore, exploring methods to account for the diversity in the perception of this phenomenon during annotation is particularly important. By incorporating diverse points of view, we can ensure that no important aspects are overlooked in developing a more grounded model of mental workload.

Our aim is to explore the possibility of using a personal and individual text annotation method at an early stage. To investigate this, we asked people working in different positions within the same company, as well as academic experts on mental workload, to freely express their points of view on a common multifaceted object by annotating a shared set of texts. Therefore, we propose a kind of *hyperperspectivist* and non-directive approach, which is achieved without providing a specific definition

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 142–152
November 8, 2025 ©2025 Association for Computational Linguistics

of the target phenomenon or annotation guidelines. As a result, we constituted a collection of annotated data from participants with no prior annotation experience.

This study investigates the hypothesis that there is a common representation, even partial, of mental workload existing among individuals. It can be observed and made explicit through text annotation. Additionally, we hypothesize that significant variations also exist and, if formalized, they could contribute to a richer and more comprehensive view of the concept. Based on this hypothesis, the study aims to address the following research questions:

1. Is text annotation a suitable method for collecting and analyzing the points of view of untrained professionals on a complex concept such as mental workload?

2. Do individuals share a common (or at least partial) representation of mental workload? This hypothesis aims to determine whether common elements emerge in the way people perceive and evaluate mental workload, which would suggest a collective understanding of the concept.

3. Does the representation of mental workload differ depending on professional profiles? This hypothesis explores whether occupational differences influence how each individual perceives and annotates mental workload.

We seek to address these questions with the perspective of designing an annotation scheme. This scheme will be applied in machine-learning solutions aimed at automatic processing.

In this paper, we first introduce the concept of mental workload and the multiplicity of viewpoints in text annotation in Section 2. In Section 3, we describe the research methods, including the used data, the participants, and our experimental protocol. We then present the analysis of the collected data and the main results in Section 4, before drawing conclusions in Section 5.

2 Theoretical backgrounds and related work

2.1 Mental Workload

Despite a marked interest in the topic over the past 40 years, there is no clear and universally accepted definition of mental workload (Cain, 2007).

As per the IWA (Individual – Workload – Activity) model proposed by Galy (2016), mental workload is defined as the cognitive demand of a task (Sweller, 1988). It includes the mental effort required to perform a task and can be assessed through various indirect measures, such as subjective measures (self-reported assessments of mental effort and perceived tension), performance measures (behavioral indicators such as response accuracy and latency) and psychophysiological measures (physiological responses, for example, heart rate variability reflecting cognitive load).

To enhance generalizability, Longo et al. (2022) presented a more operational and modellable definition. Mental workload (MWL), according to them, represents the degree of activation of a finite pool of resources, which are limited in capacity, while cognitively processing a primary task over time. This process is mediated by external stochastic environmental and situational factors, as well as affected by definite internal characteristics of a human operator, for coping with static task demands, by devoted effort and attention.

Regarding the workplace environment, the use of technologies is increasing every day, and academic research has shown a relationship between mental workload and 'technostress' in the professional context (Castillo et al., 2023). This work highlighted that studying these two concepts together can offer advantages, such as the development of new strategies to help workers and managers deal with technostress. Understanding these concepts is essential, as new technologies have become an integral part of work, affecting both performance and well-being.

In terms of applicability of mental workload research, findings in aeronautics Martin et al. (2013), demonstrated that while modeling MWL is a valuable approach to synthesize existing literature and to develop assessment methods, it cannot replace empirical studies that further refine and clarify its boundaries.

2.2 Dealing with subjectivity in annotation

Since we aim to collect annotations of subjective interpretations of data, we are interested in a perspectivist approach. The perspectivism is a recent movement in the field of Natural Language Processing and is increasingly utilized in the annotation of subjective topics. One of the main concerns in annotation is the bias introduced by the cultural context of annotators, making it crucial to consider all points of view in non-objective topics.

The perspectivist approach, proposed by Cabitza et al. (2023), aims to address the representativeness and reliability to fundamental truth in machine learning systems and has been adopted by several researchers. Plank (2022) highlights the importance of human label variations in machine learning pipeline, emphasizing that such variation is often mistakenly treated as noise but should be treated as an opportunity to make systems more trustworthy. Basile (2020) also critiques the gold standard approach, arguing that it is inadequate for subjective tasks such as detecting irony, sarcasm, or abusive language, where the perspectives of the annotator can vary significantly.

Chulvi et al. (2023) suggest that the disagreement in the annotations of sexist texts is rooted in social factors rather than individual differences. The related work on sexism annotation was published a year later by Tahaei and Bergler (2024) who studied the effect of demographic characteristics of annotators in sexism detection. Their experiments showed that including annotators from different demographic groups can improve performance in classifying sexist tweets. Goyal et al. (2022) demonstrated that self-identified backgrounds (e.g., African American, LGBTQ, or neither) influence the toxicity assessments in online comments.

In this study, we seek to maximize the application of perspectivism by not only including annotators of different profiles, but also giving them a wide range of freedom in their actions. To achieve this, we mobilize annotators from the outset, even before developing an initial annotation model (Pustejovsky et al., 2017). This early involvement allows us to capture a wide range of interpretations and insights, ensuring that the annotation process is informed by diverse viewpoints from the very beginning. In doing so, we aim to create a more robust and inclusive annotation framework that can better accommodate the complexity and variability of linguistic data.

3 Methods

3.1 Data and participants

Usually, annotation task guidelines are designed to be as precise and objective as possible. However, in this study, our objective was to explore subjectivity and to capture a wide range of perspectives on the same topic. Therefore, we intentionally limited the specific assignments provided to the annotators.

For this study, we selected eight short messages (ranging from 78 to 245 words) from various campaigns of "Micro Ouvert". "Micro Ouvert" is an internal tool developed and used by a large French company to collect employees' spontaneous opinions on a number of topics while ensuring complete anonymity. These eight texts reflect the insights of employees on topics such as their experiences with changes in the workspace, attending conferences, and their general motivation to work at the company. An example of such a message is presented in A.1 in its original French version, along with the English translation. All the language data used and collected are in French, and the experiments were conducted solely in French with our English translations provided in the paper.

We recruited four experts in mental workload, all of whom are academics in psychology with a focus on this topic, and 23 employees from the same company, including managers, human resources specialists, and other executives, particularly concerned by MWL in their team management roles. Indeed, it was crucial that the participants were motivated to perform the annotation tasks. These participants will hereafter be presented as members of the following 4 categories: 4 mental workload Experts (E), 8 Human Resources Specialists (HR), 12 Managers (M), and 3 Other specialists (O). The participants had no relationship with the experimenters, which prevented bias that could compromise the results. It is important to note that none of the participants in our study had prior experience with annotation tasks, neither experts nor employees. All participants were given information on the context of each message to be annotated and a possible clarification by the experimenters in case of misunderstanding. However, a significant difference between the experts and the employees is that the former were external to the firm and had no prior knowledge of the working context. In contrast, the employees were more familiar with the company's culture and the general content of the messages.

Having outlined the diverse participant profiles and message selection in the previous section, we now turn to the experimental protocol, which describes the methodology employed to capture the subjective interpretations of MWL.

3.2 Protocol

The participants were not informed in advance about the details of the experimentation but were only made aware of the subjects of our study and the estimated duration of the session (approximately 75 minutes). They were informed about the data collection process and they gave their written consent to participate in the study. The experimentation was conducted in the form of recorded interviews with two experimentalists (co-authors of this paper) and included two parts. The sessions lasted between 1 and 2.5 hours per participant. The final set of 8 texts was chosen after pre-testing with 5 participants, based on time demands and the observation of fatigue among the participants by the end of the study.

In the first part (which took an average of 40% of the session), the participants were assisted in defining the mental workload from their perception, during a semi-structured interview, using the explicitation techniques (Vermersch, 1994). The participants were encouraged to explain their representation of MWL without being directed towards a particular definition. Then, the experimentalists extracted keywords from their actual speech and submitted them to the participant for validation. We will refer to these labels as *prior labels*. It provides us with the initial set of data: the recorded definition of MWL as well as the keywords associated with this concept. The main purpose of this step is to define the labels for the next task: annotation.

In the second part of the task, the participants were asked to annotate eight short messages using the open-source text annotation tool *Doccano* (Nakayama et al., 2018). The tool was specifically configured to propose only their own prior labels. The participants had to annotate text parts illustrating an expression of mental workload, according to their judgment, without any constraint on the segments. The participants were also allowed to freely introduce new labels, which were then added into the annotation interface by the experimentalist. The labels used to annotate at least one segment are referred to as *used labels*. Additionally, the participants were allowed to assign multiple overlapping labels to their annotations. All annotators received the same sequence of texts, but the sequence was rotated so that each annotator began with a different text. This standard counterbalancing was made in order to minimize the learning effect and balance the performance among all the users for all the texts.

After each interview with each participant, we were able to collect the following data: 1) the list of prior and used labels 2) the annotated segments from the eight messages (start offset, end offset, and associated label).

At this stage, we did not rename or unify the labels that had similar meanings, and instead kept the exact formulation expressed by the participant. For example, we treated *priority* and *prioritization* as two distinct labels. However, the experimenters took care to consistently put down and unify the formatting of all repeated keywords and phrases across the sessions. We also ignored the polysemy in this study. The results of this methodology are presented in subsequent sections.

4 Analysis and results

To address the first research question concerning the similarity of individual representations of mental workload, we conducted a comparative analysis of the following components: labels, segments, relations between labels, segments, and user categories.

4.1 Labels

First, to identify patterns that indicate a shared understanding of MWL, we began with an examination of the prior labels (defined by the annotators during the first phase of the interview) and the used labels (actually employed in the annotation process). The maximum number of prior labels in the annotation task was 12, with a minimum of 5. On average, each participant defined 9 labels. In contrast, the maximum number of labels utilized during the annotation task was 15, while the minimum was 5, with an average of 10 labels per each participant. The average intersection between prior and used labels is 7.

We identified the labels that were commonly defined and/or used by different participants. At this stage of our analysis, we only consider strictly identical labels. The most frequently used labels are presented in Table 1, along with the number of annotators who proposed (center column) and used them (rightmost column). Labels in boldface correspond to those for which the frequency of use is at least 3 more that the frequency as prior labels.

We can see that the most frequent labels refer to individuality (*individual, personal*), demand (*temporality*, *pressure*), and task (*complexity, meaning*).

Labels	Annotators proposed	Annotators used
individual	7	7
temporality	7	6
meaning	1	5
complexity	4	5
stress	4	5
objectives	4	4
pro	4	4
professional/personal	4	4
uncertainty	1	4
context	4	4
ill-being	1	3
task	3	3
personal	4	3
permanence	4	3
environment	2	3
prioritization	3	3
time	2	3
pressure	2	3
powerlessness	0	3
recognition	0	3
volume	4	3

Table 1: The most frequently used labels (by at least 3 distinct annotators)

These three dimensions align with academic models of mental workload (Longo et al., 2022; Galy, 2016; Le Gonidec, 2022).

It is noteworthy that the individual feature of mental workload was among the first labels proposed and used by participants. While models of mental workload by Sweller, Galy, and Le Gonidec acknowledge this individual aspect, it is usually less emphasized in discussions between non-experts, especially when compared to more common aspects such as time management and task multiplicity. Indeed, the second most frequently used label corresponded to the temporal dimension of mental workload.

As indicated in bold in Table 1, some labels were frequently added by the participants during the annotation step. They reflect the fact that some aspects of mental workload were identified (or remembered) by the participants and could be considered as contingent to the topics of the selected messages. However, some of them *meaning*, *uncertainty* were also suggested as prior labels by other participants, and in any case, they were considered relevant for the annotation.

4.2 Label clusters

As mentioned earlier, many of the labels collected were semantically or formally close. To obtain a more global view of the dimensions considered by the participants, we employed a large language model (LLM) to propose a clustering of all labels.

We selected Claude Sonnet 3.5 (Anthropic, 2024) among other mainstream LLMs by prompting it to propose some categories and to regroup the 197 distinct used labels. The prompt is provided in A.2. We did not ask for a precise number of clusters but ended up with the 14 listed in Table 2, along with the number of labels and the number of participants (Annotat.) who used at least one of them for annotating. An LLM was chosen over a human annotator to ensure objectivity. The authors reviewed the associations and categories and were generally satisfied with the results. The choice of LLM over other clustering methods is defined by the fact that our approach focused on grouping labels that made sense together and, more importantly, on giving those groups clear names.

Cluster	Annotat.	Labels
Emotional and Psy. Aspects (I)	18	24
Relationships and Interactions (I)	18	19
Workload (W)	17	15
Environment and Context (A)	17	17
Temporality (W)	17	10
Organization and Management (A)	14	16
Cognitive and Mental Aspects (I)	13	16
Constraints and Difficulties (W)	13	15
Balance and Well-being (I)	13	14
Abilities and Skills (I)	11	11
Impact and Consequences (W)	11	12
Processes and Actions (A)	9	11
Adaptation and Change (A)	8	9
Management and Recognition (I)	5	8

Table 2: Label clusters and description according to Claude. IWA refers to the components of Galy's model

The resulting clusters align to the three main components of the IWA (Individual, Workload, Activity) model of mental workload Galy (2016). Therefore, we can consider that these induced categories confirm that the labels do not diverge from the models of mental workload, and that they can serve as a coarser-grained way to identify the qualitative behaviors of the participants.

4.3 Annotated segments

Having analyzed the labels, we now turn our focus to the annotated text segments. The average number of segments per annotator is 69, with an average length of 39 characters. The minimum number of segments recorded for a user was 13, while the maximum reached 248. The shortest segment had a length of 1 character (a question mark), compared to a maximum length of 1144 characters (i.e. an entire message). This variation was expected due to the lack of guidelines and constraints.

At this stage, we aimed to find specific segments of text that attracted the most attention from the participants, without considering the associated labels.

We extracted the most frequently annotated segments along with the number of users who annotated each segment. Although annotators had the option to overlap their annotations and assign multiple labels to the same text segment, we decided to consider only the number of users to avoid distorting the interpretation of markers across all users. We calculated, for each character position in each text, the number of different users who included it in at least one annotated segment, and then identified the contiguous characters that exceed a given threshold. Based on the inflection point in the curve displaying the number of segments, we selected a threshold of 14 different annotators (50% of them) and identified 52 different text segments. These segments consist of a variety of text units, ranging from single words (e.g., stress, meaning) to entire phrases (e.g., Reconnect with nature, with our environment, with humans). The complete list of these segments is provided in French in A.3.

After analyzing these segments, we noticed that the vocabulary containing words with negative connotations attracted the most attention. The annotators associated the text elements reflecting discomfort, overwhelm, overload, and disconnection with an increased mental workload.

Following a separate analysis of labels and segments, the next section explores the relationships between them.

4.4 Labels and Segments Similarity

Variability of annotation across annotators comes from both the labels used (intentional similarity) and the segments delimited (extensional). The latter similarity has been considered at the dataset level, as we have identified the main zones of interest in the target texts. On a finer grain, we aimed to identify the extent to which specific labels are used by different participants to tag the same text segments.

If we consider the set of text segments labeled L by participant P across the corpus, we can define the extensional similarity the set of segments labelled L' by another participant P' as the amount of overlapping. More precisely, we used the Jaccard index to measure the ratio between the number of characters (defined by their offsets) that the two sets have in common, to the union.

If $ext(L_P)$ is the set of characters (offsets) labeled as L by participant P (as one or several segments), we can define the extensional similarity between two labels from two participants (L_P and $L'_{P'}$ as the Jaccard index between the labeled text segments:

$$simext(L_P, L'_{P'}) = \frac{|ext(L_P) \cap ext(L'_{P'})|}{|ext(L_P) \cup ext(L'_{P'})|}$$

In other words, if the two participants used two labels (either different or identical) to tag the exact same parts of the target texts, $simext$ will be 1, while it will be 0 if there is no overlap.

We computed the $simext$ values for every pair of labels from two different users. We then focused on two specific subsets of pairs.

First, we considered the pairs of different labels with a high level of similarity ($simext > 0.4$, for a total of 169 pairs). We observed 15 cases where different annotators applied the exact same labels to the same segments. For example, on the same segment *leaving us in the dark* three annotators applied the label *uncertainty* independently. This is understandable, as the selected messages influenced the choice of labels.

Second, as expected, we found a number of synonyms used to tag the same text parts, for example *pause* and *respite*.

Third, and more interestingly, we found cases where the same segments were labeled with semantically related words, although not synonyms. For example, the labels could describe either the *causes* or the *consequences* of the same phenomenon. In this instance, the same segment of the text *a month before the show, and not 3 days before* was annotated by different persons using the label *temporality* for one and *to juggle* for the other. This means that one person describes increased MWL as a result of time constraints, while another perceives it as a consequence involving the need to juggle and manage multiple tasks simultaneously. Another example is when different users applied the labels *uncertainty* and *ability to reason* to the same text segment *leaving us in the dark*. In the first case, the annotator used a paraphrase of the segment itself to express the *cause*, while the other used a label for a *consequence*, indicating that this impacts the ability to reason. These cause-and-consequence designations are supported by participants' verbal expressions during the test. The first annotator stated: "This uncertainty is caused by the lack of answers that the person expected, as before, and that they could receive." In contrast, the second annotator said: "Leaving us in the dark means that there are no indicators or elements to be able to reason with, to face something."

This clearly indicates that although we found a number of indications that the annotators exhibit similar behavior, there are also variations in their points of view on this complex concept.

It is important to note that we didn't observe any contradiction in annotators' behavior: while there were complementary variations, there were no opposing opinions, similar to what has been observed in other related studies on perspectivism.

4.5 Overview of annotator behavior

In this last section, we examine variations among individual annotators to identify the main profiles and assess whether these variations are linked to the annotators' position and status.

We conducted a multidimensional analysis at the annotator level by computing the following variables:

- *NbDistinctLabels*: the number of distinct labels initially proposed by the participants during the first stage of the interview (before the annotation). Min:5, Max:15, Avg: 10.1.
- *UsedLabels*: the ratio of these prior labels that were actually used for annotation. Min: 41%, Max: 100%, Avg: 81%.
- *AddedLabels*: the ratio of used labels that were added by the participant during the annotation stage. Min: 0, Max: 67%, Avg: 30%.
- *NbSegments*: the total number of text segments produced by the participant during the annotation. Min: 13, Max: 248, Avg: 69.1.
- *TextCoverage*: the total amount of text (number of characters) included in at least one annotated segment. Min: 427, Max: 5065, Avg: 1987.
- *AverageOverlap*: the average number of segments in which an annotated text part is included (at the character level). Min: 1, Max: 2.7, Avg: 1.3.

These differences between the participant groups for each variable are illustrated in the boxplots in Figure 1. While there is no clear difference for

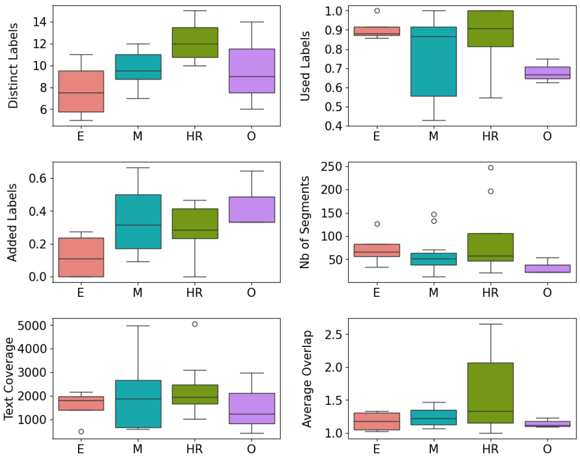


Figure 1: Boxplots of behavior variables across participants' categories (E=Mental workload experts, M=Managers, HR=Human Resources specialists, O=Others)

variables associated with text segment selection (number of segments, overlap, and text coverage), significant variations are observed in the choice of labels. Experts tend to have fewer labels, use all their initial proposals, and do not need any additional ones. HR specialists and managers exhibit roughly similar profiles, but the former use a larger number of labels. We note that the *Other* group is in the middle ground and remains inconclusive without additional participants.

To obtain a more global picture, we performed a principal component analysis (PCA) on the matrix representing each of the 27 participants across the 6 quantitative variables. The main factor map is shown in Figure 2. The variables are represented as blue arrows, and the participants are depicted as colored dots according to their group. The confidence ellipses (at the 95% level) are shown around the barycenter for each group. The first factor map is sufficient as it captures 70% of the total variance.

The first principal component (horizontal axis) is positively correlated with all variables except *AddedLabels*. On the right, there are the participants who produced a large number of segments, with a significant overlap and a high number of labels. The HR specialists are predominant, while the managers are located on the left side of the map. In other words, HR specialists exhibit a more dispersed annotation behavior with more segments and labels, in a more cumulative manner than the managers (who are less productive).

The second component (vertical axis) opposes participants who used a large amount of their ini-

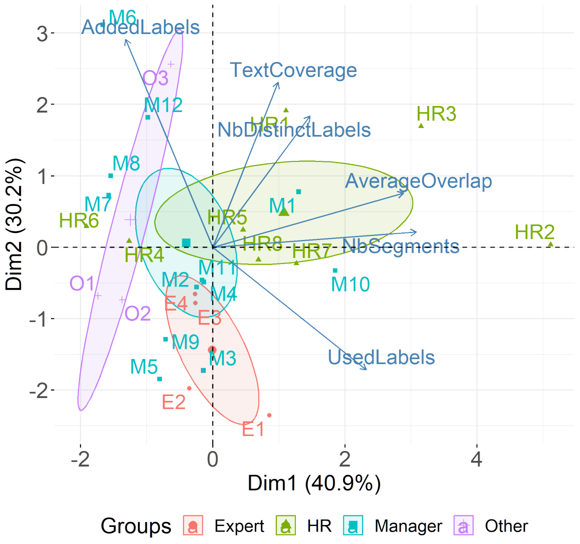


Figure 2: Principal Component Analysis of the participants

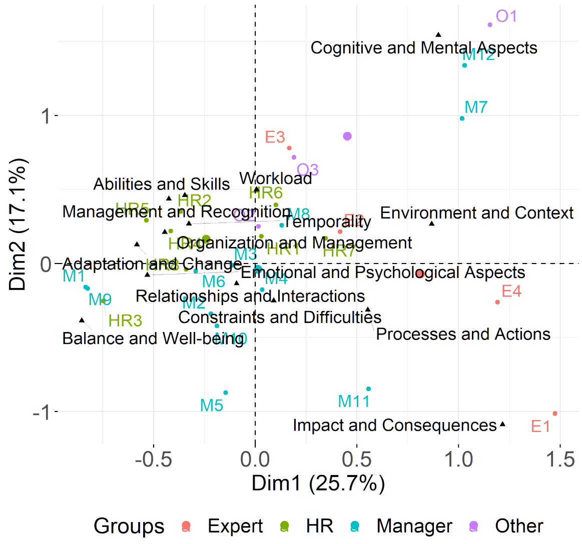


Figure 3: Correspondence Analysis: participants and label clusters

tially proposed labels (bottom) to those who had to provide additional labels in the course of the annotation process. It also appears that this distinction is correlated with the text coverage.

It is interesting to observe that the experts are very homogeneously located at the bottom of the factor map, even though none of them had prior experience with annotation and they come from diverse academic backgrounds. This seems to indicate that their knowledge allowed them to correctly anticipate the dimensions of the phenomenon, and that the labels they initially provided at the beginning of the interviews were both relevant and sufficient. Additionally, it appears that they produced a more focused set of segments, with less coverage and overlap than naive participants from other groups.

Our final analysis considers the clusters of labels that we requested an LLM to identify (see Table 2). We considered the number of labels from each cluster used by each participant (without considering the number or size of the corresponding segments) and performed a correspondence analysis. The first factor map is shown in Figure 3.

Here also, we can see that the experts exhibit a specific and coherent behavior. They all appear on the right side of the map, standing out with label clusters such as *Impact and consequences* and, to a lesser extent, *Processes and actions*. The second group that differs is the Other annotators, who focus on *cognitive and mental aspects*. On the left, Managers and HR professionals are positioned together, suggesting that they share similar preferences in the label categories used during annotation. The categories on this left side appear to be more individual-focused, encompassing aspects such as *Balance and Well-being*, *Relationships and Interaction*, *Adaptation and Change*. These clusters reflect individual experience and interpersonal dynamics within the context of mental workload.

These results seem consistent with the mission of managers and HR, more focused on individuals, while experts are more interested in mental workload processes.

Beyond this specific analysis, it appears that it remains possible to perform a qualitative analysis and to identify global tendencies in the annotations, even in the absence of specific guidelines.

5 Conclusion and future work

This study explored the subjective perceptions of mental workload through an innovative annotation method involving participants of diverse professional profiles, with different expertise related to the MWL concept. Regarding the formulated research questions, we can state that firstly, text annotation can serve various purposes, notably the analysis of different points of view on mental workload. Next, the analysis of the results showed that, even without clear instructions, people share a common representation of mental workload. Finally, despite this convergence in the representation, we observed differences in user behavior based on their professional roles. By allowing annotators to define their own labels and freely annotate text segments, we captured a variety of perspectives on MWL. We observed similarities in the annotated segments, labels, and groups, indicating a common representation of MWL, as we anticipated. Additionally, differences emerged, highlighting the influence of individual professional profiles on the perception of this topic. The comparison between experts' and non-experts' approaches allowed us to see differences in the process of identifying MWL elements in the text. Therefore, we gained a better understanding of the non-expert analysis and potentially considered new candidate facets in the MWL concept by leveraging employees' knowledge of the workplace context.

This holistic approach promotes richer and more representative annotation and can thereby improve models for analyzing and interpreting textual data. The collected data can be viewed as explicit, structured, exemplified, and tested individual models of the MWL. The same annotation protocol could be applied to topics beyond mental workload, enabling a more inclusive approach.

Since we recorded all the interviews and have a transcript of the participants' comments on their own actions (such as definition, reformulating labels, choice of segments, etc.) we possess an even richer dataset than what we have presented here, which requires further analysis and effort.

As part of a larger project, we are now considering the development of an automatic annotation process for mental workload in the messages. However, we are currently at the stage of defining the annotation scheme. This innovative approach opens the way to a better understanding of the linguistic and cultural nuances that influence the assessment of mental workload, while emphasizing the importance of integrating different perspectives to enrich textual data analysis models.

Limitations

One of the limitations of this study is the small dataset of texts, which displayed limited variability, as the messages primarily focused on a small number of topics. This limitation is caused by the collection of texts within a specific working environment, and further investigations are needed to generalize the findings to other contexts. Nevertheless, we aimed to compile a dataset from various campaigns to present annotators with messages on different topics, all still related to work.

Another limitation is that the participants were recruited from a population that was very busy at work, allowing them to dedicate only 1 to 2 hours to the study. This time constraint may have impacted their level of involvement in the process. However, all participants were engaged in the task and performed well in annotation, despite having no prior experience. We assess their performance based on the annotations they produced as well as the received feedback. All participants expressed significant interest in the task and demonstrated high motivation and self-awareness; however, they also noted that it was complicated and mentally resource-consuming.

References

Anthropic. 2024. Claude 3.5 sonnet.

Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *CEUR Workshop Proceedings*, volume 2776, pages 31–40. CEUR-WS.

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

Brad Cain. 2007. A review of the mental workload literature. Technical Report RTO-TR-HFM-121-Part-II, Defence Research and Development Canada Toronto, Toronto, Ontario, Canada.

Jose-Manuel Castillo, Édith Galy, and Pierre Thérouanne. 2023. Le technostress et sa relation avec la charge mentale en contexte professionnel. *Psychologie du Travail et des Organisations*, 29(4):197–213.

Berta Chulvi, Lara Fontanella, Roberto Labadie-Tamayo, and Paolo Rosso. 2023. Social or individual disagreement? Perspectivism in the annotation of sexist jokes. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Edith Galy. 2016. Approche intégrative de la charge mentale de travail : une échelle d'évaluation basée sur le modèle ica (individu – charge – activité). In *Actes du 51e Congrès de la Société d'Ergonomie de Langue Française (SELF)*, Marseille, France.

Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? Exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–27.

Nolwenn Le Gonidec. 2022. *Conceptualiser et évaluer la charge mentale de salariés dans un contexte d'usage d'outils numériques : Le cas d'une entreprise de télécommunications*. Ph.D. thesis, Université Côte d'Azur, France.

Luca Longo, Christopher D. Wickens, Gabriella Hancock, and P. A. Hancock. 2022. Human mental workload: A survey and a novel inclusive definition. *Frontiers in Psychology*, 13:883321.

Caroline Martin, Sylvain Hourlier, and Julien Cegarra. 2013. La charge mentale de travail : un concept qui reste indispensable, l'exemple de l'aéronautique. *Le Travail Humain*, 76(4):285–308.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

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. Association for Computational Linguistics.

James Pustejovsky, Harry Bunt, and Annie Zaenen. 2017. Designing annotation schemes: From theory to model. In *Handbook of Linguistic Annotation*, pages 21–72. Springer.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Narjes Tahaei and Sabine Bergler. 2024. Analysis of annotator demographics in sexism detection. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 376–383. Association for Computational Linguistics.

Pierre Vermersch. 1994. *L'entretien d'explicitation en formation continue et initiale*. ESF, Paris, France.

A Appendix

A.1 Sample message

French Original:

> Points à améliorer: - Avant SDLR: Faire en sorte que la Brand revoie les slides un mois avant le salon, et pas 3 jours avant. Les modifications imposées sont très importantes et nécessitent une surcharge de travail tant pour etre en conformité que pour modifier le discours pour les visiteurs. Demander des slides en français et en anglais plutôt que de nous laisser dans le flou (sur la partie anglaise).

English translation:

> Areas for improvement: - Before Research and innovation fair: Have the Brand review the slides a month before the show, and not 3 days before. The changes imposed are very significant and require a lot of work both to comply and to modify the presentation for visitors. Ask for slides in French and English rather than leaving us in the dark (about the English part).

A.2 Prompt used for clustering labels

The prompt was designed and submitted in French. Below is the translation to English by the authors:

I've interviewed people and asked them what mental workload means to them, and they've given me a list of terms that they associate with the notion of mental workload. Can you cluster these terms and group them according to their meaning? Each term must only go into one cluster and must not be repeated. I want you to use all the terms from the list in this task.
{List of 197 distinct labels in random order}.

A.3 List of text segments in French annotated by at least 14 different participants

'lourdeur de la logistique et des règles',
'ne sont pas inclus',
'points de synchronisation',
'tissage des liens sociaux',
'collaboration et le partage d'informations et d'idées',
'en-dehors des temps de travail',
'on fait quoi maintenant',
'ne sais pas trop',

'peut-être dû réfléchir un peu',
'sens',
'un peu perdus',
'on ne sait pas vers où, ni pourquoi',
'malaise général',
'pas promus',
'absolument',
'garder autant',
'fallait-il pas simplifier',
'donner du sens',
'dans les hiérarchies supérieures il n'y en a pas assez',
'numérique',
'les écrans ont raison de notre bien-être',
'Se reconnecter à la nature, à notre environnement, aux humains',
'moi',
'primordial',
'essentiel',
'revenir à des choses simples',
'reconnecter',
'non pas',
'des robots',
'nos émotions',
'réponse différente',
'tu as certainement un problème hormonal',
'gêne occasionnée',
'réponse différente',
'gêne',
'interrompre le travail',
'nous sommes très heureux',
'déshumanisé',
'perso',
'nauséabonde',
'très froid',
'heureuse',
'retrouver mes collègues',
'perdu une part de la bonne humeur, de l'ambiance',
'échange convivial',
'informel',
'un mois avant le salon',
'et pas 3 jours avant',
'imposées',
'très importantes',
'surcharge de travail',
'laisser dans le flou'

BoN Appetit Team at LeWiDi-2025: Best-of-N Test-time Scaling Can Not Stomach Annotation Disagreements (Yet)

Tomas Ruiz[1,2] **Siyao Peng**[1,3] **Barbara Plank**[1,3] **Carsten Schwemmer**[1,2]
[1]Ludwig Maximilian University of Munich, Germany
[2]Computational Social Sciences [3]MaiNLP & MCML
{t.ruiz,siyao.peng,b.plank,carsten.schwemmer}@lmu.de

Abstract

Test-time scaling is a family of techniques to improve LLM outputs at inference time by performing extra computation. To the best of our knowledge, test-time scaling has been limited to domains with verifiably correct answers, like mathematics and coding. We transfer test-time scaling to the LeWiDi-2025 tasks to evaluate annotation disagreements. We experiment with three test-time scaling methods: two benchmark algorithms (Model Averaging and Majority Voting), and a Best-of-N (BoN) sampling method. The two benchmark methods improve LLM performance consistently on the LeWiDi tasks, but the BoN method does not. Our experiments suggest that the BoN method does not currently transfer from mathematics to LeWiDi tasks, and we analyze potential reasons for this gap.

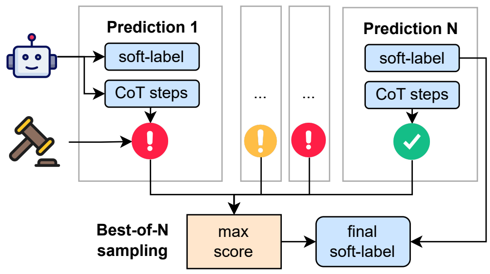


Figure 1: **Best-of-N sampling with step-wise scores.** For each problem a reasoning LLM generates N soft-labels and Chains-of-Thought (CoTs). Next, an LLM-as-a-judge scores each step in the CoT for correctness, and BoN selects the soft-label with the highest total score. *Takeaway*: Sampling multiple times increases the chances for a good prediction.

1 Introduction

Supervised learning typically assumes a single fixed label per example. However, prior work documents substantial interpretative variability in human annotations, with annotators often disagreeing on labels (Roß et al., 2016; Warner and Hirschberg, 2012; Baan et al., 2022), especially for subjective Natural Language Processing (NLP) tasks (Ovesdotter Alm, 2011). Plank (2022) and Cabitza et al. (2023) argue that this variability is informative rather than problematic and Röttger et al. (2022) suggests that variability should be explicitly integrated into the annotation processes.

The shared task **Learning With Disagreement** (LeWiDi) 2025 (Leonardelli et al., 2025) tackles this opportunity and provides four datasets with annotator-level metadata and label variation. We document the datasets in detail in subsection 3.6. The datasets support two different tasks: (1) Perspecivist task: Predicting the label of each individual annotator. (2) Soft-label task: Predicting the distribution of human annotations for a single problem instance. This distribution is known as a **soft-label**, or a *human judgement distribution* (Nie et al., 2020).

In the previous iteration of the LeWiDi shared task (Leonardelli et al., 2023), many teams trained encoder-based models like BERT (Devlin et al., 2019) directly on soft-labels. However, the innovations in generalist Large Language Models (LLMs) and the rise of "reasoning" capabilities (Wei et al., 2022; OpenAI, 2025; Yang et al., 2025; DeepSeek-AI et al., 2025) motivated us to answer the following question:

"*Can reasoning LLMs handle interpretative variability and annotation disagreement effectively at inference time?*"

To answer this question, we turn to **test-time scaling** methods, like BoN sampling, which improve the LLMs performance by spending more compute per problem (Cobbe et al., 2021; Shen et al., 2021). These methods have been very successful in mathematics and coding, but have not been applied yet to NLP tasks with annotation disagreement, as far as we know. In this paper, we

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 153–170
November 8, 2025 ©2025 Association for Computational Linguistics

take established test-time scaling methods and apply them to the LeWiDi tasks.

Our **contributions** are:

- A metric named prediction diversity, used to analyze the performance of test-time scaling methods on soft-label tasks. We show that it tracks problem difficulty on the LeWiDi tasks.
- We show that Model Averaging and Majority Voting consistently improve LLMs performance across all LeWiDi datasets.
- Finally, we show that BoN sampling with step-wise scores *does not work well* on the LeWiDi tasks, and analyze potential causes.

2 Related Work

2.1 Learning Interpretative Variability

Modeling the diverse perspectives that human annotators have on the same problems is important to prevent minority voices from being ignored (Leonardelli et al., 2021). Prior work on modeling annotator disagreement has explored various techniques, such as using separate model heads for each annotator (Mostafazadeh Davani et al., 2022), learning specific representations for annotators (Mokhberian et al., 2024), separating stable opinions from annotation mistakes (Gordon et al., 2021; Weber-Genzel et al., 2024), and using soft-labels to aid learning (Fornaciari et al., 2021; Uma et al., 2020). To evaluate models on soft-labels, Rizzi et al. (2024) propose using the Manhattan or Euclidean distance rather than the Cross-Entropy loss. In terms of quantifying the diversity of soft-labels, Singh et al. (2024) proposed the Jensen-Shannon Divergence in the context of ensemble classification.

2.2 Test-Time Scaling

Test-time scaling methods improve the performance of LLMs by spending more compute per problem instance. One approach is to refine an initial response iteratively with self-feedback (Madaan et al., 2023), or improve the response by following a set of rules (a *constitution*, Bai et al. 2022). Another common test-time scaling approach is **Best-of-N sampling**, where multiple solutions are sampled in parallel, and a verifier model scores or ranks the solutions to select the best one (Cobbe et al., 2021; Shen et al., 2021). The scores are computed based solely on the outcome (correct or incorrect) of the task (*Outcome Reward Model)*. But the scores can also be computed for the correctness of individual reasoning steps used to arrive at the answer (*Process Reward Model*). Lightman et al. (2024) showed that scoring individual steps in a Chain-of-Thought (CoT) for correctness, and discarding CoTs with faulty steps improves the performance of LLMs on the MATH dataset (Hendrycks et al., 2021). In their work, the scoring annotations were provided by humans. Follow-up work replaced the human scores with *automated scoring*, using either Monte Carlo (MC) sampling (Wang et al., 2024) or an LLM judge (*LLM-as-a-judge*, Zheng et al., 2023). Research by Zhang et al. (2025) and Zheng et al. (2025) showed that using LLMs to provide these scores generalizes well and is competitive with training a custom model.

A different approach that leverages diversity plus selection is *Mixture-of-Experts (MoE)*: multiple parallel expert subnetworks, with a gate that selects a few experts per input. Both MoE and test-time scaling are independent approaches that can be combined during model evaluation, *e.g.* as did Comanici et al. (2025) for the SWE-Bench (Jimenez et al., 2024).

3 Method

Nomenclature: A dataset is a collection of problem instances (*problem* in short). We sample a reasoning LLM N times to solve a problem. Each *sample* contains a *prediction* and a *CoT*. A prediction could be text, a soft-label or a list of integers (perspectivist task).

Our test-time scaling method is not novel but rather a combination of methods already established in the literature. Our innovation is to apply it to a new domain: the LeWiDi-2025 tasks. We refer to the method as *BoN sampling with step-wise scores*, or just *BoN sampling* in short. The method, shown in Figure 1, consists of three steps: (1) A reasoning LLM generates N samples for a problem. (2) A judge LLM scores each CoT-step in each sample for correctness. (3) We choose the sample with the best score for the final prediction. We explain all method details in subsection 3.1 and subsection 3.2. Table 1 is an overview of all methods we run experimentally, which include different baselines and benchmarks.

Method	Samples	Use CoT
Most Frequent	-	-
Simple Sampling	1	✗
Model Averaging	N	✗
Majority Voting	N	✗
BoN Oracle	N	✗
BoN + SWS	N	✓

Table 1: **Methods Overview.** The first two methods are baselines (subsection 3.4). The next two methods are our own benchmarks (subsection 3.5). The BoN Oracle (subsection 3.3) is a performance upper bound on our proposed method BoN + SWS (Step-Wise Scores, subsection 3.2). The models we submitted to the shared task are in *italic*.

3.1 LLM Setup

Prompts We prompt a reasoning LLM to solve the soft-label and perspectivist tasks directly. For example, in the soft-label task, we present the dataset (*e.g.* sarcasm detection), and instruct the model to predict the human soft-labels (snippet in Listing 1, full prompt in Listing 12).

Listing 1: Prompt Snippet (Soft-label Task)

```
Below is a context+response pair where
human annotators rated the sarcasm
level of the 'response' ranging from 1
(not at all) to 6 (completely). Please
guess the distribution of ratings and
output it in the final_response field
in JSON format [...]
```

In the perspectivist task, we instead instruct the model to predict the label for each annotator (snippet in Listing 2, full prompt in Listing 13).

Listing 2: Prompt Snippet (Perspectivist Task)

```
Below is a context+response pair [...].
Please guess the rating given by each
annotator and output them all in a
list, in the same order as the
annotators. [...]
```

We include a prompt section that explicitly instructs the model to reason about the *diverse perspectives and interpretations* that annotators could have. This improved performance by a small, but statistically significant margin, so we included it in all later experiments (see Appendix H).

3.2 BoN Sampling with Step-Wise Scores

In BoN sampling we score each of the N model samples and select the best for the final prediction. The score for a sample depends on the correctness of each step in its CoT. Our BoN sampling method borrows heavily from Lightman et al. (2024), so in Table 2 we summarize the differences and similarities between both.

	Lightman et al.	Ours
Domain	Math	LeWiDi Tasks
Model	GPT-4	Qwen3-32B
Scorer	Human(s)	LLM-as-a-judge
Reduction	Product	Mean
Sampling	Best-of-N	
Scores	bad=0, okay=0, good=1	

Table 2: **Comparison to Lightman et al. (2024).** First 4 rows are differences, last 2 rows are similarities.

First, we split the CoT into logical steps (details in Appendix B), and then score each step as either "great", "okay", or "bad" in line with Lightman et al. (2024). They used human annotations to train a scoring model, but we use an *LLM-as-a-judge* instead to provide the scores directly, as suggested by Zheng et al. (2025). The prompt for the LLM-as-a-judge is based on their scoring instructions (snippet in Listing 3, full prompt in Listing 14).

Listing 3: Prompt Snippet (LLM-as-a-judge)

```
Your goal is to grade an LLM's
step-by-step solution to a problem. The
model will often say things that look
ok at first, but will turn out to be
wrong on closer inspection - stay
vigilant!
Please mark each step with (great,
okay, bad). [...]
```

We follow Lightman et al. (2024) to convert the three scores to numbers (bad=0, okay=0, good=1), and average all the step-wise scores to compute a *prediction-level score*. Lightman et al. (2024) used a product reduction, but in our experiments, mean reduction outperformed product. This step-wise scoring is repeated for $N = 10$ model samples and their corresponding CoTs. We select the one with the highest prediction-level score as the final prediction.

Models Previous research in hate speech detection and Natural Language Inference (NLI) showed that *explanations* are useful to judge the plausibility and correctness of annotations and model predictions (Mathew et al., 2020; Jiang et al., 2023; Weber-Genzel et al., 2024). We hypothesize that reasoning about annotator disagreements also requires a *deliberative* and *explanatory* approach that considers multiple interpretations and weights their likelihoods. Therefore, we use a reasoning LLM for our experiments (Qwen3-32B, Yang et al. 2025). For LLM-as-a-judge we use a model from a different family besides Qwen3 (DeepSeek-R1-0528-Qwen3-8B, DeepSeek-AI et al. 2025). Sampling parameters are detailed in Appendix A.

3.3 Upper Bound on Performance

We determine the upper bound on performance of *any* BoN sampling method by computing a so-called **BoN oracle**. The BoN oracle is a hypothetical model that always selects the best prediction among N predictions (in our case, lowest distance). We compute the oracle in the *training* set by choosing the soft-label among the N predictions with the lowest distance to the human soft-label. However, for a dataset with unknown human soft-labels, we cannot compute the oracle. The oracle is an analytical tool to determine the BoN performance ceiling, rather than an algorithm to use in practice. The oracle soft-label p_o for a set of predictions P (size N) is defined as:

$$p_o = \arg\min_{p \in P} W(p, p_h) \tag{1}$$

where p_h is the human soft-label. We compute it for both the soft-label and perspectivist tasks.

3.4 Baselines

The LeWiDi 2025 shared task proposed the **Most Frequent Baseline**. In the soft-label task, this is the mean label value for each label across all training problems. In the perspectivist task, it is the most frequent label for each individual annotator. Our own basic baseline is the performance of the LLM without any test-time scaling, *i.e.* with a single sample per problem ($N = 1$). We call this **Simple Sampling**.

3.5 Test-Time Scaling Benchmarks

BoN sampling uses a lot more compute per problem than Simple Sampling (N times more). To benchmark BoN sampling fairly, we compare it with two test-time scaling algorithms that also create a single prediction out of N predictions.

Soft-label task We benchmark against *Model Averaging*, where all N soft-labels p^n are averaged into a single soft-label $\bar{p}$. The resulting soft-label $\bar{p}$ is a valid probability distribution. Each entry i of $\bar{p}$ is defined as:

$$\bar{p}_i = \frac{1}{N} \sum_{n=1}^{N} p_i^n \tag{2}$$

Perspectivist task We benchmark against *Majority Voting*, where we sample the model N times per problem, each prediction resulting in a label per annotator, and then select the most frequent label (within the N predictions) for each annotator.

3.6 LeWiDi-2025 Datasets

We report the datasets as provided by the LeWiDi-2025 shared task. All datasets provide some level of annotator-level metadata like gender, age, nationality, education and more.

The Conversational Sarcasm Corpus (CSC): The CSC dataset by Jang and Frassinelli (2024) is a dataset for sarcasm detection with 7,036 entries (5,628 train, 704 dev, 704 test). Each entry consists of a context+response pair, where the reponse is rated for sarcasm on a 6-point Likert scale, by either 4 or 6 annotators.

The MultiPico dataset (MP): The MP by Casola et al. (2024) is a dataset for irony detection with 18,778 entries (12,017 train, 3,005 dev, 3,756 test). Each entry consists of a post-reply pair from Twitter and Reddit, and the reply's irony is rated as either ironic (1) or not ironic (0) by between 2 and 21 annotators.

The Paraphrase Detection dataset (PAR): The Paraphrase is a dataset by the MaiNLP lab[1] for paraphrasing detection with 500 entries (400 train, 50 dev, 50 test). Each entry has two questions from Quora Question Pairs (QQP), and annotators rate how strongly the questions are paraphrases of one another from -5 to 5. Each entry is rated by 4 annotators.

The VariErr NLI dataset (VEN): VariErrNLI by Weber-Genzel et al. (2024) is a dataset for NLI with 500 samples (400 train, 50 dev, 50 test). Annotators can assign any and multiple of the NLI categories (entailment, contradiction, neutral) for each entry. Each entry is annotated by 4 annotators.

3.7 Metrics

Soft-label Task As suggested by the LeWiDi task, we report *Manhattan Distance* for the MP and VEN datasets, and *Wasserstein Distance* for the CSC and PAR datasets. Both distances are exactly equivalent when applied to binary datasets (Rizzi et al., 2024). The Wasserstein distance measures the minimum "work" needed to transform one

[1]https://mainlp.github.io/

probability distribution into another, where "work" equals the amount of mass moved times the distance.

Perspectivist Task For the perspectivist task, we report *Error Rate* (1 - accuracy) for the MP and VEN datasets, and *Absolute Distance* for the CSC and PAR datasets. We divide the Absolute Distance by the range of the Likert scale, in line with the LeWiDi organizers. Both metrics are exactly equivalent when applied to binary datasets.

Prediction Diversity BoN sampling requires a diverse set of predictions for each problem. Otherwise, if all predictions were the same (or very similar), it would not matter which one is selected, and BoN sampling would provide no improvement over Simple Sampling. Therefore, we quantify the variability of the soft-labels across the N predictions for each problem, and call this the *prediction diversity*.

We implement this as the average pair-wise distance between all N soft-labels for a single problem. For the LeWiDi datasets, we use the Wasserstein distance because it can capture distances in Likert scales. We do not compare soft-labels to themselves, because the distance is 0. This is why we divide by $N(N-1)$ rather than by N^2. The formula for diversity D is:

$$D(P) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i}^{N} W(p^i, p^j) \quad (3)$$

where P is the set of N soft-labels, and $W(p^i, p^j)$ is the Wasserstein distance between the soft-labels p^i and p^j. Note that measuring the diversity of a set of predictions P is different from measuring the spread of a single soft-label (*i.e.* measuring the entropy of the soft-label).

Problem Difficulty Classically, problem difficulty is measured as the percentage of correct answers over N attempts. For soft-label tasks, it can instead be defined as the distance between predictions and human soft-label across N attempts. In the LeWiDi task, we posit a relationship between prediction diversity and problem difficulty: low diversity arises when the model perceives no ambiguity (the problem is easy or only one interpretation is considered), while high diversity arises when multiple plausible interpretations exist and the model's N predictions vary. Since both prediction diversity and distances (Wasserstein, Manhattan) are computable, their correlation is empirically measurable.

4 Results

In Table 3 we summarize the results on the test, taken from the LeWiDi leaderboard[2], since we have no access to the test set ground truth. In Table 4 we present a performance overview of all methods for the LeWiDi datasets. We did not train on the train set, so we used it for evaluation.

	Task	
Dataset	Soft-label (↓)	Perspectivist (↓)
CSC	0.928	0.231
MP	0.466	0.414
PAR	1.797	0.228
VEN	0.356	0.272
Avg. Rank	5th	6th
Out of	15	11
Method	Model Averaging	Majority Voting

Table 3: Results on the **test** set of the LeWiDi datasets (lower is better). Values are from the LeWiDi leaderboard. We submitted our best performing methods, Model Averaging and Majority Voting.

4.1 Test-Time Scaling Benchmarks

We first report the performance of all methods except BoN sampling. The orange bar in Figure 2 is the performance *Simple Sampling* with Qwen3-32B. It outperforms the *Most Frequent Baseline* on 3 out of 4 datasets in the soft-label task, but only in 2 out of 4 datasets in the perspectivist task. The test-time scaling benchmarks (green) are *Model Averaging* and *Majority Voting*. Both methods consistently improve performance over Simple Sampling across datasets and tasks. We discuss the effects of Model Averaging on soft-label entropy in Appendix E and compare it with naive soft-label smoothing. The *BoN oracle* (red) is meant to show the performance ceiling of any BoN sampling method. Its strong performance indicates that, at least *theoretically*, a good BoN sampling method can achieve very good performance on the LeWiDi tasks.

4.2 Best-of-N Sampling with Step-Wise Scores

The BoN sampling method has inconsistent performance in the LeWiDi datasets, as shown in Figure 3. Performance is often flat with the number of samples N, or varies wildly with judge model. *E.g* in the MP dataset, the Deepseek judge is consistently *worse* (higher distance) than Simple Sampling on both tasks (soft-label and perspectivist). BoN sampling is only competitive with the benchmarks (red

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

Method	Soft-label Task CSC	Par	MP	VEN	Perspectivist Task CSC	Par	MP	VEN
Most Frequent Baseline	1.14	2.89	0.26	0.27	**0.21**	0.36	**0.30**	0.33
Simple Sampling	1.00	1.96	0.26	0.22	0.24	**0.25**	0.43	0.32
Model Averaging	**0.91**	**1.78**	**0.24**	**0.20**	-	-	-	-
Majority Voting	-	-	-	-	0.23	**0.25**	0.40	**0.30**
BoN Sampling + SWS	1.01	1.93	0.26	0.22	0.24	**0.25**	0.42	0.32
BoN Oracle	0.51	1.29	0.11	0.11	0.15	0.18	0.18	0.16
Metric (↓)	Wasserstein		Manhattan		Abs. Dist.		Error Rate	

Table 4: Results on the **train** set of the LeWiDi datasets. In **bold** is the best performing method by column. BoN sampling underperforms the test-time scaling benchmarks, even though the BoN oracle suggests a high performance ceiling. We submitted to the LeWiDi shared task the Model Averaging (soft-label task) and Majority Voting (perspectivist task) methods, since they perfomed best.

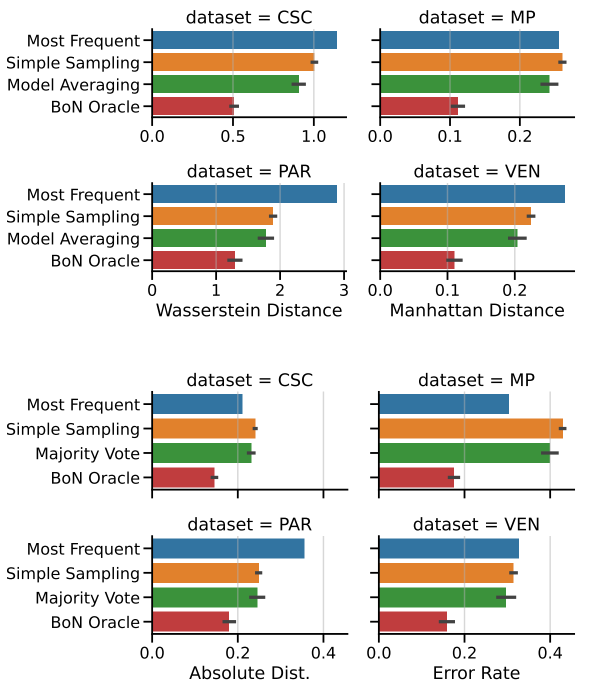


Figure 2: **Test-time Scaling Benchmarks**. Top: soft-label task. Bottom: perspectivist task. Distance metric on the x-axis (lower is better). Model Averaging and Majority Voting (green) are consistently better than Simple Sampling (orange) in the soft-label and perspectivist tasks, respectively.

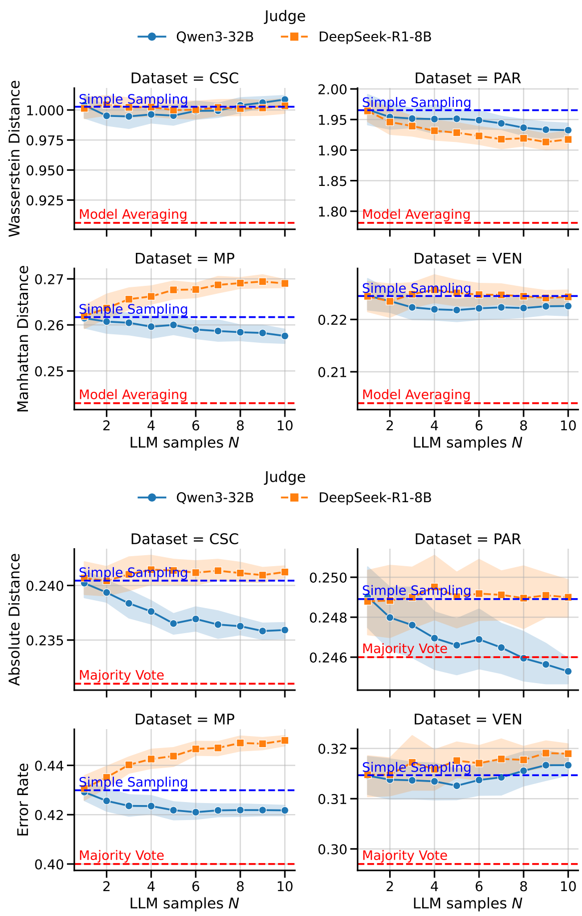


Figure 3: **Best-of-N Sampling on LeWiDi Tasks**. Top: soft-label task. Bottom: perspectivist task. Distance metric on the y-axis (lower is better). Higher N should lead to better performance, but does not. No consistent pattern emerges across datasets and tasks. In red are the test-time scaling benchmarks, which BoN generally does not beat. The shaded areas show the 0.25 and 0.75 quantiles.

horizontal lines) in a single case (perspectivist task, PAR dataset, Qwen3-32B judge). These inconsistent results raise the question why step-wise scoring is not effective in the LeWiDi tasks. For the BoN sampling numbers in Table 3, we report the Qwen3-32B judge, because it performs slightly better than the Deepseek judge on the perspectivist task. For the LeWiDi shared task, **we submitted the predictions for Model Averaging and Majority Voting**, rather than BoN sampling.

4.3 Prediction Diversity

Back to the LeWiDi tasks, we empirically observe that prediction diversity correlates with model performance (Figure 4): diversity increases for difficult problems and decreases for easier ones. For analysis, we binned prediction diversity into five quantiles (but the trends hold for any number of bins). We document the distribution of prediction diversity across datasets in Appendix F. Prediction diversity strongly affects test-time scaling methods, as shown in Figure 5: the BoN oracle performance (the upper bound for any BoN sampling method) increases with diversity. The same applies for Model Averaging.

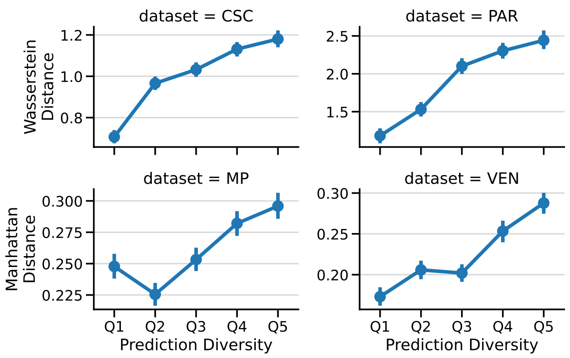


Figure 4: Model performance (lower is better) varies with prediction diversity and is related to the difficulty of the problem.

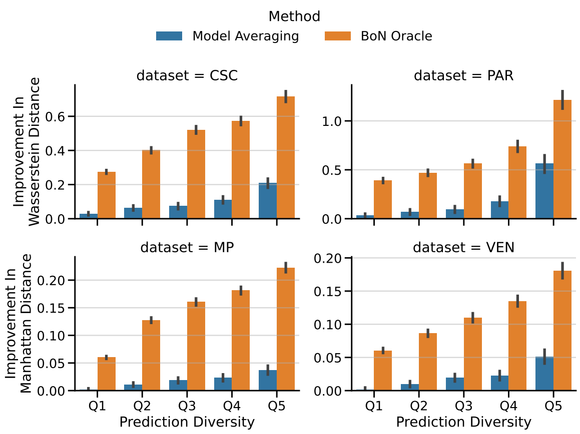


Figure 5: High prediction diversity leads to better performance of test-time scaling methods. Both the upper bound (BoN oracle) and Model Averaging benefit from higher prediction diversity. The y-axis shows the improvement over Simple Sampling.

Table 5 shows that Model Averaging achieves a significant fraction of the theoretical performance gains dictated by the BoN oracle. For example, in the top quantile of the PAR dataset, Model Averaging achieves 46% of the performance gains of the BoN oracle.

	Prediction Diversity				
Dataset	Q1	Q2	Q3	Q4	Q5
CSC	0.11	0.16	0.14	0.19	**0.29**
PAR	0.09	0.14	0.17	0.24	**0.46**
MP	0.03	0.09	0.12	0.13	**0.17**
VEN	0.03	0.11	0.18	0.18	**0.27**

Table 5: Fraction of the BoN oracle performance gains that Model Averaging achieves for different datasets and prediction diversities.

5 Discussion

5.1 BoN Sampling Underperformance

We were surprised by the underperformance of BoN sampling in the LeWiDi datasets. To verify that we had not made a mistake in our implementation of BoN sampling, we ran our method on two math datasets (PRM800K and AIME), as shown in Figure 6.

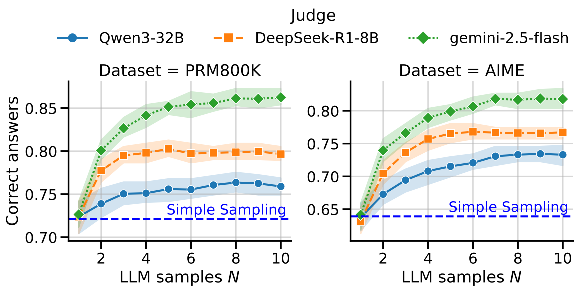


Figure 6: **Best-of-N Sampling in Mathematics**: The performance of BoN sampling (Correct answers, higher is better) improves with the number of samples N and with stronger judges. The shaded area shows the 0.25 and 0.75 quantiles: improvements are consistent.

The results are in line with Lightman et al. (2024). Using the best judge and $N = 10$ samples, the rate of correct answers jumps by 14% on PRM800K and by 18% in AIME. More information about both datasets is in Appendix C.

Why is BoN sampling effective in math, but not in the LeWiDi tasks? We think the LeWiDi tasks are not inherently harder or more intractable than math problems. The gap we observe is a failure of *cross-domain generalization*. For example, we observed that the shift in domain introduces unexpected side-effects:

1. We found qualitative evidence of the LLM being **more vague** in its formulation of CoT

steps in LeWiDi tasks (see Appendix I). When steps are vague, it is harder for a judge to discriminate between good and bad steps. During *post-training*, the Qwen3 model was likely never rewarded for summarizing precise arguments around interpretative variation and different perspectives. In contrast, we know that Qwen3 has been post-trained to reason on *"[...] math, code, logical reasoning, and general STEM problems."* (Yang et al., 2025). We find same in the technical reports for Deepseek R1 and Gemini-2.5 (DeepSeek-AI et al., 2025; Comanici et al., 2025).

2. We empirically observe that LLMs and judges both spend a higher **compute budget** (*i.e.* they produce more tokens) on reasoning when solving the mathematical tasks than on the LeWiDi tasks as shown in Figure 7. Since reasoning capabilities are learned during post-training, we hypothesize that this difference is also caused by the standard post-training recipe.

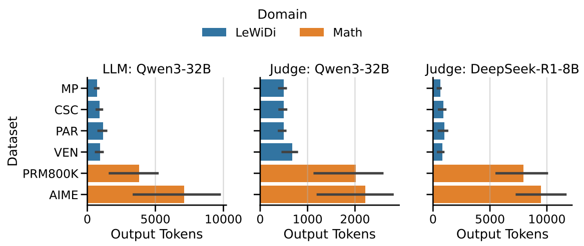


Figure 7: **Compute Budget** used by LLM and judges on different domains. Error bars are the 0.25 and 0.75 quantiles: they show large variability in output length. The models invest an **order of magnitude** more compute budget into solving AIME problems than in the LeWiDi tasks. Both Qwen3 and Deepseek-R1-8B show this bias.

5.2 Logical Steps in LeWiDi Tasks

One might argue that step-wise scoring requires a clear boundary between correct and incorrect steps, which is lacking in tasks with strong interpretative variation. We argue against this for two reasons:

1. Mathematical problem solving is also not always clear-cut. Lightman et al. (2024) show many steps add no insight or progress, leading them to use an "okay" label alongside "great" and "bad".

2. LeWiDi tasks define correctness precisely (*e.g.*, Wasserstein Distance 0). Steps that are logical, plausible, and advance a prediction are "great", while vague or unsound steps are "bad". Previous perspectivist research has also leveraged explanations to judge the validity of annotations (Weber-Genzel et al., 2024).

We see no theoretical conflict between perspectivism and step-wise scoring. Rather, adjusting the method to incorporate perspectivist principles is an avenue for future work. For example, using different step labels like "plausible", "implausible", "vague", etc.

6 Conclusion

We present a systematic evaluation of three test-time scaling methods on the LeWiDi tasks. Our key findings are: (1) Our prediction diversity metric correlates with test-time scaling performance and problem difficulty on the LeWiDi soft-label task. (2) Model Averaging and Majority Voting consistently improve LLM performance across the LeWiDi tasks. (3) BoN sampling with step-wise scores does not transfer from the domain of mathematics to the LeWiDi tasks, potentially due to vague reasoning steps and lower reasoning compute used. We hypothesize that this difference is caused by the post-training recipes of current reasoning LLMs, which lean towards mathematical and logical reasoning. The performance on datasets with annotation disagreements could potentially be improved by including similar tasks in the post-training recipe.

Limitations

We articulated the limitations of BoN sampling with step-wise scores in the LeWiDi tasks. We do not explore prompt optimization thoroughly, because we think that methods should be robust over different prompts. In terms of the prediction diversity metric, we suggest that authors evaluate the correlation with problem difficulty on their own datasets, since we showed an empirical rather than theoretical relationship.

Acknowledgements

This research is funded by the Bavarian Research Institute for Digital Transformation (bidt).

References

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

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

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*.

W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Symposium on Operating Systems Principles*.

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.

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.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. In *AAAI Conference on Artificial Intelligence*, pages 14867 – 14875.

Negar Mokhberian, Myrl Marmarelis, Frederic Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2024. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, Mexico City, Mexico. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

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.

OpenAI. 2025. OpenAI o3 and o4-mini System Card. Technical report, OpenAI. Technical report — system card.

Cecilia Ovesdotter Alm. 2011. Subjective natural language problems: Motivations, applications, characterizations, and implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, Portland, Oregon, USA. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 Röttger, 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.

Björn Roß, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the european refugee crisis.

Jianhao Shen, Yichun Yin, Lin Li, Lifeng Shang, Xin Jiang, Ming Zhang, and Qun Liu. 2021. Generate &

rank: A multi-task framework for math word problems. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2269–2279, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Avyav Singh, Ekaterina Shutova, and Helen Yannakoudakis. 2024. Learning new tasks from a few examples with soft-label prototypes. In *Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024)*, pages 215–236, Bangkok, Thailand. Association for Computational Linguistics.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. 8:173–177.

Hemish Veeraboina. 2023. Aime problem set 1983-2024.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. 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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10495–10516, Vienna, Austria. Association for Computational Linguistics.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. ProcessBench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1024, Vienna, Austria. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Sampling Parameters

Yang et al. (2025) suggest two different parameter configurations for Qwen3: for thinking and non-thinking modes. In early expriments we found almost no difference in performance between both configurations, but observed less variation with the non-thinking configuration, so we used thoese parameters in our experiments: top-k=20, top-p=0.8, temperature=0.7, presence-penalty=1.5. We used the same parameters for the Deepseek-R1-8B model. For Gemini-2.5-flash we used the default parameters documented in Google's documentation: top-k=64, top-p=0.95, temperature=1.0[3].

B Splitting the CoT into Steps

To score each step of a CoT for correctness, it must be first split into steps. We instruct the model to answer using a structured format (JSON) with separate fields for the prediction and the CoT steps, as shown in Listing 4.

Listing 4: Output format for the LLM

```
{
  'steps': [
    '<step 1 text>',
    '<step 2 text>',
    ...
  ],
  'final_response': '<text>'
}
```

We found that this approach to get logical steps is more robust than two alternatives: (1) Using *string matching* (*e.g.* on double line breaks) to split a CoT into steps, because it produces overly granular, incoherent steps, where *e.g.* a bulleted list becomes a step on its own. (2) Using a separate LLM to *reformat* the CoT into logical steps, because the reformatting model sometimes rephrases and truncates the original CoT instead of only reformatting it. We think that the original LLM is best positioned to split its own reasoning process into coherent, logical steps.

One might argue these steps are *constructed ex-post* and do not reflect the model's true reasoning. However, during a math exam, students are allowed to sketch ungraded work on separate sheets, and then present a clean step-by-step solution. We follow this same principle, and our math experiments show that the ex-post steps are expressive enough to discriminate good and bad reasoning.

[3]https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro

C Mathematical Datasets

Our method for providing step-wise scores is *completely automated* and requires no human annotations for the CoT steps at all. As **a sanity check** that our test-time scaling implementation is correct, we also include in our BoN evaluation two datasets with mathematical problems, where we expect step-wise scoring to perform very well. The datasets are: (1) High-school math problems and solutions compiled in the **PRM800K** dataset by Lightman et al. (2024). The problems are originally from the MATH dataset by Hendrycks et al. (2021). (2) Mathematical problems given to the top 2.5% to 5% of high-school students in the US from the American Invitational Mathematics Examination **(AIME)** compiled by Veeraboina (2023) and ranging from 1983 to 2024. The AIME problems are generally more difficult than those in PRM800K. Many math problems are solved by Qwen3-32B in 10/10 samples, which makes BoN sampling unnecessary. We skipped these problems in our BoN evaluation, which is why the horizontal line for Simple Sampling is relatively low in Figure 6.

D LLM Compliance

When using an LLM with structured output, we need to measure its adherence to the output format of the prompt. We call this **compliance**. As we show in Table 6, the compliance level varies by dataset and by task.

Dataset	Perspectivist	Soft-label
MP	100.0	100.0
CSC	100.0	99.3
VEN	100.0	93.9
PAR	100.0	86.2

Table 6: Percentage of compliant predictions sorted by dataset from highest to lowest. The PAR soft-label task is difficult because the weight of 11 classes must sum to 1.0.

We observe near-perfect compliance for the CSC and MP datasets. The VEN dataset has lower compliance because of the nested strucure of the predictions (one for each NLI category). The lowest compliance is in the PAR datasets, which has 11 classes (-5 to 5, including 0). We found that the LLM outputs correct JSON for PAR, but often the soft-labels did not sum exactly to 1. We experimented with enforcing **structured outputs** in vLLM, but initial experiments showed that the LLM would sometimes output infinite newline characters until it reached the output token limit, which is valid JSON, so we dropped this constraint.

E Model Averaging and Entropy

Model Averaging has an **adaptive** flattening effect on the model's soft-labels: When the model identified a consensus interpretation (regime of low prediction diversity), Model Averaging keeps soft-labels intact (*e.g.* peaky). And when the model's answers are diverse, Model Averaging flattens the soft-labels, which has a hedging effect. We compare Model Averaging with a naive **smoothing** method, which flattens a soft-label by averaging it with the uniform distribution, therefore increasing the entropy of the soft-label.

Figure 8 shows the entropy of the soft-labels for Qwen3-32B, for different datasets and sampling methods. It shows that smoothing the soft-labels does not automatically improve the model performance and that Model Averaging is much more adaptive than the naive smoothing.

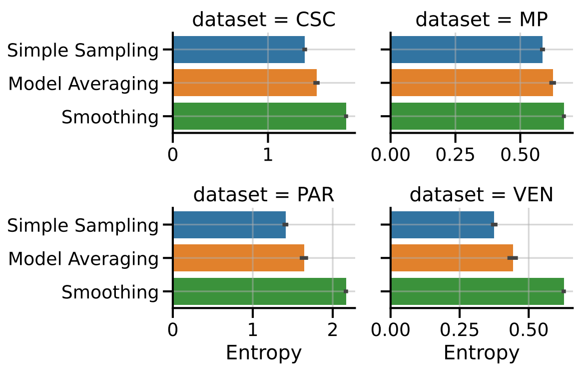


Figure 8: **Entropy of soft-labels**: We observe that both smoothing (green) and Model Averaging (orange) increase the entropy of the soft-labels, but only Model Averaging improves the model performance.

F Prediction Diversity

The distribution of prediction diversity by dataset is shown in Figure 9. We observe that it is distributed with a single peak in the center, and sometimes has a right tail.

G Compute Infrastructure

We use the vLLM engine (Kwon et al., 2023) to run the models, because of its high throughput, which help us compute N samples per example in parallel. vLLM can also be configured to parse the CoT and return them separately from the final answer. All

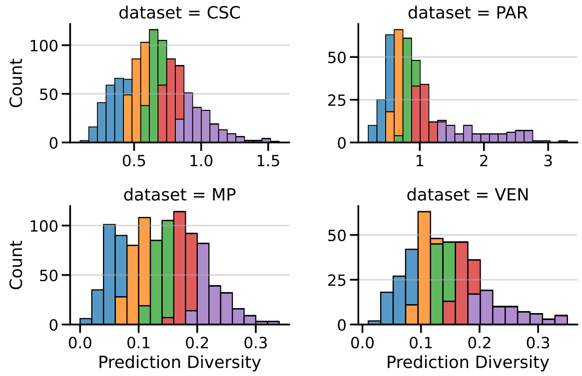


Figure 9: Distribution of prediction diversity by dataset. The distributions follow a normal-like distribution, and the PAR dataset shows a longer tail to the right. The colors indicate the quantiles of the distribution.

our experiments are run on a single NVIDIA H100 GPU, except for the Qwen3-32B model, which is run on two GPUs. We called Gemini-2.5-flash over the Google Cloud API.

H Prompt Ablations

We created two prompt variations that could potentially affect performance for interpretative tasks: (1) One variant provides a *dictionary definition*[4] of sarcasm (for CSC), or irony (for MP). (2) The second variant explicitly instructs the model to consider different *perspectives and interpretations*.

Listing 5: Prompt Section Defining Sarcasm

```
Use this definition for sarcasm: "The
use of remarks that clearly mean the
opposite of what they say, made in
order to hurt someone's feelings or to
criticize something in a humorous way"
```

Listing 6: Prompt Section to Consider Perspectives

```
Think about the perspectives that
different annotators might have and how
they could potentially interpret the
post-reply pair.
```

We perform an ablation analysis to determine the impact of the two prompt variants: (1) first, we remove the prompt section that defines irony and sarcasm and (2) we remove the prompt section about considering diverse perspectives.

As shown in Table 7 **for ablation 1**, we observe mixed effects: In the CSC dataset, including the definition of sarcasm improves performance, while in the MP dataset, including the definition of irony

[4]Online dictionary: https://dictionary.cambridge.org/

decreases performance. We compute the 95% confidence interval of mean performance using the bootstrap method to rule out the possibility that performance differences between prompts are a sampling artifact.

For ablation 2, we observe that prompting the model to consider diverse perspectives improves performance in 3 out of 4 datasets (CSC, MP, VEN). In the PAR dataset, performance is not affected by the prompt section on diverse perspectives.

I Vague CoT Steps

Below is a qualitative comparison of two responses to the same problem by Qwen3-32B. Listing 7 describes the problem (ID=637 of the CSC dataset on sarcasm detection). Listing 8 shows a response with very vague CoT steps, which are not strictly wrong, but are so general that they could be applied to any or all problems. Listing 9 shows a different response with very specific CoT steps that directly refer to the problem statement and are easier to judge.

Note that these are the structured logical steps, rather than the raw CoT. The raw CoT for the vague answer is in Listing 10 and the one for the concrete answer is in Listing 11. Neither of the raw CoT are vague, so there is no reason why the model should generate vague logical steps from either of them. We observed this "vagueness" behavior particularly in the CSC dataset.

Listing 7: Problem Description (Sarcasm Detection)

```
{
    "context": "You and Steve went out
    last night and drank a lot. The
    next day, Steve excitedly says that
    he fell in love with the girl he
    met last night and that he is going
    to quit his job and join her on a
    backpacking trip.  Steve says,
    \"this girl is totally amazing and
    I'm not gonna miss the opportunity
    to be with her. I'm quitting my job
    and traveling with her!\"",

    "response": "Man, you are fucking
    serious right now? You're going to
    quit your job and travel with some
    girl you just met last night while
    drunk? No girl is worth quitting
    your job and screwing your life
    over, especially not a girl you
    just met last night."
}
```

Listing 8: Vague Logical Steps

```
"Analyze the context and response to
identify elements of sarcasm based on
the given definition.",

"Determine if the response uses remarks
that mean the opposite of what they say
to criticize or hurt someone
humorously.",

"Assess the likelihood of different
annotator interpretations, considering
variations in perception of sarcasm.",

"Assign probabilities to each rating
(1-6) based on the analysis, ensuring
the sum equals 1.0."
```

Listing 9: Specific Logical Steps

```
"Analyze the context and response to
determine if sarcasm is present. The
response directly criticizes Steve's
decision without evident irony or mock
praise.",

"Consider that the response uses
rhetorical questions and direct
statements to express disapproval,
which aligns more with blunt criticism
than sarcasm.",

"Assess potential annotator
perspectives: most may view the
response as non-sarcastic (rating 1),
while others might detect mild sarcasm
in the tone (ratings 2-3).",

"Distribute probabilities accordingly,
allocating higher weights to lower
sarcasm ratings and minimal
probabilities to higher ones."
```

J Prompts

	Dataset							
	CSC		MP		PAR		VEN	
Prompt	low	high	low	high	low	high	low	high
Default	**1.002**	**1.013**	0.256	0.258	1.838	1.920	**0.217**	**0.227**
-Definition (ablation 1)	1.057	1.071	**0.246**	**0.248**	-	-	-	-
-Perspectives (ablation 2)	1.027	1.040	0.263	0.265	1.840	1.924	0.236	0.247

Table 7: Two Prompt Ablations: The numbers are mean performance in terms of Wasserstein Distance (lower is better). The column "low" and "high" are the bounds of the 95% confidence intervals on the performance computed with the bootstrap method. Ablation 1 shows that including/excluding the definition of irony / sarcasm has mixed effects. Ablation 2 shows that not instructing the model to consider diverse perspectives has negative consequences in 3 out of 4 datasets.

Listing 10: Raw CoT for Vague Answer

```
Okay, let's tackle this problem step by step. First, I need to understand the task. The user provided a
context and a response pair, and I have to guess how human annotators would rate the sarcasm level of the
response on a scale from 1 to 6. The final answer should be a probability distribution over these ratings,
with each probability summing to 1.0.

The definition given for sarcasm is important here: using remarks that mean the opposite to hurt or
criticize humorously. So, I need to analyze if the response fits this definition and how different
annotators might perceive it differently.

Looking at the context: Steve met a girl after drinking a lot and decided to quit his job to travel with
her. The response is a friend (the speaker) reacting to Steve's decision. The response says, "Man, you are
fucking serious right now? You're going to quit your job and travel with some girl you just met last night
while drunk? No girl is worth quitting your job and screwing your life over, especially not a girl you just
met last night."

First, I need to determine if there's any sarcasm in the response. Sarcasm often involves irony or mock
praise. In this case, the speaker is clearly criticizing Steve's impulsive decision. The phrases like
"you're fucking serious" and "screwing your life over" indicate frustration or disapproval. However,
there's no direct use of words that mean the opposite. Instead, the speaker is being straightforward in
their criticism. They are pointing out the irrationality of Steve's actions without obvious irony.

But maybe some annotators could interpret parts of the response as sarcastic. For example, when the speaker
says "especially not a girl you just met last night," they might be emphasizing the absurdity, which could
be seen as a form of understatement or rhetorical question. However, under the given definition, sarcasm
requires the remark to clearly mean the opposite. Here, the speaker is directly stating their opinion
without such inversion.

Considering different perspectives: Some annotators might think that the speaker is being sarcastic in
their tone, even if the words aren't technically ironic. Others might see it as direct criticism without
sarcasm. Since sarcasm can be subtle, there's a chance that a portion of annotators might rate it as
low-level sarcasm (like 2 or 3), but most might see it as non-sarcastic (rating 1). A few might
misinterpret and give higher ratings, thinking the speaker is being sardonic, but that's less likely given
the explicit criticism.

I should distribute the probabilities accordingly. Most of the probability mass would be on 1 (not at all
sarcastic) and 2 (slightly sarcastic). Maybe a small portion on 3 for those who see some underlying
sarcasm. Higher ratings (4-6) would be minimal since the response is pretty straightforward. Let me check
if the sum adds up to 1.0 and adjust the probabilities.
```

Listing 11: Raw CoT for Concrete Answer

```
Okay, let's tackle this problem. So, the task is to guess the distribution of sarcasm ratings given a
context and response pair. The possible ratings are from 1 to 6, with 0 not allowed. The final answer needs
to be a JSON object where each key is the rating (as a string) and the value is the probability that an
annotator gave that rating.

First, I need to understand what sarcasm means here. The definition provided says it's when someone says
the opposite of what they mean, usually to criticize or hurt someone in a humorous way. So, I have to
analyze the response to see if it uses sarcasm and how different people might rate it.

Looking at the context: Steve met a girl while drunk, fell in love, and wants to quit his job to travel
with her. The response is from someone else reacting to Steve's decision. The response says, "Man, you are
fucking serious right now? You're going to quit your job and travel with some girl you just met last night
while drunk? No girl is worth quitting your job and screwing your life over, especially not a girl you just
met last night."

Now, I need to check if this response is sarcastic. Sarcasm often involves irony or mock praise. Here, the
speaker is clearly criticizing Steve's impulsive decision. They're questioning his seriousness and pointing
out the impracticality. There's no obvious use of words that mean the opposite. Instead, it's
straightforward criticism without humor. The tone is more disapproving than sarcastic.

So, most annotators might rate this as low sarcasm (like 1 or 2). Some might think there's a bit of sarcasm
in the rhetorical question ("you are fucking serious"), but that's more shock than sarcasm. Others might
see it as direct advice rather than sarcastic. Maybe a few could interpret the repeated emphasis on "just
met" as mocking, but that's a stretch.

I should consider possible distributions. If the response isn't sarcastic, the majority would rate 1 (not
at all). Some might give 2 for slight sarcasm. Very few might go higher if they think there's hidden
mockery. Let's say 70% rate 1, 20% rate 2, and 10% split between 3 and maybe 4. But wait, the sum has to be
1.0. Let me adjust. Maybe 0.7 for 1, 0.2 for 2, 0.05 for 3, and 0.05 for 4. That adds up to 1.0.
Alternatively, perhaps even less for higher numbers. Another approach: since the response is direct and not
using irony, most will rate 1. A small percentage might see some sarcasm in the phrasing. Maybe 0.8 for 1,
0.1 for 2, and 0.1 distributed among others. But ensuring the total is 1.0. Let's try:

- 1: 0.7
- 2: 0.2
- 3: 0.05
- 4: 0.03
- 5: 0.01
- 6: 0.01

That sums to 1.0. But maybe even lower for higher numbers. Alternatively, maybe only 1 and 2. Let's check
again. The response is critical but not sarcastic. So most would rate 1. Some might rate 2 if they think
there's a hint of sarcasm. Let's say 0.8 for 1, 0.15 for 2, and 0.05 for 3. That's 1.0. Or maybe 0.9 for 1,
0.05 for 2, 0.05 for 3. Hmm. It's tricky. To be safe, I'll go with a high concentration on 1, some on 2,
and minimal elsewhere. So final distribution might look like {'1': 0.7, '2': 0.2, '3': 0.05, '4': 0.03,
'5': 0.01, '6': 0.01}. That adds up to 1.0.
```

Listing 12: LLM prompt for the CSC dataset (soft-label task)

```
[Instructions]

Solve the task below. Format your response into logical steps that can be verified by a third-party, and
format each step in JSON format like this:

{{
    'steps': [
        '<text of step 1>',
        '<text of step 2>',
        ...,
        '<text of step n>'
    ],
    'final_response': '<response>'
}}

Please simplify your final response as much as possible, and avoid any unnecessary text there.

[Task]

Below is a context+response pair where human annotators rated the sarcasm level of the 'response' ranging
from 1 (not at all) to 6 (completely). Please guess the distribution of ratings and output it in the
final_response field in JSON format like this: {{'0': p0, '1': p1, '2': p2, '3': p3, '4': p4, '5': p5, '6':
p6}}, where each p0, ..., p6 is a probability. Note that p0 is always 0.0 because the rating 0 is not
allowed. The sum of probabilities must equal 1.0.

Use this definition for sarcasm: "The use of remarks that clearly mean the opposite of what they say, made
in order to hurt someone's feelings or to criticize something in a humorous way"

Think about the perspectives that different annotators might have and how they could potentially interpret
the context+response pair.

[Context+Response Pair]

{text}
```

Listing 13: LLM prompt for the CSC dataset (perspectivist task)

```
[Instructions]

Solve the task below. Format your response into logical steps that can be verified by a third-party, and
format each step in JSON format like this:

{{
    'steps': [
        '<text of step 1>',
        '<text of step 2>',
        ...,
        '<text of step n>'
    ],
    'final_response': '<response>'
}}

Please simplify your final response as much as possible, and avoid any unnecessary text there.

[Task]

Below is a context+response pair where human annotators rated the sarcasm level of the 'response' ranging
from 1 (not at all) to 6 (completely). Please guess the rating given by each annotator output them all in a
list in the final_response field, in the same order as the annotators.

Use this definition for sarcasm: "The use of remarks that clearly mean the opposite of what they say, made
in order to hurt someone's feelings or to criticize something in a humorous way"

Think about the perspectives that different annotators might have and how they could potentially interpret
the context+response pair.

[Annotator Metadata]

{annotator_metadata}

[Context+Response Pair]

{text}
```

Listing 14: Judge prompt for scoring reasoning steps (soft-label task)

```
[Chain-of-Thought Evaluation]

Overview: Your goal is to grade an LLM's step-by-step solution to a problem. The model will often say
things that look ok at first, but will turn out to be wrong on closer inspection - stay vigilant!
Please mark each step with (great, okay, bad).

Instructions: A "great" step is anything a smart student would try.
Most of the time it's a clear cut step forward towards solving the problem. But it
could also be a sub-optimal choice, as long as it looks like something a reasonably smart
human might say while trying to solve the problem. An "okay" step is anything that's
reasonable for a person to say, but it's not offering any insight, doesn't further the solution
by exploring an option, performing a calculation, or offering an idea for the next step. A
"bad" step is one that confidently says something incorrect, is off-topic/weird, leads
the solution into a clear dead-end, or is not explained clearly enough for a human to follow
along with (even if it is correct).

[Output Format]
Only output the step idx and the rating, like this example below.
[
    {{"idx": 0, "rating": "great"}},
    {{"idx": 1, "rating": "ok"}},
    {{"idx": 2, "rating": "bad"}},
    {{"idx": 3, "rating": "great"}},
    ...
]
Please verify that the number of steps is the same as in the input.

[LLM Problem]

<problem>
{PROBLEM}
</problem>

[LLM Chain-of-Thought Steps]

<steps>
{STEPS}
</steps>

<final_response>
{FINAL_RESPONSE}
</final_response>
```

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

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

Abstract

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

1 Introduction

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

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

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

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

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

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 171–181
November 8, 2025

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

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

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

2 Datasets, tasks, and evaluation metrics

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

2.1 Datasets

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

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

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

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

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

2.2 Tasks and evaluation metrics

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

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

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

3 In-context learning

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

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

3.1 System pipeline

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

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

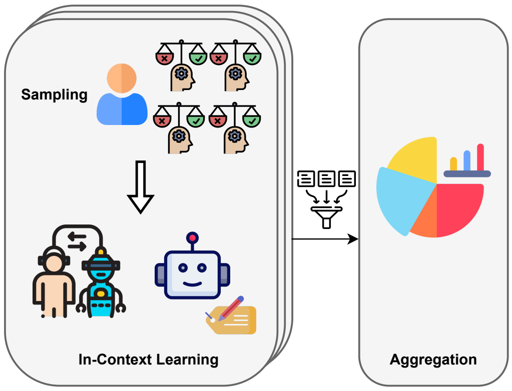


Figure 1: Our two-step pipeline to solve both tasks, based on ICL. In the first step (Task B), we sample examples from an annotator's past annotations and prompt the LLM to model annotator-specific behavior and predict labels for test inputs. In the second step, we aggregate these predictions into soft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

Prompt Template

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

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

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

{EXAMPLES}
[/INST]

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

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

3.2 Example selection strategies

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

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

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

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

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

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

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

3.3 Model performance

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

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

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

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

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

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

3.4 Discussion

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

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

MP–dev–1597 (snippet)

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

[Reply]: How old are you?

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

At the same time, we note that the tested models are generally successful in mimicking specific annotators' labeling strategies. This is best illustrated in the VariErrNLI and Par datasets, since the annotators' motivations are directly available for analysis. For example, in the Par dataset, annotator `Ann3` uses label `0` considerably more often than their peers, consistently labeling most non-contradicting examples with 0 rather than negative values even when they are non-relevant. This reasoning is also reflected in Ann3's explanations. The models, particularly when combined with label-based sampling, tend to imitate this peculiarity while also never predicting 0 for annotators Ann1 and Ann2 (who rarely use it). Likewise, the predictions also reflect less subtle differences, like the annotators' inclination toward positive or negative scale values (for Par) and entailment or neutral labels (for VariErrNLI). For example, Ann3's preference for positive values is also discernible in the predicted labels. In this respect, it can be argued that ICL can be successful when used for perspectivist modeling of individual perspectives.

4 Fine-tuning approaches

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

4.1 Approach 1: Cumulative distances for Likert scales

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.2 Approach 2: Population-level label distributions

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

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

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

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

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

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

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

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

Results Our best score with this approach is a Wasserstein distance of 1.66 for the Par dataset. We achieved this by classifying the dataset into 3 clusters. While the performance is above the baseline by a notable margin, this method still underperformed compared to the other fine-tuning method described in Section 4.1 and the in-context learning method described in Section 3.

4.3 Discussion

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

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

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

5 Conclusion

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

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

Ethical Considerations

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

Acknowledgments

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

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, and 260 others. 2023. Gpt-4 technical report.

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

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

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

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

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

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

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

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

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

Hyewon Jang and Diego Frassinelli. 2024. Generalizable sarcasm detection is just around the corner, of course! In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4238–4249, Mexico City, Mexico. Association for Computational Linguistics.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

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

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

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

Elisa Leonardelli, Silvia Casola, Siyao Peng, Giulia Rizzi, Valerio Basile, Elisabetta Fersini, Diego Frassinelli, Hyewon Jang, Maja Pavlovic, Barbara Plank, and Massimo Poesio. 2025. Lewidi-2025 at nlperspectives: third edition of the learning with disagreements shared task. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.

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

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

Nicole Meister, Carlos Guestrin, and Tatsunori B Hashimoto. 2025. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 24–49.

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

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

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

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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

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

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

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

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

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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

A Example of an ICL prompt

CSC–test–2143–Ann743

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


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

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

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

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

...

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

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

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

LeWiDi-2025 at NLPerspectives: The 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

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 182–195
November 8, 2025 ©2025 Association for Computational Linguistics

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		MP		PAR		VEN	
			SOFT EVALUATION							
			WS	*(rank)*	MD	*(rank)*	WS	*(rank)*	MMD	*(rank)*
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		MP		PAR		VEN	
			PERSPECTIVIST EVALUATION							
			MAD	*(rank)*	ER	*(rank)*	MAD	*(rank)*	MER	*(rank)*
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:

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

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

LPI-RIT at LeWiDi-2025: Improving Distributional Predictions via Metadata and Loss Reweighting with DisCo

Mandira Sawkar*, **Samay U. Shetty***, **Deepak Pandita**,
Tharindu Cyril Weerasooriya, **Christopher M. Homan**,

Rochester Institute of Technology
{ms7201, ss4711, cmhvcs}@rit.edu, {deepak, cyril}@mail.rit.edu

Abstract

The Learning With Disagreements (LeWiDi) 2025 shared task aims to model annotator disagreement through soft label distribution prediction and perspectivist evaluation, which focuses on modeling individual annotators. We adapt DisCo (Distribution from Context), a neural architecture that jointly models item-level and annotator-level label distributions, and present detailed analysis and improvements. In this paper, we extend DisCo by introducing annotator metadata embeddings, enhancing input representations, and multi-objective training losses to capture disagreement patterns better. Through extensive experiments, we demonstrate substantial improvements in both soft and perspectivist evaluation metrics across three datasets. We also conduct in-depth calibration and error analyses that reveal when and why disagreement-aware modeling improves. Our findings show that disagreement can be better captured by conditioning on annotator demographics and by optimizing directly for distributional metrics, yielding consistent improvements across datasets.

1 Introduction

As machine learning systems increasingly mediate social, legal, and civic decision-making, their alignment with human values becomes paramount. However, as any participant in a democratic process knows well, human disagreement is always present. This includes many existing problems, such as hate speech detection, intent classification, or moral judgment. The LeWiDi 2025 shared task (Leonardelli et al., 2025) directly addresses this need by evaluating models on their ability to (1) predict soft label distributions derived from annotator disagreement and (2) approximate individual annotator behavior in a perspectivist setting.

Supervised learning typically resolves annotation disagreement by aggregating labels into a single ground truth, often via plurality vote. However, doing so can obscure valuable minority perspectives, especially on subjective or contentious content (Basile et al., 2021; Prabhakaran et al., 2021; Uma et al., 2021b; Plank, 2022; Cabitza et al., 2023; Homan et al., 2023; Weerasooriya et al., 2023a; Prabhakaran et al., 2023; Pandita et al., 2024). However, preserving and modeling this disagreement can improve system robustness, fairness, and social accountability. Tasks such as MultiPICo (Casola et al., 2024), Paraphrase, VariErrNLI, and CSC (Jang and Frassinelli, 2024) exemplify domains where capturing nuanced human perspectives, rather than just the majority opinion, is essential for ethical and practical deployment. LeWiDi-2025 challenges systems to go beyond single-label classification and instead model the full distribution of possible human responses.

The core challenge lies in modeling disagreement when annotation is both sparse and noisy. Annotators may vary in reliability, background, and interpretation, and most datasets provide only a few annotations per item. Moreover, models must predict not only soft aggregate distributions but also simulate individual annotator responses, requiring them to generalize from partial supervision over complex, entangled signal sources. Compounding this difficulty is the need for robust evaluation across both soft (e.g., Manhattan, Wasserstein) and perspectivist (e.g., Error Rate, Normalized Absolute Distance) metrics, which test a model's fidelity to human-like prediction under both collective and individual frames. The four datasets utilized in the shared task are Conversational Sarcasm Corpus (CSC), MultiPico (MP), Paraphrase (Par), and VariErr NLI (Ven).

We adapt the DisCo (Weerasooriya et al., 2023b) model to the LeWiDi 3rd Edition datasets. DisCo consumes item–annotator pairs as input and jointly predicts three interconnected distributions: the specific label an individual annotator would assign,

*Equal contribution.

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 196–207
November 8, 2025 ©2025 Association for Computational Linguistics

the soft label distribution over all annotators for that item, and the annotator's own distribution over all items.

While DisCo demonstrated the value of jointly modeling item- and annotator-level distributions, it treated annotators as one-hot IDs and optimized losses misaligned with evaluation. We address both limitations by embedding annotator metadata and by designing loss functions directly tied to disagreement-aware metrics, enabling more interpretable and robust models.

For the post-evaluation phase, we made the following contributions:

1. The original DisCo model relied solely on simple annotator ID mappings, limiting its ability to understand annotator characteristics and biases. We modified it to account for annotator metadata features such as age, nationality, gender, education, etc.

2. We extended DisCo's preprocessing capabilities to process a wider range of data formats.

3. We updated the underlying sentence transformer models on which DisCo may depend.

4. We modified the loss functions to align with the evaluation for soft label distribution prediction and perspectivist modeling.

5. We perform extensive failure mode analysis on the model.

With these updates, we observed a substantial improvement in the scores for three datasets: CSC, MP, and Par. Additionally, this placed us as rank 4 instead of 7 for Par and Rank 5 instead of 9 for MP in the post-evaluation phase.

2 Background

The LeWiDi shared task has emerged as a focal point for advancing methods that embrace, rather than suppress, annotator variation, since its inception (Uma et al., 2021a). The third edition, LeWiDi-2025 (Leonardelli et al., 2025), further extends these efforts by evaluating both distributional and perspectivist modeling across diverse datasets.

LeWiDi-2025 focuses on four core benchmark datasets, each designed to probe different facets of human interpretative variation. Please refer to Appendix A.1 for further information on the datasets.

The LeWiDi evaluation draws on two complementary research traditions. First, item–annotator modeling, the goal is to explicitly account for individual annotator behaviors when aggregating labels. Dawid and Skene (1979)'s foundational model represents each annotator's reliability via a latent confusion matrix, enabling joint estimation of true item labels and per-annotator error rates. Subsequent work extended this framework with fully Bayesian treatments (Raykar et al., 2010; Kim and Ghahramani, 2012) and introduced clustering techniques to group annotators by shared labeling patterns (Lakkaraju et al., 2015).

In the second paradigm, label distribution learning (LDL) reframes "ground truth" not as a single class but as a probability distribution over all possible labels. Under this view, models are trained to match the full annotator-derived distribution rather than just the majority vote. Early LDL work demonstrated strong performance in tasks like facial age estimation (Geng, 2016; Gao et al., 2017) and has since been applied to diverse applications, from short text parsing (Shirani et al., 2019) to climate forecasting (Yang et al., 2020), showing that distributional targets can yield richer, more nuanced predictions.

By learning shared embeddings for both items and annotators, DisCo effectively regularizes sparse annotation settings and pools context across related examples. In experiments on six publicly available datasets, DisCo matched or exceeded state-of-the-art LDL approaches, such as multinomial mixture models combined with CNNs, and outperformed annotator-modeling baselines like CrowdLayer across both single-label and distributional evaluation metrics.

Since SemEval-2023, researchers have continued to push toward richer annotator-aware modeling. IREL (Maity et al., 2023) conditions toxicity predictions on anonymized user metadata—integrating each annotator's identity embedding directly into both the model input and the loss function to improve alignment with individual judgments. CICL_DMS (Grötzinger et al., 2023), by contrast, builds on large pre-trained language models and explores ensemble learning, multi-task fine-tuning, and Gaussian process calibration to better match the full distribution of annotator labels. Together, these contributions underscore a growing emphasis on leveraging demographic, behavioral, and contextual signals to capture the nuances of human disagreement.

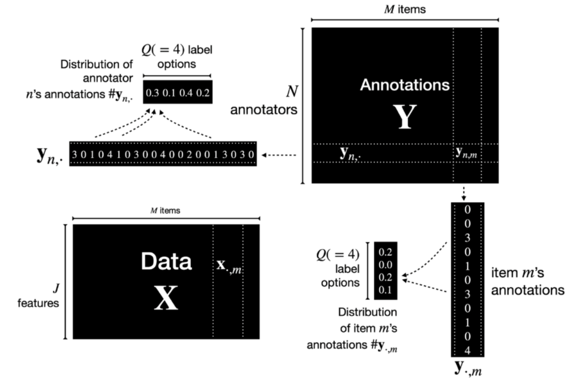


Figure 1: Data representation for DisCo: each item $\mathbf{x}_m$ is paired with per-annotator responses $\mathbf{y}_{\cdot,m}$ and their empirical distribution $\#\mathbf{y}_{\cdot,m}$, and each annotator n has a response vector $\mathbf{y}_{n,\cdot}$ with distribution $\#\mathbf{y}_{n,\cdot}$.

3 System Overview

Our system builds upon the DisCo (Distribution from Context) architecture originally proposed by Weerasooriya et al. (2023b). To adapt it for the LeWiDi-2025 task, we introduced several targeted enhancements, including the use of task-specific sentence encoders, integration of annotator metadata via pretrained embeddings, and modified loss functions to reflect task evaluation metrics. These adaptations enable the model to generalize more effectively from sparse supervision and better capture the complexity of annotator behavior and disagreement.

DisCo is designed to jointly model individual annotator responses, aggregate item-level label distributions, and annotator-level behavior distributions in a unified probabilistic framework.

Each data item $\mathbf{x}_m \in \mathbb{R}^J$ is represented as a column vector of J features, and its associated annotations from N annotators are collected in the matrix $\mathbf{Y} \in \mathbb{Z}^{N\times M}$. We denote the vector of responses for item m as $\mathbf{y}_{\cdot,m}$ and the histogram of these responses as $\#\mathbf{y}_{\cdot,m}$. Similarly, each annotator n's behavior across all items is summarized by $\mathbf{y}_{n,\cdot}$ and its histogram $\#\mathbf{y}_{n,\cdot}$. This setup is illustrated in Figure 1.

In the encoder (Figure 2), item and annotator inputs are mapped into separate subspaces. The item vector $\mathbf{x}_m$ is projected via a learnable matrix $\mathbf{W}_I \in \mathbb{R}^{J_I\times J}$ to yield the embedding $\mathbf{z}_I = \mathbf{W}_I\mathbf{x}_m$, while the one-hot annotator identifier $\mathbf{a}_n$ is projected through $\mathbf{W}_A \in \mathbb{R}^{J_A\times N}$ to produce $\mathbf{z}_A = \mathbf{W}_A\mathbf{a}_n$. These embeddings are concatenated and passed through a two-layer MLP with softsign activations

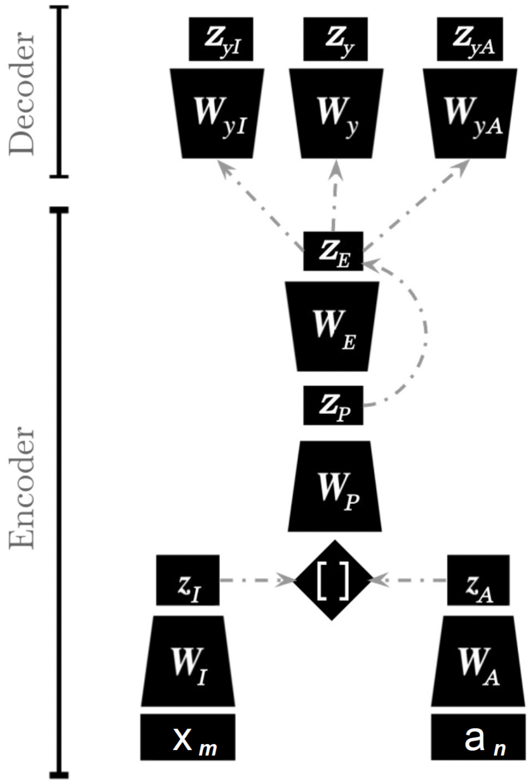


Figure 2: Block diagram of the DisCo encoder and decoder. The encoder maps item and annotator inputs into a joint latent code $\mathbf{z}_E$, and the decoder produces three parallel distributions via softmax heads.

and a residual connection:

$$\mathbf{z}_P = \phi(\mathbf{W}_P \cdot \phi\mathbf{z}_I,\ \mathbf{z}_A), \tag{1}$$

$$\mathbf{z}_E = \phi(\mathbf{W}_E \cdot \mathbf{z}_P\ \ \mathbf{z}_P), \tag{2}$$

where $\mathbf{W}_P$ and $\mathbf{W}_E$ are learned projection matrices.

The decoder takes the joint code $\mathbf{z}_E$ and outputs three softmax-normalized vectors: $\mathbf{z}_y = \mathrm{softmax}\mathbf{W}_y\mathbf{z}_E$ for the per-annotator label distribution $Py_{n,m} \mid \mathbf{x}_m, \mathbf{a}_n$, $\mathbf{z}_{yI} = \mathrm{softmax}\mathbf{W}_{yI}\mathbf{z}_E$ for the item-level distribution, and $\mathbf{z}_{yA} = \mathrm{softmax}\mathbf{W}_{yA}\mathbf{z}_E$ for the annotator-level distribution. Training minimizes a composite loss that combines the negative log-likelihood of observed annotator responses with KL divergence terms that align predicted and empirical label distributions at both the item and annotator levels.

At inference time, for an unseen item $\mathbf{x}_m$ without a specific annotator ID, we embed $\mathbf{x}_m$ to obtain $\mathbf{z}_I$ and tile it across all annotator embeddings in $\mathbf{W}_A$ to form N joint codes. Each code is decoded to yield per-annotator distributions, which are then aggregated by expectation or majority vote to pro-

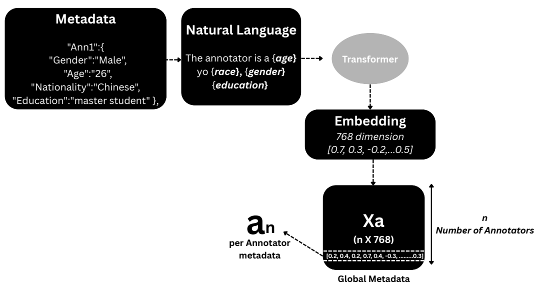


Figure 3: Metadata Embedding Pipeline for DisCo_New: After converting raw metadata into Natural language, it is passed through a transformer to generate embeddings and eventually generate a_n

duce the final item-level prediction. This procedure preserves the learned annotator diversity even when specific annotator metadata is unavailable.

In the post-evaluation phase, we extended the DisCo architecture to better leverage annotator and item information. Annotators were no longer represented by one-hot identifiers but instead by metadata derived from structured JSON inputs. The metadata preprocessing pipeline (Figure 3) concatenated demographic attributes into a textual description, which was then encoded using a transformer-based sentence embedding model $f_{\text{meta}}\cdot$. This produced annotator embeddings $\tilde{\mathbf{a}}_n = f_{\text{meta}}\text{JSON}n \in \mathbb{R}^D$, which were projected through a learnable matrix $\mathbf{W}_A \in \mathbb{R}^{d_A \times D}$ to yield the annotator representation $\mathbf{z}_A = \mathbf{W}_A\tilde{\mathbf{a}}_n$. On the item side, the generic encoder was replaced with a task-specific transformer encoder $f_{\text{item}}\cdot$, producing item vectors $\mathbf{z}_I = \mathbf{W}_I f_{\text{item}}\mathbf{x}_m$. Both item and annotator vectors were mapped into semantically aligned subspaces and concatenated into a joint latent representation $\mathbf{z}_E$, which was decoded following the original DisCo framework.

In parallel, we revised the training objective to incorporate additional distributional and per-annotator losses. Beyond categorical negative log-likelihood and KL divergence, we explored Wasserstein distance for soft-label alignment and mean absolute error for per-annotator alignment, as well as combined and alternating formulations. These revisions aligned optimization more closely with the evaluation metrics. Full implementation details, loss formulations, and dataset-level hyperparameter configurations are described in Section 4.4.

These modifications to the DisCo architecture are not cosmetic but address fundamental gaps: richer annotator modeling and task-aligned optimization.

4 Experimental Setup

4.1 Datasets

Experiments are conducted on three datasets provided by LeWiDi-2025: Conversational Sarcasm Corpus (CSC), MultiPico (MP), and Paraphrase (Par). Each dataset is provided in a unified JSON format, including item-level features, per-annotator labels, and annotator identifiers. The datasets and their evaluation metrics are discussed further in Appendix A.1.

4.2 Tasks

The system is evaluated on the two complementary tasks defined in the LeWiDi-2025 shared task framework. In **Task A (Soft Label Prediction)**, a probability distribution over the label space must be output for each instance. Evaluation is conducted using the Manhattan distance for MP and Ven, and the Wasserstein distance for Par and CSC. In **Task B (Perspectivist Prediction)**, the individual labels assigned by each annotator must be predicted. Evaluation is performed using Error Rate for MP, and Normalized Absolute Distance for Par and CSC. This setup reflects the task's emphasis on modeling annotator disagreement rather than collapsing it into a single ground-truth label.

4.3 Model Configuration: DisCo_OG

The original DisCo model was adapted to the LeWiDi-2025 tasks with minimal modifications. Annotators were represented using simple identifiers, and the model jointly optimized soft-label and perspectivist objectives. Training used a composite loss combining negative log-likelihood of annotator responses with KL divergence against empirical distributions. Hyperparameters such as activation function, optimizer, dropout rate, learning rate, and fusion strategy were tuned based on validation performance.

4.4 Model Configuration: DisCo_New

Building on the architectural extensions described above, we implemented several systematic modifications.

First, the metadata preprocessing pipeline was redesigned to extract annotator attributes (age, gender, nationality, education, etc.) from structured JSON files. These attributes were verbalized into natural language templates and embedded using transformer-based sentence encoders such as

Hyperparameter	Par Value	MP Value	CSC Value
Activation	ReLU	Softsign	elu
Annotator Latent Dim	64	64	256
Item Latent Dim	128	256	256
Fusion Type	Concat	Concat	Concat
Optimizer	Adam	Adam	Adam
Learning Rate	0.001	0.001	0.001
Embedding	paraphrase-mpnet-base-v2	paraphrase-multilingual-mpnet-base-v2	all-mpnet-base-v2
Loss	Wasserstein + MAE (α = 0.6)	KL Divergence	KL Divergence
Weight Init	Gaussian	Uniform	Gaussian

Table 1: Best hyperparameters.

paraphrase-mpnet and all-mpnet. Each annotator's metadata embedding was 768-dimensional and projected into the model space via a learnable transformation matrix, replacing the simple one-hot identifier scheme used in DisCo_OG. This richer representation enabled the model to capture systematic annotator behavior beyond identity-level patterns.

Second, the training objectives were expanded. In addition to KL divergence and categorical cross-entropy, we introduced multi-objective loss functions: (i) Wasserstein distance for aligning predicted and true soft-label distributions (applied to Par and CSC), (ii) mean absolute error (MAE) for per-annotator alignment (also on Par and CSC), (iii) a weighted combined loss that optimized both simultaneously, and (iv) an alternating formulation that switched objectives between epochs. The combined loss proved most effective, defined as:

$$\mathbf{L} = \alpha \cdot \mathbf{L}_{\text{Wasserstein}} \;\; 1 - \alpha \cdot \mathbf{L}_{\text{MAE}},$$

with $\alpha = 0.6$ favoring the soft-label component. This formulation produced the most consistent improvements across datasets.

Finally, extensive hyperparameter sweeps were conducted per dataset. The optimal configurations covering activation functions, latent dimensions, fusion strategies, optimizers, learning rates, embedding models, loss functions, and weight initialization schemes are reported in Table 1.

4.5 Reproducibility

To ensure reproducibility, all experiments were conducted with fixed random seeds and repeated five times per dataset. The optimal hyperparameter settings for each dataset are reported in Section 4.4. Source code is publicly available at `https://github.com/Homan-Lab/lewidi3_public`. The metadata prompt templates are included in Section A.3 in the appendix to facilitate end-to-end replication of our results.

5 Results

We report the official results of our submitted system (under the name "LPI-RIT") on the final leaderboard of the LeWiDi 2025 shared task. Table 2 presents our ranking and evaluation metrics across the three datasets, under both tasks. Our team, "LPI-RIT", placed tenth in both soft and perspectivist tasks among fifteen and eleven teams (including LeWiDi baselines), respectively.

Compared to the two official baselines, our system outperformed the random baseline across all submitted tasks except for Paraphrase, but performed worse than the most frequent label baseline. For soft labels, our results were 1.45 (CSC), 0.54 (MP), and 3.71 (Par) while in the perspectivist task, they were 0.33 (CSC), 0.32 (MP), and 0.44 (Par).

Despite not achieving top rankings, our system provided a consistent output across tasks and served as a solid implementation of the DisCo modeling framework. These results highlight several areas for improvement—particularly in soft-label prediction on CSC and in modeling individual annotator behavior under the perspectivist setup—while affirming the feasibility of generalizing DisCo to the LeWiDi setting without extensive task-specific modifications.

In the post-evaluation phase, we introduced several improvements to the DisCo model, including the use of annotator metadata, expanded preprocessing support, stronger sentence encoders, and loss functions better aligned with soft-label and perspectivist objectives. These changes led to consistent gains across all datasets. Table 3 summarizes these results; further analysis is provided in Section 6.

6 Discussion

Having established that DisCo_NEW consistently outperforms both OG and baselines, we now analyze how and why these improvements occur. In the

Participant	TASK A - Soft Evaluation			TASK B - PE Evaluation		
	CSC	MP	Par	CSC	MP	Par
taysor	0.746	0.422	1.200	0.156	0.288	0.120
dignatev	0.792	0.469	1.12	0.172	0.300	0.130
azadis2	0.803	0.439	1.610	0.213	0.311	0.200
aadisanghani	0.803	0.439	3.050	0.213	0.311	0.490
twinhter	0.835	0.447	0.980	0.228	0.319	0.080
tomasruiz	0.928	0.466	1.800	0.231	0.414	0.230
LeWiDi_mostfrequent	*1.169*	*0.518*	*3.230*	*0.238*	*0.316*	*0.360*
aadisanghani	0.803	0.439	3.051	0.213	0.311	0.491
funzac	1.393	0.551	3.140	0.291	0.326	0.420
LPI-RIT (DisCo_OG)	**1.451**	**0.540**	**3.710**	**0.331**	**0.324**	**0.440**
LeWiDi_random	*1.549*	*0.689*	*3.350*	*0.355*	*0.500*	*0.380*

Table 2: Final leaderboard scores for LeWiDi 2025. Scores reflect error or distance metrics (lower is better).

subsequent comparisons and analyses, the original and updated models are referred to as DisCo_OG and DisCo_New, respectively.

Across all datasets and both tasks, the post-evaluation model (DisCo_NEW) consistently outperforms both our original submission (DisCo_OG) and the strongest LeWiDi baselines. On CSC and Par, DisCo_New reduces error substantially in both soft-label and perspectivist metrics, while on MP the gains are smaller but still clear. These results demonstrate that the proposed extensions—metadata embeddings and task-aligned loss functions—yield tangible improvements over the baseline DisCo architecture and most frequent baselines.

6.1 MultiPICo Analysis

Evaluation: A modest but consistent reduction in Manhattan distance was observed for DisCo_New compared to DisCo_OG (evaluation score reduced from 0.54 to 0.45), indicating that tighter predicted distributions around human soft labels were achieved. A comparison of soft-label confusion matrices (Figure 4) shows a clear improvement in recall for the Ironic class—true positives increased from 92 to 116, while false negatives decreased from 711 to 687. We interpret this shift as evidence of improved sensitivity to sarcastic and ironic instances, which is a core objective of the MP task. Importantly, these gains were achieved with only a small increase in false positives, suggesting that minority perspectives were captured more effectively without over-predicting irony. The error-rate distribution for individual annotator predictions also improved from 0.32 to 0.31. Overall, stronger alignment at the class level and consistency through replication were demonstrated by DisCo_New.

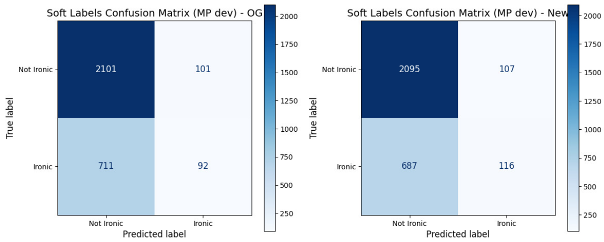


Figure 4: Soft-label confusion matrix for MP dev set (DisCo_New). Improved recall for the Ironic class is shown compared to DisCo_OG.

Confidence Calibration: Improvements in model calibration were also observed. Figure (5), a scatterplot of prediction error versus modal label probability, compares model performance DisCo_OG and DisCo_New using Manhattan distance against modal prediction confidence. In the original submission, the model exhibited numerous high-error predictions even at high confidence, and the error spread remained large across the confidence spectrum. After improvements, the updated model shows a tighter error distribution, particularly in the 0.7–0.95 confidence range, and fewer catastrophic failures at high confidence. This indicates improved calibration and reliability, although low-confidence predictions continue to produce erratic errors, suggesting room for further refinement in uncertain regions of the prediction space.

6.2 Paraphrase Analysis

Evaluation: For the Par dataset, the largest improvement in soft-label matching was recorded,

Dataset	Task	_OG Score	_New Score	LeWiDi Most Frequent Label	LeWiDi Random Label
CSC	Soft	1.45	0.87	1.17	1.54
	PE	0.33	0.22	0.24	0.36
MP	Soft	0.54	0.45	0.52	0.69
	PE	0.32	0.31	0.32	0.5
Par	Soft	3.71	2.21	3.23	3.35
	PE	0.44	0.28	0.36	0.38

Table 3: Original vs. new scores across datasets.

Figure 5: Prediction error vs. modal label probability for the MP dev set. Fewer high-error outliers at high confidence are seen for DisCo_New.

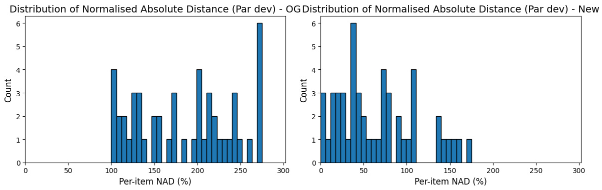


Figure 7: Distribution of Normalized Absolute Distance (NAD) for the Par dev set. DisCo_New exhibits a sharper peak and lower error across the board.

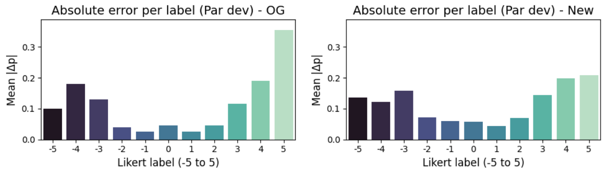


Figure 6: Mean absolute error per Likert label on the Par dev set. DisCo_New (blue) shows a more balanced and lower error profile, especially at the extremes.

with the Wasserstein distance decreasing from 3.71 to 2.21. This indicates substantially better alignment with annotator distributions. The absolute distance was also reduced from 0.44 to 0.28, showing that gains in the soft-label space translated to higher accuracy under the perspectivist evaluation metric. We believe these results demonstrate that DisCo_New can capture annotator-specific variations more effectively.

Error Calibration by Label: To assess model behavior across the Likert scale, mean absolute error per label was examined. As shown in Figure 6, predictions from DisCo_OG were highly skewed, with excessive probability mass assigned to label +5, producing sharp error peaks. A more balanced error profile was seen in DisCo_New, with reduced overcommitment to extreme positive labels while calibration error in the mid-range was maintained or slightly increased. This suggests that output bias was corrected in a way that more faithfully reflects the true distribution of paraphrase strength.

Normalized Error Distribution: Overall soft-label alignment was further assessed using Normalized Absolute Distance (NAD), which measures deviation from the gold distribution relative to total mass. As shown in Figure 7, lower and more concentrated NAD scores were achieved by DisCo_New, with most predictions deviating less than 75%. In contrast, DisCo_OG exhibited inflated NAD values due to label scale mismatches and miscalibration. We view this as evidence that DisCo_New better captures the inherent ambiguity and subjectivity in paraphrase judgments.

6.3 Conversational Sarcasm Corpus (CSC)

Evaluation: For CSC, clear gains in soft-label alignment were recorded. The Wasserstein distance decreased from 1.45 in DisCo_OG to 0.87 in DisCo_New, indicating a closer approximation to gold label distributions. This improvement was especially evident for examples with low annotator consensus. The absolute distance also fell from 0.33 to 0.22, showing significant enhancement in the perspectivist task.

Confidence Sensitivity: The effect of gold label certainty on model performance was examined by plotting prediction error against modal label probability. As shown in Figure 8, lower error for cases with low modal confidence (high annotator disagreement) was achieved by DisCo_New. While DisCo_OG exhibited the highest Wasserstein error in these ambiguous cases, DisCo_New maintained

Figure 8: Prediction error vs. modal label probability on the CSC dev set. Reduced error on low-agreement cases is observed for DisCo_New.

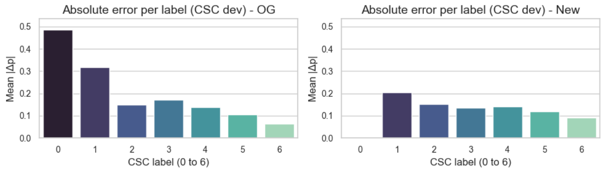


Figure 9: Mean absolute error per Likert label on the CSC dev set. DisCo_New reduces overprediction of non-sarcastic responses (label 0) and achieves smoother calibration overall.

greater stability and resilience, capturing soft-label nuances even when consensus was weak. We see this as further support for the model's improved perspectivist capabilities and robustness in handling disagreement.

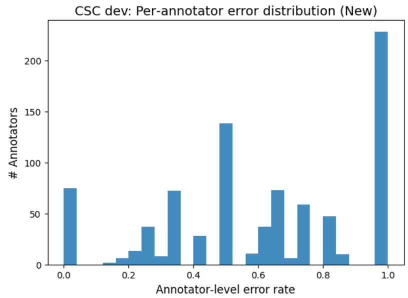
Error Calibration by Label: Mean absolute error per Likert label (Figure 9) showed that DisCo_OG over-predicted label `0`—non-sarcastic interpretations—resulting in large mismatches. This overcommitment was reduced by more than half in DisCo_New. A smoother error profile across all sarcasm intensities was also observed, avoiding the sharp asymmetries seen in DisCo_OG. These findings indicate a more balanced and context-aware handling of literal and sarcastic language, with improved soft-label calibration overall.

6.4 Cross-Dataset Insights

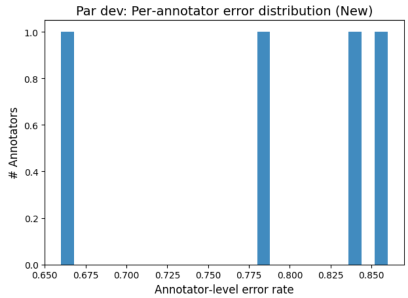
Several cross-cutting patterns emerged across CSC, MP, and Par, providing broader insight into the handling of label ambiguity, annotator disagreement, and error sensitivity.

Annotator-Level Evaluation: Annotator error distributions (Figure 10) showed that for CSC, virtually all annotators were predicted incorrectly by DisCo_OG—error rates clustered at 1.0. In contrast, a more varied distribution was seen for DisCo_New, with many annotators achieving error rates below 0.6. We interpret this as evidence of better alignment with annotator-specific viewpoints. MP remained largely stable, with a slightly tighter distribution under DisCo_New. For Par, high error persisted in both models, driven by strong prior bias in predictions. These findings confirm that while overall system-level scores improved modestly, substantial gains in modeling annotator diversity and disagreement were achieved for CSC.

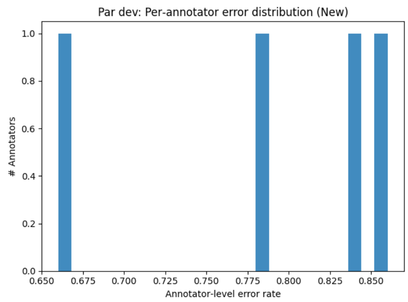
Additional linguistic and entropy-based analyses in Appendix A.2 further support these findings.



(a) CSC (New)



(b) MP (New)



(c) Par (New)

Figure 10: Annotator-level error distributions for the DisCo_New model. Each histogram shows the distribution of absolute error per annotator across the dataset.

These analyses show that beyond leaderboard scores, disagreement-aware modeling yields interpretable and socially relevant gains.

7 Conclusion

We presented enhancements to the DisCo architecture in the context of the LeWiDi-2025 shared task, addressing key limitations in annotator modeling, input representation, and loss formulation. By embedding annotator metadata, refining item encoders, and introducing task-aligned multi-objective losses, our post-evaluation system achieved consistent improvements across CSC, MP, and Par in both soft-label and perspectivist evaluations.

Beyond leaderboard performance, our analyses revealed important behavioral patterns: improved calibration under uncertainty, stronger alignment with annotator-specific perspectives, and greater robustness to label ambiguity. These findings demonstrate that modeling disagreement is not only a technical challenge but also an opportunity to capture the diversity inherent in human annotation.

Looking ahead, we see promising directions in scaling demographic-aware modeling, developing systematic ablation studies, and exploring methods that safeguard fairness and privacy while leveraging annotator metadata. Our work underscores the value of moving beyond aggregated ground truth toward systems that better reflect the complexity of human judgment.

Limitations

Our study has some limitations. First, we did not evaluate on the VariErrNLI dataset, primarily due to time constraints and the additional modeling adjustments the dataset features would require. As a result, our findings are restricted to CSC, MP, and Par, and may not fully generalize to NLI-style disagreement tasks.

Second, while our system integrates multiple extensions to DisCo, including metadata embeddings and revised loss formulations, we did not conduct full ablation studies. Consequently, it is difficult to isolate the contribution of each component, and future work should aim to quantify their relative impact more systematically.

Finally, the use of annotator metadata raises ethical considerations. Demographic information such as age, gender, and nationality can be valuable for modeling disagreement, but also introduces potential risks around privacy and fairness if applied in real-world systems. These aspects warrant further investigation before deployment in sensitive applications.

Future work should address these limitations by extending evaluation to broader datasets, performing systematic ablations, and developing methods that leverage annotator metadata while safeguarding privacy and fairness.

Acknowledgments

We thank the organizers of the LeWiDi-2025 shared task for providing the datasets, evaluation framework, and leaderboard infrastructure. We are also grateful to the anonymous reviewers for their constructive feedback, which helped improve this work. Additionally, we thank Ayo Owolabi for the insightful discussions.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

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.

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

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. 2017. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838.

Xin Geng. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748.

Dennis Grötzinger, Simon Heuschkel, and Matthias Drews. 2023. CICL_DMS at SemEval-2023 task 11:

Learning with disagreements (le-wi-di). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1030–1036, Toronto, Canada. Association for Computational Linguistics.

Christopher M Homan, Greg Serapio-Garcia, Lora Aroyo, Mark Diaz, Alicia Parrish, Vinodkumar Prabhakaran, Alex S Taylor, and Ding Wang. 2023. Intersectionality in conversational ai safety: How bayesian multilevel models help understand diverse perceptions of safety.

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.

Hyun-Chul Kim and Zoubin Ghahramani. 2012. Bayesian classifier combination. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 619–627, La Palma, Canary Islands. PMLR.

Himabindu Lakkaraju, Jure Leskovec, Jon Kleinberg, and Sendhil Mullainathan. 2015. *A Bayesian Framework for Modeling Human Evaluations*, pages 181–189.

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.

Ankita Maity, Pavan Kandru, Bhavyajeet Singh, Kancharla Aditya Hari, and Vasudeva Varma. 2023. IREL at SemEval-2023 task 11: User conditioned modelling for toxicity detection in subjective tasks. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2133–2136, Toronto, Canada. Association for Computational Linguistics.

Deepak Pandita, Tharindu Cyril Weerasooriya, Sujan Dutta, Sarah K. Luger, Tharindu Ranasinghe, Ashiqur R. KhudaBukhsh, Marcos Zampieri, and Christopher M. Homan. 2024. Rater cohesion and quality from a vicarious perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5149–5162, Miami, Florida, USA. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Christopher Homan, Lora Aroyo, Alicia Parrish, Alex Taylor, Mark Díaz, and Ding Wang. 2023. A framework to assess (dis) agreement among diverse rater groups.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research*, 11(43):1297–1322.

Amirreza Shirani, Franck Dernoncourt, Paul Asente, Nedim Lipka, Seokhwan Kim, Jose Echevarria, and Thamar Solorio. 2019. Learning emphasis selection for written text in visual media from crowd-sourced label distributions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1167–1172, Florence, Italy. 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 N 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.

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 Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher Homan, and Ashiqur KhudaBukhsh. 2023a. Vicarious offense and noise audit of offensive speech classifiers: Unifying human and machine disagreement on what is offensive. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11648–11668, Singapore. Association for Computational Linguistics.

Tharindu Cyril Weerasooriya, Alexander Ororbia, Raj Bhensadadia, Ashiqur KhudaBukhsh, and Christopher Homan. 2023b. 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, Toronto, Canada. Association for Computational Linguistics.

Xuebing Yang, Yajing Wu, Wensheng Zhang, and Wei Tang. 2020. Label distribution learning with climate probability for ensemble forecasting. *Intelligent Data Analysis*, 24(1):69–82.

A Appendix

A.1 Datasets

Conversational Sarcasm Corpus (CSC): It comprises roughly 7,000 context–response pairs, each annotated for sarcasm intensity on a six-point scale by both the original response generators ("speakers") and subsequent external observers (Jang and Frassinelli, 2024). In an initial online experiment, speakers wrote a reply to a given situational context and self-rated the sarcasm of their own utterance from 1 ("not at all") to 6 ("completely"). In follow-up studies, fresh cohorts of observers provided independent ratings for the same context–response pairs—six observers per item in Part 1 and four in Part 2—yielding rich soft label distributions that reflect both insider and outsider perspectives.

MultiPico (MP): The dataset is a multilingual irony-detection corpus built from short post–reply exchanges drawn from Twitter and Reddit (Casola et al., 2024). For each entry, crowdsourced annotators judged whether the reply was ironic in light of the preceding post, producing a binary label. Crucially, MP includes sociodemographic metadata (gender, age, nationality, race, student/employment status) for each annotator, and covers eleven languages—among them Arabic, Dutch, English, French, German, Hindi, Italian, Portuguese, and Spanish. On average, each post–reply pair receives five independent annotations, making MP a challenging benchmark for cross-lingual and demographic-aware perspectivist modeling. The paper describing this dataset is available here.

Paraphrase Detection (Par): The benchmark adapts the Quora Question Pairs (QQP) format to a fine-grained judgment task. Four expert annotators each assigned an integer score from –5 ("completely different") to +5 ("exact paraphrase") for 500 question pairs, and provided brief justifications for their ratings. Unlike typical NLI-style datasets, Par uses scalar labels and limits each annotator to one judgment per item, emphasizing inter-annotator variance in graded semantic similarity. This dataset is maintained by the MaiNLP Lab and is not yet formally published.

VariErr NLI ((VariErrNLI)): The corpus was specifically designed to disentangle genuine human label variation from annotation errors in Natural Language Inference (NLI) tasks (Weber-Genzel et al., 2024). In the first round, annotators relabeled 500 premise–hypothesis pairs drawn from the MNLI corpus, providing both labels (Entailment, Neutral, or Contradiction) and free-text explanations for their choices. In the second round, these same annotators validated each label–explanation pair, yielding 7,732 judgments that pinpoint error versus variation. LeWiDi-2025 focuses on the Round 1 soft label distributions, challenging systems to model nuanced NLI judgments at the intersection of semantics and annotator reasoning. The paper describing this dataset is available here.

A.2 Supplementary Analysis

This section provides additional analyses for the three datasets, supplementing the main results discussed in Section 6. The figures below explore linguistic complexity, annotator alignment, and perspective variance in greater detail.

A.2.1 Qualitative Insights from Word Clouds:

Word clouds (from the top 25% hardest and easiest examples (by error) (Figure 11) in each dataset provided further interpretability. In CSC, hard examples in the new system reflected more nuanced social situations (e.g., *"borrowed," "paid," "trust"*), while easy examples featured clear sentiment or tonal markers (e.g., *"congrats," "hang," "job"*). The new system appeared to better distinguish pragmatic cues of sarcasm. In MP, multilingual word clouds remained dense and difficult to interpret visually, but no major shifts were observed in the most frequent hard/easy terms. Par's clouds showed consistent emphasis on mechanical or structured terms (e.g., *"support," "contact"*) in hard cases and evaluative language in easy ones (e.g., *"best," "make," "win"*). These patterns support the conclusion that the new system is sensitive to social and tonal variation, particularly in CSC.

A.2.2 Error vs. Token Length and Entropy:

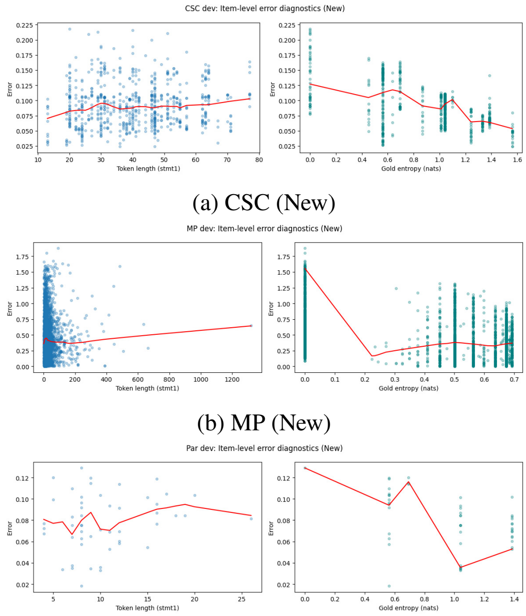
Across datasets, we examined how item-level error varied with input length and gold label en-

(a) CSC (New)

(b) MP (New)

(c) Par (New)

Figure 11: Word clouds.



(a) CSC (New)

(b) MP (New)

(c) Par (New)

Figure 12: Error vs. token length and gold entropy across datasets.

tropy, refer Figure 12. In CSC, the updated model showed improved behavior on high-entropy items—error steadily decreased as label entropy increased, whereas the original model incurred the highest errors for ambiguous cases. This suggests that the revised model better approximates human uncertainty. A similar trend was observed in MP, although gains were more moderate. For Par, error increased slightly with entropy in the new model, possibly reflecting persistent overfitting to majority-label patterns. Overall, the improved system is more robust to uncertainty in CSC and MP, a key desideratum in perspectivist modeling.

A.3 Reproducibility - Metadata Prompts

For full transparency, we provide the exact templates used to verbalize annotator metadata into natural language prompts. These were applied consistently across datasets to ensure reproducible results.

Par: The annotator is `gender`, `age` years old, from `nationality` with an education level of `education`.

MP: The annotator is a `gender`, `age` years old, of nationality `nationality`, born in `country_birth` and residing in `country_of_residence`, with student status `student_status` and employment status `employment_status`, and of ethnicity `ethnicity`.

CSC: The annotator is a `gender` and `age` years old.

These templates allow consistent regeneration of metadata embeddings and support faithful reproduction of our experiments.

McMaster at LeWiDi-2025: Demographic-Aware RoBERTa

Aadi Sanghani†, Sarvin Azadi†, Virendra Jethra, Charles Welch
McMaster University
{sanghana,azadis2,jethrav,cwelch}@mcmaster.ca

Abstract

We present our submission to the Learning With Disagreements (LeWiDi) 2025 shared task. Our team implemented a variety of BERT-based models that encode annotator meta-data in combination with text to predict soft-label distributions and individual annotator labels. We show across four tasks that a combination of demographic factors leads to improved performance, however through ablations across all demographic variables we find that in some cases, a single variable performs best. Our approach placed 4th in the overall competition.

1 Introduction

The shift in natural language processing toward more perspectivist approaches has been positive, in that it allows us to incorporate a variety of viewpoints for subjective tasks and construct models that are more aligned with, and useful for individuals. The number of available disaggregated corpora is small but growing, allowing us to test more techniques in annotator modeling. While the number of available corpora has increased, the amount and type of meta-data about annotators has not significantly changed.

Sociodemographic variables are sometimes collected with annotations for analysis or modeling of the annotators. Without this information, we are often left with only the set of annotations themselves from which to learn patterns. While these demographic variables are not sufficient to represent people or populations and their diverse viewpoints, they give us a starting point to building annotator models that can be expanded in future work as more relevant information becomes available.

With this in mind, we developed a demographic-aware RoBERTa model for the shared task competition. We chose RoBERTa as the transformer of choice as it is well established, and well finetuned which offers strong baseline results, with relatively easy to finetune. With RoBERTa also widely used for NLP tasks, it increased our speed of iteration, and allowed to focus more on demographic adaptions. The shared task we submitted our system to is the 3rd edition of the Learning with Disagreements (LeWiDi) competition (Casola et al., 2025). Our system uses embeddings of demographic features and encodings of text together to predict annotator labels. We evaluate using both soft-label and perspectivist metrics, showing that our model outperforms several baselines, including the Mistral-7b large language model (LLM). Mistral-7b was specifically choosen as it is lightweight enough to run on our hardware, and has demonstrated strong performance across benchmarks. We further perform ablations, exploring the significance of individual demographic variables and discuss directions for future work.

† Denotes equal contribution.

2 Background

The learning with disagreements shared task is motivated by recent efforts in annotator modeling, pluralistic alignment, and data perspectivism. We first describe work along these directions and follow this with an in-depth description of the shared task.

2.1 Related Work

The past decade of work in natural language processing has seen a shift from understanding ground truth as an absolute to be uncovered through annotation, to a subjective value that varies across individuals with different backgrounds and perspectives (Aroyo and Welty, 2015; Frenda et al., 2024). Majority voting can take voices away from underrepresented groups, e.g. older crowdsource workers (Díaz et al., 2019). This kind of aggregation removes perspectives of sociodemographic groups and makes it difficult to discern causes of model underperformance (Prabhakaran et al., 2021).

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 208–218
November 8, 2025

Many recent works have begun releasing disaggregated labels, supporting perspectivist work (Cabitza et al., 2023). These can be used to model annotators using a variety of approaches. Works have used disagreements in Bayesian models to identify unreliable annotations in single ground-truth scenarios (Hovy et al., 2013) and in corpora with differing labels across subpopulations (Ivey et al., 2025). Others have examined the most efficient way to label data, requesting more labels from more uncertain annotators to more efficiently model a spectrum of viewpoints (Golazizian et al., 2024). Perspectivism and personalization have been applied simultaneously in cases where extra annotator information is available (Plepi et al., 2022) with extensions from classification to perspectivist generation (Plepi et al., 2024).

Fornaciari et al. (2021) predicted soft-label distributions for all annotators and found that their model was more robust and higher performing even on the aggregated labels (through majority vote comparison). Mostafazadeh Davani et al. (2022) implemented models with varying degrees in the number of shared parameters across annotators, with some fully independent models, or only shared layers, showing improved performance. They also showed how models that predict multiple labels can be used to measure uncertainty. Mokhberian et al. (2023) proposed a similar approach, which compares multi-task models to a model that embeds individual annotators. These approaches are possible when the set of annotators is not disjoint across the train and test splits.

Deng et al. (2023) studied annotator modeling on eight datasets, finding that demographics correlated with annotation patterns but only explained a fraction of the variance in annotations. While demographic factors are not adequate predictors of differences in opinion, an individuals lived experience can be viewed as a form of expertise which informs their annotation (Fleisig et al., 2024). There is a more meaningful connection between model performance and individual annotator perception than with sociodemographic factors (Orlikowski et al., 2025).

2.2 Shared Task

Our system was built to address the shared task for the 2025 Learning with Disagreements (LeWiDi) competition (Casola et al., 2025). This task invited submissions to build classifiers for tasks not previously addressed in earlier versions of the shared task, including natural language inference, irony detection, and sarcasm detection.

A variety of distributional and information-theoretic metrics have been proposed for modeling human label distributions (Kurniawan et al., 2025). Previous versions of the LeWiDi shared task used cross-entropy and other soft evaluation metrics (Rizzi et al., 2024). This shared task similarly uses soft label predictions for evaluation, where the system outputs the distribution over labels and uses Manhattan distance to measure distance between distributions. It also requires a perspectivist evaluation, where performance is measured as the percentage of correct instances classified at the individual annotator level.

3 System Description

We fined tuned RoBERTa-Large (Liu et al., 2019) to develop a general model and apply this to all datasets through finetuning. The model architecture consisted of many different layers. This model encodes a variable-length text sequence as input and produces embeddings of each token, and a sequence embedding, represented with the [CLS] token.

For all datasets we used all available demographics. We embedded these demographics as follows. To simplify age, we binned the ranges into groups of: $18-24$, $25-35$, $35-44$, $45-55$, $55+$, for the datasets that provided age. Other demographic variables had predefined sets of categorical values from their original work. These are further listed in the demographic breakdowns for each dataset in the Appendix. For each field, a learnable embedding matrix is created, and the text embedding and the demographic embedding are concatenated into a single feature vector. This vector is then normalized using LayerNorm, regularized with dropout and also passed through a linear classifier to produce the logits for classification.

The MP and CSC corpora had many annotators with no instances annotated by all annotators, whereas the Par and VarErr NLI datasets each had only four annotators who annotated all instances. This allows for a slightly different approach. For the first two corpora, we predict each annotators label individually and aggregate them afterward to compute evaluation metrics. In this case, there are no parameters that are specifically designated to any individual annotator. For the latter two corpora, we take a different approach, predicting all annotator labels at the same time. This is similar to the multi-label model described in Mostafazadeh Davani et al. (2022). This approach is tractable due to the very small number of annotators in these corpora. The training time substantially increases with the number of annotators.

We also compared our approach to the Mistral-7b model (Jiang et al., 2023). This is a large language model shown to outperform similar sized models across reasoning, mathematics, and code generation tasks using several recent optimization techniques. This is an instruction-tuned model that is more receptive to prompting.

CSC. For the CSC (Conversational Sarcasm Corpus) dataset (Jang and Frassinelli, 2024), only age and gender were provided as demographic metadata for the annotator model. A notable difference in CSC compared to other datasets is the presence of a context situation paired with a generated "response" from a speaker. The corpus consists of 7k pairs. To preserve their distinct roles in sarcasm, we concatenated the context and response fields into a single input string, delimiting each section with special tokens. This allowed the model to better understand the situation (context) and interpret the reply (response), helping it detect the mismatch or ironic twist between them. Unlike other datasets, the goal for this dataset was to predict the provided sarcasm ratings, which ranged from 0 (not sarcastic at all) to 6 (extremely sarcastic).

MP. Specifically, for the MultiPico (MP) dataset, all of the demographic information wasn't used. The following wasn't used for the final submission: country_birth, nationality, and student status. In preliminary experiments, we found that performance decreased when using all demographic variables. We found that using a combination of country_birth, nationality, country_residence, and ethnicity decreased performance, perhaps due to the inclusion of a redundant but noisy signal. The final model we submitted used Age, Gender, Ethnicity, Country_residence, and Employment as the embeddings. Similarly, the student status meta-data didn't provide any valuable information during preliminary tests and was also omitted. This dataset contains multilingual social media data with post-reply pairs (Casola et al., 2024). The posts are labeled for irony using 0s and 1s, where 1 means that the response is ironic.

Par. The paraphrase detection dataset (Par), consists of question pairs from Quora. We implemented an approach that significantly enhanced the general model architecture. We incorporated SBERT embeddings as a layer alongside RoBERTa-Large to capture semantic similarities between paraphrase pairs more effectively. Specifically, we used the pretrained "all-MiniLM-L6-v2" SBERT model as a frozen feature extractor, concatenating its 384-dimensional embeddings with RoBERTa's 1024-dimensional [CLS] token representation. The model architecture for Par consisted of three main embedding components: RoBERTa-Large embeddings (1024 dimensions), SBERT embeddings (384 dimensions), and demographic embeddings. We used a reduced set of demographic fields (age, gender, nationality, and education) rather than the full available set, as this improved performance by reducing noise from redundant features. Age was binned into discrete ranges, and each demographic field was embedded using learnable 8-dimensional vectors. The final concatenated representation (totaling 1424 dimensions plus demographic embeddings) was processed through LayerNorm and dropout for regularization before being passed to a linear classifier for the 11-class Likert scale prediction (-5 to +5). This approach allowed the model to leverage both syntactic patterns from RoBERTa and semantic similarities from SBERT while accounting for individual annotator perspectives through demographic embeddings.

VarErr NLI. For the Variable Error Natural Language Inference, (NLI) dataset, our approach closely followed the general model architecture. Where it differed is in the output distribution. Each annotator can assign more than one label, making each output a prediction of all three labels for each of the four annotators. This dataset consists of around 1.9k explanations and 7.7 validity judgments of NLI labels (Weber-Genzel et al., 2024). The dataset presented natural language inference tasks with context-statement pairs, where annotators classified relationships as entailment, contradiction, or neutral. We maintained the standard RoBERTa-Large text encoding approach, concatenating the context and statement using the separator token and extracting the [CLS] token representation. Similar to the Par dataset, we predicted the labels of all four annotators at the same time, using separate labels in the output layer. The demographic information available in VariErr NLI included gender, age, nationality, and education level for four annotators. We utilized all avail-

Demographic	CSC	MP	Par	NLI	Values
Age	✓	✓	✓	✓	5
Gender	✓	✓	✓	✓	3
Nationality		✓	✓	✓	33
Education			✓	✓	2
Ethnicity		✓			6
Co. Birth		✓			48
Co. Residence		✓			23
Student Status		✓			2
Emp. Status		✓			7

Table 1: Inclusion of each demographic feature across datasets, showing for which datasets the metadata is present and the number of possible discrete values associated with that feature. Co. stands for *country of* and Emp. for *employment*.

able demographic features without reduction, as the limited number of annotators and demographic diversity made each feature valuable for capturing annotator-specific biases. Age was handled using the same binning strategy as other datasets, and each demographic field was embedded using 8-dimensional learnable vectors. The model produced soft label distributions across the three NLI classes (entailment, contradiction, neutral) rather than hard classifications, allowing it to capture the inherent disagreement and uncertainty in human annotations.

4 Experimental Setup

CSC. The CSC model was trained using soft label cross-entropy loss based on the annotator distributions. We optimized the model using the AdamW optimizer with a learning rate of $2e-5$, weight decay of 0.01, and applied a linear learning rate scheduler with warm-up.

The primary evaluation metric was Manhattan Distance, ranging from 0 to 1, with lower values indicating better performance. We also calculated Absolute Distance (Mean Absolute Error) as a secondary metric to assess the degree of convergence between the annotators' labels and the predicted mean label.

To test model robustness, we experimented with alternative architectures, such as Mistral large language model. However, RoBERTa consistently outperformed these alternatives across both evaluation metrics. Therefore, we carried out an extensive hyperparameter tuning process to further enhance performance, testing with factors including batch size, weight decay, dropout rate, and the number of frozen layers in RoBERTa. After determining an effective value for one of the parameters, we tuned the others while maintaining the same value.

The prompt used for Mistral is as follows:

```
You are a sarcasm detection expert.

Given the following conversation, rate
how sarcastic, the response is on a
scale from 1 (not sarcastic at all) to
6 (extremely sarcastic). Respond only
with a number between 1 and 6.

Context:
{context}

Response:
{response}

How sarcastic is the response?
```

To create this final prompt, we applied prompt engineering techniques. First, we specified the role ("You are a sarcasm detection expert") to encourage analytical reasoning. We then constrained the output ("Respond only with a number") for machine-readability. Next, we separated Context and Response to highlight their distinct roles in sarcasm interpretation. Finally, we ended with a direct question to focus the model. These changes improved clarity, reduced variability, and ensured consistent outputs.

MP. Training for the MP model was based on soft-label cross-entropy loss using annotator distributions with AdamW optimization (learning rate of $2e-5$, and weight decay of 0.01). Similarly, the primary evaluation was Manhattan Distance between the predicted and the true probability distributions, where 0 is the best possible score. The submitted model had a Manhattan Distance of 0.442.

During training, we used plots of the loss in training and validation, learning rate schedule, and performance metrics to inform our tuning of hyperparameters. Parameters were tuned individually, including the dimension of demographic embeddings of size 8, weight decay of 0.01, warm-up ratio of 0.1 and dropout of 0.3 on the concatenated feature vector.

We used prompting the same way as in the MP task. The prompt used for Mistral is as follows:

```
Analyze this social media conversation for
irony:
    Post: "{post}"
```

```
    Reply: "{reply}"
    Is the reply ironic? Consider:
    - Does it say something positive
      about a negative situation?
    - Does it use obvious exaggeration
      or contradiction?
    - Does it mean the opposite of
      what it literally says?

Answer with ONLY a number:
    0 = Not ironic/sarcastic
    1 = Ironic/sarcastic
```

Par. Training for the Par model utilized cross-entropy loss with hard labels rather than soft distributions, as the paraphrase ratings were converted to discrete classes on the Likert scale (-5 to +5, mapped to 11 classes). We used AdamW optimization with a learning rate of $1e-5$, weight decay of 0.01, batch size of 16, and a maximum of 15 training epochs. The learning rate scheduler employed a warmup ratio of 0.15 followed by linear decay. The primary evaluation metric was Manhattan Distance between predicted and true probability distributions, calculated after converting logits to softmax probabilities. Early stopping was implemented with a patience of 5 epochs to prevent overfitting. We also employed gradient clipping (max norm of 0.5) and a dropout rate of 0.3 for regularization. During training, we generated comprehensive analysis plots for each epoch including: prediction vs target scatter plots, prediction distribution comparisons, error distribution histograms, and error vs target relationships. These visualizations helped track model performance and identify potential issues like prediction bias. Key hyperparameters that we tuned included the demographic embedding dimension (8), SBERT embedding dimension (384), dropout rate (0.3), and the specific set of demographic fields used. The reduced demographic field strategy improved performance over using all available features.

The prompt for the Mistral model is as follows:

```
You are an expert at determining semantic
similarity between question pairs. Rate
how similar these questions
are on a scale from -5 to +5, where:
    -5 = Completely different meanings
    -4 = Very different meanings
    -3 = Somewhat different meanings
    -2 = Slightly different meanings
    -1 = Minor differences in meaning
    0 = Neutral/unclear relationship
    +1 = Minor similarities in meaning
    +2 = Slightly similar meanings
    +3 = Somewhat similar meanings
    +4 = Very similar meanings
    +5 = Identical or nearly identical
    meanings

Examples:
Question 1: "How do I learn Python?"
Question 2: "What's the best way to
study Python programming?"
Rating: +4 (Very similar meanings)

Question 1: "What is machine learning?"
Question 2: "How do I bake a cake?"
Rating: -5 (Completely different meanings)

Now rate this pair:
Question 1: "{question1}"
Question 2: "{question2}"
Rating:
```

VariErr NLI. Training for the VariErr NLI model followed a similar approach to other datasets, using soft-label cross-entropy loss based on the three-class probability distributions (entailment, contradiction, neutral). We maintained the AdamW optimizer configuration with appropriate hyperparameters for the NLI task structure. The evaluation was primarily based on Manhattan Distance between predicted and ground truth soft label distributions across the three NLI classes. This metric effectively captured the model's ability to predict not just the most likely class, but the full distribution of annotator disagreement. The model's performance was assessed by how well it could reproduce the uncertainty and variability inherent in human NLI judgments. Given the limited number of annotators, and the importance of capturing individual perspectives in NLI tasks, we utilized all available demographic features without reduction. The hyperparameter tuning focused on balancing the model's capacity to learn individual annotator patterns while maintaining generalization across the three-class output space. The perspectivist approach was particularly important for this dataset, as legitimate disagreement between annotators is common in natural language inference tasks where context interpretation can vary based on background knowledge and reasoning patterns (Pavlick and Kwiatkowski, 2019).

The prompt for the Mistral model is as follows:

```
You are an expert at natural language
inference. Given a context and a
statement, determine the logical
relationship.
Choose from:
- ENTAILMENT: The statement is definitely
true given the context
- CONTRADICTION: The statement is
definitely false given the context
- NEUTRAL: The statement might be true or
false; can't be determined from context

Examples:
Context: "The cat is sleeping on the couch."
Statement: "There is an animal on the
furniture."
Answer: ENTAILMENT

Context: "All birds can fly."
Statement: "Penguins cannot fly."
Answer: CONTRADICTION

Context: "John went to the store."
Statement: "John bought milk."
Answer: NEUTRAL

Now analyze:
Context: "{context}"
Statement: "{statement}"
Answer:
```

5 Results

For all tasks we evaluated using multiple different architectures to understand the impact of various ways of find an optimal model. The summary of results can be found in Table 2, with the comparison against the Majority Baseline. Our main approach, which incorporates the *Demographic Embeddings* for the annotators performs well for the given tasks. This row represents our submission to the shared task competition, which landed us in fourth place when results were computed using the grand average. This scoring approach assigned a rank the same as the random baseline for any particular dataset for which a team performed below that baseline or did not submit any results. Demographic embeddings generally improved model performance. Our model outperformed the simple baseline, the RoBERTa base model, and the Mistral LLM model. The Mistral LLM was prompted to generate responses for each instance in each corpus.

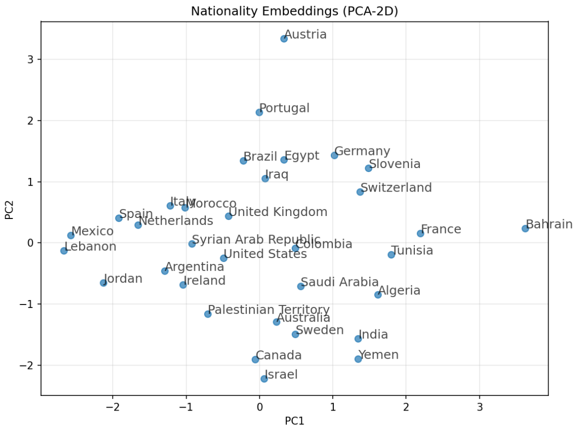


Figure 1: PCA plot showing similarity of embeddings of nationality for the MP task.

We found that even though neither the RoBERTa-base nor Mistral models incorporated annotator-specific features, the LLM performed much worse than RoBERTa.

We performed an ablation by each demographic factor, including only one piece of information at a time. We found that some variables have a much more significant impact on the model than others. The nationality/ethnicity variables appeared to perform best. Gender performed best for the Par and VariErr NLI corpora on the perspectivist evaluation. Surprisingly, we found that some of the single demographic models outperformed our submission to the shared task, showing that even better performance with a demographic-aware RoBERTa model is possible. The VariErr NLI task was the most difficult for our model, as our model underperformed on the soft evaluation and was close to the baseline on the perspectivist evaluation. Future work should explore these relationships in more detail.

6 Discussion

We noted that the LLM performance was substantially worse than the RoBERTa-based models. It is possible that the LLM could perform better with more effort put into prompt-tuning, though this remains to be shown. The added computational overhead and tuning efforts pose barriers to their practical use, over much more readily high-performing, and smaller BERT-based models.

The much smaller RoBERTa models were successful in this task, placing high on the leaderboard and showing greater improvement in our subsequent ablation experiments. Where a person is from, which is partially covered by four

Method	Soft Eval. ↓ CSC	MP	Par	VariErr NLI	Perspectivist Eval. ↓ CSC	MP	Par	VariErr NLI
Majority Baseline	1.169	0.518	3.23	**0.590**	0.238	0.316	0.360	0.340
Demographic Embeddings	**0.803**	0.439	1.610	0.640	**0.213**	0.311	0.200	0.340
- Age Only	0.809	0.443	1.118	0.635	0.216	0.314	0.190	0.335
- Country of Residence Only	-	0.470	-	-	-	0.329	-	-
- Country of Birth Only	-	0.442	-	-	-	0.309	-	-
- Employment Only	-	0.442	-	-	-	0.313	-	-
- Ethnicity Only	-	**0.435**	-	-	-	0.311	-	-
- Gender Only	0.811	0.444	1.145	0.633	0.215	0.310	**0.188**	**0.333**
- Education Only	-	-	1.114	0.650	-	-	0.250	0.400
- Nationality Only	-	**0.435**	**1.063**	0.630	-	**0.307**	0.270	0.380
- Student Only	-	0.449	-	-	-	0.315	-	-
RoBERTa Base	0.821	0.450	1.64	0.645	0.225	0.318	0.380	0.350
LLM (Mistral)	1.020	0.536	2.300	0.680	0.352	0.326	0.450	0.360

Table 2: Breakdown of results for the majority baseline, our submission to the LeWiDi competition, the RoBERTa base model, the large language model Mistral, and an ablation for all demographics. Empty cells mean the demographic is not available for that dataset according to Table 1. Results are shown for both the soft and perspectivist evaluations. Lowest (best) results for each column are shown in bold.

different demographic variables, appeared to have the strongest effect. As participants in the studies which collected the four datasets come from many different countries (see Appendix for details), it makes sense that this would be a variable that correlates strongly with differences in viewpoints or opinion. A PCA plot of the embeddings learned by our best RoBERTa model is shown in Figure 1, showing some regional clusters.

Sarumi et al. (2025) found that of the datasets for this shared task, VariErr NLI had the lowest annotator agreement measured by Krippendorff's alpha, $\alpha = 0.06$, while Par agreement was $\alpha = 0.09$, MP $\alpha = 0.26$ and CSC $\alpha = 0.34$. The low agreement for VariErr NLI, coupled with the low number of annotators may contribute to our lower performance on this task.

As noted in previous work, it is important to emphasize that demographics do not and cannot tell the full story (Fleisig et al., 2024). Given the historical context in which data as been collected and annotated for building NLP models, it is often the case that no meta-data is available for annotators, and when data is available it is often in the form of a handful of demographic variables. This provides us a rough starting point for beginning to explore annotator modeling, but future work must find ways to gather or infer more individual annotation patterns or those that do not directly align with sociodemographic factors.

7 Conclusion

We developed a demographic-aware RoBERTa model for annotator modeling on four tasks, including irony detection, sarcasm detection, paraphrase detection, and NLI. We found that our model could outperform baselines including a large language model; Mistral-7b. In an ablation of demographic factors, we found that nationality and ethnicity led to the biggest performance increases. We note that although demographics provide a starting point to exploring annotator modeling approaches, more individualized approaches will be needed to fully capture differences in annotation patterns.

Limitations

Our experiments with LLMs used only one type of model, which limits the generalizability of the findings, but nonetheless provides a point-of-reference for future exploration. Furthermore, our budget for hyperparameter tuning and further optimization was relatively low given our time constraints and higher performance of the BERT-based models is likely achievable as well.

Importantly, while demographics show that we can improve the model to some extent, they do not provide the full picture. We believe that more individualized approaches will be necessary to improve performance on perspectivist NLP tasks. Applications and developers should not assume that demographics are a sufficient proxy for modeling stakeholders in any scenario. Doing so poses risks to users, the severity of which depend on the specific application, but include both harms of representation and allocation (Blodgett et al., 2020).

References

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

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

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

Silvia Casola, Elisabetta Fersini, Diego Frassinelli, Hyewong Jang, Elisa Leonardelli, Maja Pavlovic, Siyao Peng, Massimo Poesio, and Giulia Rizzi. 2025. Learning with disagreements (LeWiDi) 3rd edition. In *Proceedings of the 4rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)*. Association for Computational Linguistics.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2019. Addressing age-related bias in sentiment analysis. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. 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 2279–2292, Mexico City, Mexico. 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.

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*, pages 1–28.

Preni Golazizian, Alireza S Ziabari, Ali Omrani, and Morteza Dehghani. 2024. Cost-efficient subjective task annotation and modeling through few-shot annotator adaptation. *arXiv preprint arXiv:2402.14101*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Jonathan Ivey, Susan Gauch, and David Jurgens. 2025. Nutmeg: Separating signal from noise in annotator disagreement. *arXiv preprint arXiv:2507.18890*.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kemal Kurniawan, Meladel Mistica, Timothy Baldwin, and Jey Han Lau. 2025. Training and evaluating with human label variation: An empirical study. *arXiv preprint arXiv:2502.01891*.

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

Negar Mokhberian, Myrl G Marmarelis, Frederic R Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2023. Capturing perspectives of crowd-sourced annotators in subjective learning tasks. *arXiv preprint arXiv:2311.09743*.

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.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. *arXiv preprint arXiv:2502.20897*.

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

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.

Joan Plepi, Charles Welch, and Lucie Flek. 2024. Perspective taking through generating responses to conflict situations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6482–6497, Bangkok, Thailand. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, Punta Cana, Dominican Republic.

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.

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.

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.

Appendix

The following tables in this appendix describe the demographic breakdowns for all datasets used in the shared task.

Table 3: Age distribution of MP dataset annotators

Age Group	Count	Percentage
18–24	133	26.3
25–34	219	43.3
35–44	88	17.4
45–54	42	8.3
55+	24	4.7

Table 4: Gender distribution of MP dataset annotators

Gender	Count	Percentage
Male	274	54.2
Female	230	45.5
<UNK>	1	0.2

Table 5: Ethnicity distribution of MP dataset annotators

Ethnicity	Count	Percentage
White	315	62.3
Other	66	13.0
Mixed	64	12.6
Asian	44	8.7
Black	13	2.6
<UNK>	4	0.8

Table 6: Country of residence distribution of MP dataset annotators

Country	Count	Percentage
United States	66	13.0
United Kingdom	54	10.7
Germany	43	8.5
Spain	38	7.5
Canada	37	7.3
Portugal	36	7.1
Netherlands	34	6.7
France	31	6.1
Italy	30	5.9
Mexico	27	5.3
Austria	25	4.9
Switzerland	21	4.2
Australia	20	4.0
Ireland	18	3.6
Hungary	12	2.4
South Africa	5	1.0
Sweden	2	0.4
Israel	2	0.4
Poland	1	0.2
New Zealand	1	0.2
Belgium	1	0.2
Greece	1	0.2
Czech Republic	1	0.2

Table 7: Nationality distribution of MP dataset annotators

Nationality	Count	Percentage
United States	42	8.3
India	39	7.7
Canada	27	5.3
Germany	27	5.3
Netherlands	27	5.3
France	25	4.9
Austria	25	4.9
Portugal	25	4.9
Mexico	25	4.9
Colombia	25	4.9
Italy	24	4.7
Brazil	24	4.7
Spain	24	4.7
Argentina	24	4.7
Switzerland	21	4.2
United Kingdom	18	3.6
Australia	15	3.0
Ireland	15	3.0
Egypt	14	2.8
Syrian Arab Republic	8	1.6
Lebanon	6	1.2
Morocco	5	1.0
Jordan	4	0.8
Palestinian Territory	4	0.8
Saudi Arabia	3	0.6
Algeria	2	0.4
Israel	2	0.4
Slovenia	1	0.2
Bahrain	1	0.2
Tunisia	1	0.2
Sweden	1	0.2
Iraq	1	0.2
Yemen	1	0.2

Table 8: Employment status distribution of MP dataset annotators

Employment Status	Count	Percentage
Full-Time	178	35.2
<UNK>	109	21.5
Part-Time	74	14.6
Unemployed (and job seeking)	74	14.6
Other	36	7.1
Not in paid work (e.g. home-maker, retired)	24	4.7
Due to start a new job within next month	11	2.2

Table 9: Student status distribution of MP dataset annotators

Student Status	Count	Percentage
No	260	51.4
Yes	165	32.6
<UNK>	81	16.0

Table 10: Country of birth distribution of MP dataset annotators

Country of Birth	Count	Percentage
India	34	6.7
United States	31	6.1
Mexico	27	5.3
Colombia	26	5.1
Germany	25	4.9
Austria	25	4.9
Portugal	25	4.9
Netherlands	24	4.7
Brazil	24	4.7
Spain	24	4.7
Argentina	24	4.7
Canada	23	4.5
Italy	23	4.5
France	23	4.5
United Kingdom	17	3.4
Switzerland	17	3.4
Ireland	15	3.0
Egypt	14	2.8
Australia	11	2.2
Syrian Arab Republic	10	2.0
Lebanon	9	1.8
<UNK>	7	1.4
Morocco	7	1.4
Jordan	5	1.0
Saudi Arabia	5	1.0
UAE	4	0.8
Algeria	3	0.6
Togo	2	0.4
Israel	2	0.4
Iraq	2	0.4
Haiti	1	0.2
New Zealand	1	0.2
Hong Kong	1	0.2
South Africa	1	0.2
Dominican Republic	1	0.2
Martinique	1	0.2
Bosnia and Herzegovina	1	0.2
Romania	1	0.2
China	1	0.2
Nicaragua	1	0.2
Chile	1	0.2
Puerto Rico	1	0.2
Kuwait	1	0.2
Bahrain	1	0.2
Somalia	1	0.2
Tunisia	1	0.2
Palestinian Territory	1	0.2
Yemen	1	0.2

Table 11: Age distribution of CSC dataset annotators

Age Group	Count	Percentage
18–24	134	16.5
25–34	273	33.6
35–44	217	26.7
45–54	106	13.0
55+	83	10.2

Table 12: Gender distribution of CSC dataset annotators

Gender	Count	Percentage
Male	418	49.8
Female	397	47.3
<UNK> (Nan, Data_expired, Consent_revoked)	17	2.9

Table 13: Age distribution of Paraphrase dataset annotators

Age Group	Count	Percentage
25–34	3	75.0
35–44	1	25.0

Table 14: Gender distribution of Paraphrase dataset annotators

Gender	Count	Percentage
Male	2	50.0
Female	2	50.0

Table 15: Nationality distribution of Paraphrase dataset annotators

Nationality	Count	Percentage
Chinese	3	75.0
German	1	25.0

Table 16: Education distribution of Paraphrase dataset annotators

Education	Count	Percentage
Master student	4	100.0

Table 17: Age distribution of VariErrNLI dataset annotators

Age Group	Count	Percentage
18–24	1	25.0
25–34	2	50.0
35–44	1	25.0

Table 18: Gender distribution of VariErrNLI dataset annotators

Gender	Count	Percentage
Male	2	50.0
Female	2	50.0

Table 19: Nationality distribution of VariErrNLI dataset annotators

Nationality	Count	Percentage
Chinese	3	75.0
German	1	25.0

Table 20: Education distribution of VariErrNLI dataset annotators

Education	Count	Percentage
Master student	3	75.0
Postdoc	1	25.0

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
November 8, 2025 ©2025 Association for Computational Linguistics

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-α
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*).

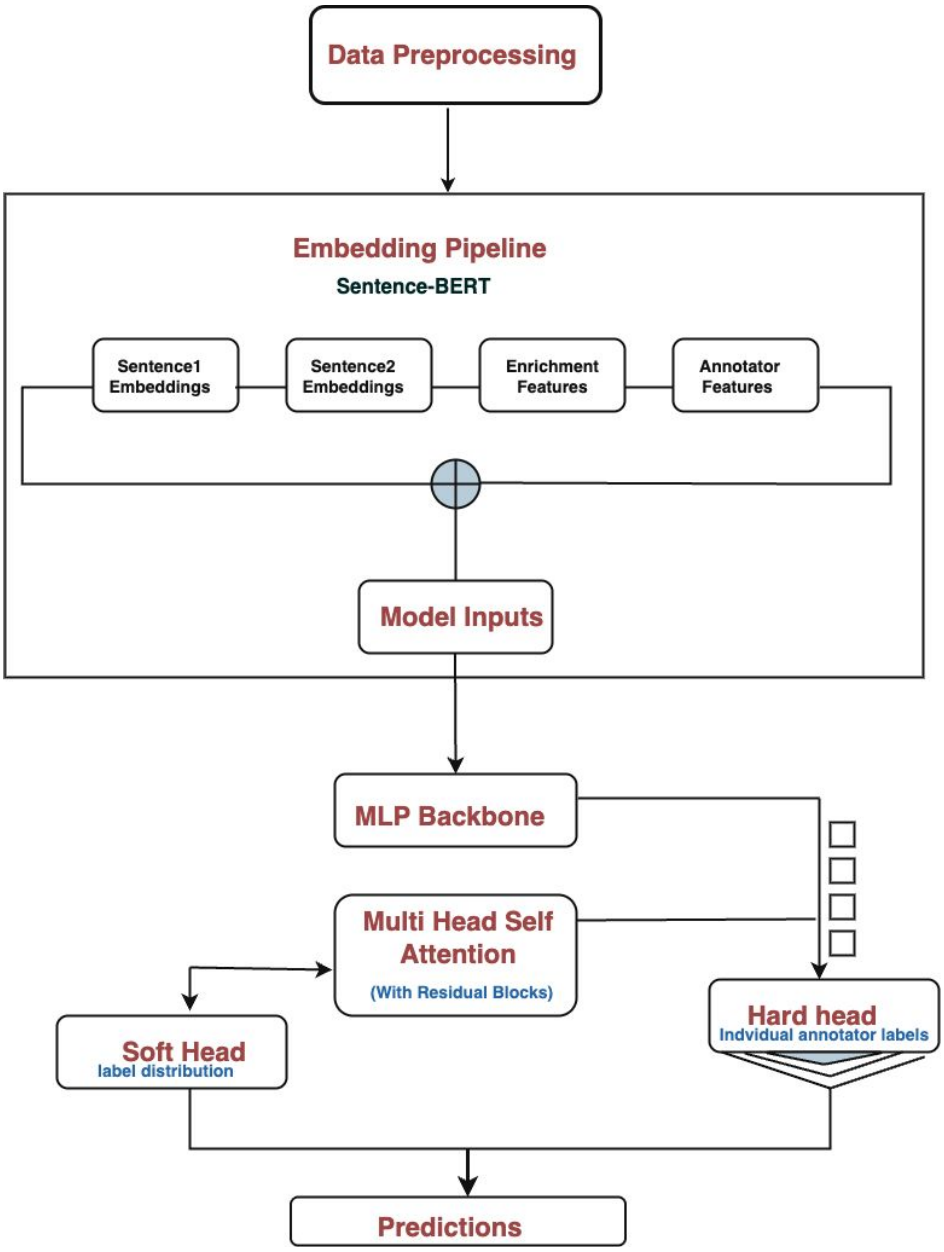


Figure 1: System architecture

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 performance 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 1×10^{-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

Method	Task A (Soft)				Task B (Hard)			
	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**.

Method	Task A (Soft)				Task B (Hard)			
	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	<u>0.746</u>	<u>0.422</u>	<u>0.983</u>	<u>0.233</u>
Perspectivist Task				
Baseline (Random)	0.352	0.499	0.367	0.497
Ours	**0.291**	**0.326**	0.418	**0.345**
Top Submission	<u>0.156</u>	0.289	<u>0.080</u>	<u>0.124</u>

Table 6: Leaderboard Evaluation Results. Best overall results are <u>underlined</u>.

	Error Rate	**MAD**	**F1**	**ROC**
User ID Token			$\color{purple}{\checkmark}$	$\color{red}{\checkmark}$
User Passport Token			$\color{purple}{\checkmark}$	$\color{purple}{\checkmark}$
Composite Token	$\color{purple}{\checkmark}$	$\color{green}{\checkmark}$	$\color{green}{\checkmark}$	
Composite User ID			$\color{red}{\checkmark}\color{purple}{\checkmark}$	
Composite Passport	$\color{blue}{\checkmark}$	$\color{red}{\checkmark}$	$\color{blue}{\checkmark}\color{purple}{\checkmark}$	$\color{blue}{\checkmark}$

Legend	
CSC	$\color{red}{\checkmark}$
MP	$\color{blue}{\checkmark}$
PAR	$\color{green}{\checkmark}$
VAR	$\color{purple}{\checkmark}$

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.

[3]GitHub

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.

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.

Opt-ICL at LeWiDi-2025: Maximizing In-Context Signal from Rater Examples via Meta-Learning

Taylor Sorensen
Department of Computer Science
University of Washington
tsor13@cs.washington.edu

Yejin Choi
Department of Computer Science
Stanford University
yejinc@stanford.edu

Abstract

Many natural language processing (NLP) tasks involve subjectivity, ambiguity, or legitimate disagreement between annotators. In this paper, we outline our system for modeling human variation. Our system leverages language models' (LLMs) in-context learning abilities, along with a two-step meta-learning training procedure for 1) post-training on many datasets requiring in-context learning and 2) specializing the model via in-context meta-learning to the particular data distribution of interest. We also evaluate the performance of our system submission to the Learning With Disagreements (LeWiDi) competition, where it was the overall winner on both tasks. Additionally, we perform an ablation study to measure the importance of each system component. We find that including rater examples in-context is crucial for our system's performance, dataset-specific fine-tuning is helpful on the larger datasets, post-training on other in-context datasets is helpful on one of the competition datasets, and that performance improves with model scale.

1 Introduction

Natural language processing (NLP) evaluations typically assume that there is a single correct answer (a.k.a., "ground truth") and view annotator disagreement as a source of *noise* to be eliminated, generally attributing rating variation to poor instructions, incomplete task specification, or noisy data. However, oftentimes annotator disagreement can be a useful *signal* of subjectivity, ambiguity, or multiple reasonable interpretations (Aroyo and Welty, 2015). Properly integrating this disagreement can be important for robustness, uncertainty calibration, and representing multiple viewpoints. To address this, more and more have argued for focusing on methods for integrating human variation into evaluation and modeling (Basile et al., 2021; Gordon et al., 2022), including annotations from people from diverse backgrounds (Kirk et al., 2024; Aroyo et al., 2023), and aligning AI systems with pluralistic values (Sorensen et al., 2024).

In order to inspire work towards these goals, the Learning With Disagreements (LeWiDi) competition (Leonardelli et al., 2025) consists of four datasets across two tasks for modeling disagreement: one task for predicting how a particular annotator's ratings ("perspectivist" task) and one for predicting the distribution of labels that a pool of annotators gave ("soft label" task). In this system paper, we outline our system submission.

Our system (Opt-ICL, for Optimizing In-Context Learning) takes a fully perspectivist approach, trying to predict how an individual annotator rated each instance and then aggregating individual predictions into a distribution for the soft task. It primarily leverages LLMs' in-context learning ability (Brown et al., 2020; Xie et al., 2022), including an annotator's train ratings directly in-context at inference time. On top of a pre-trained autoregressive language model, we additionally perform two steps of training: post-training in order to enhance the models' in-context learning abilities and teach a unified prompt format (or, Spectrum Tuning, see Sorensen et al. 2025b), and dataset-specific fine-tuning. Both training steps can be seen as forms of meta-learning (Vanschoren, 2018; Min et al., 2022), where the model is tasked with learning how best to fit to the in-context rater examples.

Our main contributions include: our proposed system for modeling disagreement (§3), which was **the overall winner on both competition tasks**, and an ablation study outlining the effect of each system component (§4).

In particular, we find that:

- Including rater examples in-context is crucial for performance;
- Dataset-specific fine-tuning is helpful on larger datasets;

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 228–241

Perspectivist Task	MP (error rate ↓)	CSC (abs. dist. ↓)	Par (abs. dist. ↓)	VEN (error rate ↓)	Average Rank
Ours	**.289 (1)**	**.156 (1)**	.119 (2)	.270 (2)	-
Best other team	.300 (2)	.172 (2)	**.080 (1)**	**.124 (1)**	-
Most frequent label baseline	.316	.239	.362	.345	-
Random label baseline	.499	.352	.367	.497	-
Ours (rank)	**1**	**1**	2 (2-way tie)	2 (4-way tie)	**1.5 (1)**
Best other team (name)	DeMeVa (2)	DeMeVa (2)	**twinther (1)**	**twinther (1)**	DeMeVa 2 (2)
Soft Task	**MP (Manh. dist. ↓)**	**CSC (Wass. dist. ↓)**	**Par (Wass. dist. ↓)**	**VEN (Manh. dist. ↓)**	**Average Rank**
Ours	**.422 (1)**	**.746 (1)**	**1.20 (1)**	.449 (3)	-
Best other team	**.428 (1)**	**.792 (1)**	**.983 (1)**	**.233 (1)**	-
Most frequent label baseline	.518	1.17	3.23	.595	-
Random label baseline	.687	1.54	3.35	.676	-
Ours (rank)	**1 (2-way tie)**	**1 (2-way tie)**	**1 (3-way tie)**	3 (3-way tie)	**1.5 (1)**
Best other team (name)	**PromotionGo (1)**	**DeMeVa (1)**	**twinther (1)**	**twinther (1)**	DeMeVa 2.75 (2)

Table 1: Competition final results. Our system had an average rank of 1.5 on both the perspectivist and soft tasks, and was the **overall winner for both tasks**. First place result bolded, second place underlined for each dataset. The competition organizers determined ties by a two-sided Wilcoxon signed-rank test with the rank leader on item-level scores, failing to reject a difference above $\alpha = .05$ (see Appendix B).

- Post-training on other in-context perspectivist datasets (Spectrum Tuning) significantly helped on one dataset;
- Performance scales with model size, but size alone does not compensate for dataset-specific training.

2 Background and Task Summary

The Learning With Disagreements competition (LeWiDi, Leonardelli et al. 2025) aims to evaluate machine learning systems' ability to engage with and model human variation. The competition spans four datasets which contain subjective judgments where raters may disagree:

- the MultiPIco dataset (MP) (Casola et al., 2024), in which workers label whether or not a short exchange from Twitter/Reddit is ironic (binary);
- the Conversational Sarcasm Corpus (CSC) (Jang and Frassinelli, 2024), involving a 1-6 Likert scale for rating the level of sarcasm of a response given a context (6-way classification);
- A paraphrase detection dataset (Par) (as of yet unpublished, shared by conference organizers) from Quora Question Pairs where annotators rate how strongly the questions are paraphrases of each other on a Likert scale from -5 to 5, along with an explanation (11-way classification); and
- the VariErrNLI (VEN) dataset (Weber-Genzel et al., 2024), on which annotators reannotate premise/hypothesis pairs for entailment. Annotators could assign one or more labels from entailment, neutral, and contradiction and provide an explanation (3 binary classifications, with at least one positive label).

In addition, some basic demographic information is provided about annotators for all datasets.

For dataset statistics, see Table 2. Notably, MP and CSC are much larger datasets than Par and VEN: the MP/CSC train data contains 50k/25k ratings from 506/872 annotators, while Par/VEN contain 1.6k/1.5k ratings from 4/4 annotators respectively.

Using these datasets, the competition constitutes two tasks: a *“soft labeling" task*, where the goal is to predict a probability distribution over possible labels that best match the human annotator label distribution and a *“perspectivist" task*, where the goal is to take on the perspective on an individual annotator and predict that particular annotator's label given prior demonstrations from that rater and (optionally) some demographic information.

For scoring submissions, the two binary datasets (MP/VEN) evaluate the soft task with Manhattan

TRAIN SPLIT	MP	CSC	Par	VEN
# Ratings	60,471	25,574	1,600	1,505
# Instances	12,017	5,628	400	388
# Annotators	506	872	4	4
# Mean Rat./Ann.	119.5	29.4	400	360.8
# Min Rat./Ann.	10	21	400	348
# Max Rat./Ann.	147	38	400	373
DEV SPLIT	MP	CSC	Par	VEN
# Ratings	15,178	3,186	200	187
# Instances	3,005	704	50	50
# Annotators	506	850	4	4
TEST SPLIT	MP	CSC	Par	VEN
# Ratings	18,693	3,224	200	199
# Instances	3,756	704	50	50
# Annotators	506	860	4	4

Table 2: Dataset statistics across train, dev, and test splits for the four LeWiDi datasets. MP and CSC are much larger across the total number of ratings and the number of annotators.

distance and the perspectivist task with error rate. The two Likert scale datasets (CSC/Par) are evaluated using Wasserstein distance for the soft task and absolute distance for the perspectivist task.

For additional information on the competition setup, please refer to Leonardelli et al. (2025).

3 System Overview

Our system consists of three components:

1. Spectrum Tuning (or SpecT, Sorensen et al. 2025b): Post-training an autoregressive large language model (LLM) on a collection of datasets with human variation, stochasticity, or epistemic uncertainty;

2. Dataset-specific fine-tuning on in-context demonstrations from each rater; and

3. Inference with in-context annotator information and training demonstrations.

Specifically, our system uses the `google/gemma-3-12b-pt` (Gemma Team et al., 2025) language model.

3.1 Prompt Structure

Our method depends on LLMs' ability to do in-context learning (Brown et al., 2020; Xie et al., 2022). We adopt the prompting structure from Sorensen et al. (2025b), which has three components: a description (including a task description/any annotator demographics), inputs (the instance to rate), and outputs (the given rating). For example, here is a prompt from Par:

```
Given a pair of questions from Quora Question
  ↪ Pairs (QQP), assign a Likert scale score
  ↪ from -5 to 5 indicating how strongly the
  ↪ questions are paraphrases of one another,
  ↪ and provide a short explanation for your
  ↪ score.
Annotator demographics: annotator_id: Ann1;
  ↪ Gender: Male; Age: 26; Nationality: Chinese
  ↪ ; Education: master student
{"question1": "What are some things new
  ↪ employees should know going into their
  ↪ first day at Exact Sciences?", "question2":
  ↪  "What are some things new employees should
  ↪  know going into their first day at Garmin
  ↪ ?", "lang": "en"}
<start_of_turn>{"paraphrase_rating": -1,
  "explanation": "The companies are
  different."}<end_of_turn>
{"question1": "Who are the everyday heroes and
  ↪ heroines of life?", "question2": "What was
  ↪ everyday life like under Nazi rule?", "lang
  ↪ ": "en"}
<start_of_turn>{"paraphrase_rating": -5,
  "explanation": "Q1 asks about everyday heroes
  and heroines. Q2 is aobut everyday life under
  nazi rule"}<end_of_turn>
{"question1": "What does 'sandiaga' mean?", "
  ↪ question2": "What does \u064a\u0639\u0646\
  ↪ u064a mean?", "lang": "en"}
<start_of_turn>{"paraphrase_rating":...
```

The output of interest (in this case, a paraphrase rating and explanation) is wrapped in special tokens `<start_of_turn>/<end_of_turn>`. While the LeWiDi competition evaluates only a systems' ability to predict the Likert/binary score, we include all rating data in the prompt (including the explanations) with the reasoning that 1) the rater's stated reasoning may contain predictive information for new examples and 2) training on the rating and the explanation concurrently may be helpful.

When predicting how a given rater may respond to a particular instance (e.g., the "perspectivist" approach), we include their demographics at the beginning of the prompt, put as many example train ratings as will fit into context, and append the instance to evaluate at the end of the context.

Throughout the paper, we use a maximum context length of 3,000 tokens. With this limit, we are able to fit about 16 in-context examples for MP, 29 for CSC, 35 for Par, and 29 for VEN (See Table 3).

3.2 Spectrum Tuning: Post-Training for In-Context Steerability

Given this prompt structure, we post-train a language model on a large collection of >40 datasets involving human variation, epistemic uncertainty, or stochasticity, as described in Sorensen et al. (2025b). The post-training technique consists of

Dataset	MP	CSC	Par	VEN
In-Context Examples per Rater Prompt				
Mean	15.8	28.6	35.0	29.1
Min	1	21	32	27
Max	32	37	41	31
Prompt Length (tokens)				
Mean	2,542.1	2,492.1	2,717.1	2,707.3
Min	182	1,647	2,688	2,649
Max	2,798	2,776	2,757	2,769

Table 3: Prompt length and number of in-context examples used during inference across datasets.

unifying the datasets into the common description/input/output format, removing any local dependencies by shuffling the in-context examples, and fine-tuning with cross-entropy loss *only* on the output/<end_of_turn> tokens (a.k.a., the highlighted tokens in the example Par prompt). This post-training is meant to enhance the models' in-context learning abilities, teach the model to focus on predicting only the output tokens wrapped in the scaffolding, and improve calibration. For additional details, please refer to Sorensen et al. (2025b) and App. D.

3.3 Dataset-Specific Training

Once we have the post-trained ICL model, we specialize the model to the particular dataset on which we plan to do inference. We do so by templatizing the entire train dataset in our prompt format, where all ratings in a given context are from the same rater, and performing additional supervised fine-tuning with cross-entropy loss on *just* the output tokens (same format and loss as SpecT, just with data only from target inference dataset). On MP/CSC, we include one training sequence per annotator, and on Par/VEN, which only have four annotators each, we batch into groups of 20 (Par)/30 (VEN) ratings per prompt and train on multiple sequences per annotator.

This could be seen in a way as meta-learning for the specific dataset (Vanschoren, 2018; Min et al., 2022), with each rater being a different "task" to which the model has to adapt in-context.

3.4 In-Context Inference

With the dataset-specific specialized model, we then do inference for each test instance / rater pair 1) by adding randomly-selected train examples into context until we hit a maximum token budget and 2) putting the target test instance at the end. We then directly calculate the model's probability of each label given the rater prompt, which is tractable due to there only being a small set of possible outputs.

Since MP and CSC's possible outputs all differ by only the initial token, only one forward pass per test rating was required. However, VEN and Par's outputs span multiple tokens, and thus required multiple forward passes in order to estimate the entire output probability distribution. Finally, we normalize the probability distribution to sum to one, removing probability mass on any token sequences that do not result in a valid label.

At the end, we have a probability estimate for all possible outputs for each test rater/instance combination.

3.5 From Probabilities To Submission

Up until this point we have taken a wholly perspectivist approach to predicting (a distribution over) how each rater will respond to each test instance. However, the perspectivist task requires a single answer candidate, and the soft task requires an distributional estimate of the entire population of raters will rate an instance.

For the perspectivist task, we submit the single response that minimizes the corresponding evaluation loss. For the two binary datasets, we submit the argmax response.[1] For the two Likert datasets, we make the assumption that our label distribution estimate is well-calibrated, and submit the 50th percentile (median) Likert response of the distribution as this minimizes the expected absolute distance given draws from our distribution estimate.

For the soft task, the optimal distribution to submit under the evaluation criteria (Manhattan/Wasserstein) depends on how well-calibrated our probability estimates are. Here, rather than assuming a well-calibrated distribution, we experiment with a few approaches and submit the one that has the best dev set performance, which are as follows: MP/Par: the averaged distributions for all test raters who annotated the instance; CSC/VEN: an equal average of 1) the averaged distributions and 2) the averaged perspectivist single-answer submissions.

4 Results

We now outline how our system performed compared to others in the competition. Then, we ablate

[1] With the added constraint for VEN that each rater submits at least one positive annotation from entailment, neutral, and contradiction.

PERSPECTIVIST TASK	MP (error rate ↓)	CSC (abs. dist. ↓)	Par (abs. dist. ↓)	VEN (error rate ↓)
Opt-ICL (SpecT + SFT + Demographics + ICL)	**.289 (1)**	**.156 (1)**	**.119 (1)**	**.270 (1)**
Prompt ablations				
no demographics	<u>.295 (2)</u>	**.156 (1)**	**.122 (1)**	**.268 (1)**
no many-shot ICL (one example)	.305 (3)	.185 (3)	.216 (3)	<u>.321 (2)</u>
Training Ablations				
no SFT	.316 (4)	.191 (3)	**.123 (1)**	**.257 (1)**
no SpecT	.303 (3)	**.157 (1)**	**.120 (1)**	**.247 (1)**
no SFT, no SpecT (12B-pt)	.336 (5)	.192 (3)	**.129 (1)**	**.243 (1)**
Model Size ablations (no train)				
1B-pt (no SFT, no SpecT)	.341 (6)	.219 (5)	.308 (4)	.429 (3)
4B-pt (no SFT, no SpecT)	.351 (7)	.201 (4)	<u>.174 (2)</u>	<u>.314 (2)</u>
12B-pt (no SFT, no SpecT)	.336 (5)	.192 (3)	**.129 (1)**	**.243 (1)**
27B-pt (no SFT, no SpecT)	.312 (4)	<u>.176 (2)</u>	**.120 (1)**	**.246 (1)**

SOFT TASK	MP (Manh. dist. ↓)	CSC (Wass. dist. ↓)	Par (Wass. dist. ↓)	VEN (Manh. dist. ↓)
Ours (SpecT + SFT + Demographics + ICL)	**.422 (1)**	**.746 (1)**	**1.20 (1)**	**.449 (1)**
Prompt ablations				
no demographics	<u>.430 (2)</u>	**.751 (1)**	**1.17 (1)**	**.458 (1)**
no many-shot ICL (one-example)	.448 (3)	<u>.851 (2)</u>	2.27 (3)	**.484 (1)**
Training Ablations				
no SFT	.486 (5)	.963 (3)	**1.15 (1)**	**.446 (1)**
no SpecT	.450 (3)	**.749 (1)**	**1.21 (1)**	**.418 (1)**
no SFT, no SpecT (12B-pt)	.507 (6)	.959 (3)	**1.21 (1)**	**.427 (1)**
Model Size ablations (no train)				
1B-pt (no SFT, no SpecT)	.511 (7)	1.13 (5)	3.24 (4)	.703 (3)
4B-pt (no SFT, no SpecT)	.526 (8)	1.03 (4)	<u>1.75 (2)</u>	<u>.519 (2)</u>
12B-pt (no SFT, no SpecT)	.507 (6)	.959 (3)	**1.21 (1)**	**.427 (1)**
27B-pt (no SFT, no SpecT)	.462 (4)	<u>.875 (2)</u>	**1.11 (1)**	**.413 (1)**

Table 4: Ablation study results for a hypothetical competition between all entries shown, with the rank in parentheses. First place is bolded, second place is underlined. Ties are determined sequentially by a two-sided Wilcoxon signed-rank test on item-level scores, failing to reject a difference with the rank leader above $\alpha = .05$ significance, as in the actual competition (see Appendix B for details). To see the results presented visually, also see Fig. 1.

the components of our system to determine the effect of each on task performance.

4.1 LeWiDi Competition Results

Our system was the overall winner on both tasks. The final results can be seen in Table 1. For MP and CSC, our system had the lowest (best) scores for both the perspectivist and the soft tasks. We tied for second across the perspectivist evaluations for Par and VEN, tied for first on Par (soft), and got third on VEN (soft). Our average rank for the perspectivist and soft tasks was 1.5/1.5, which was the lowest overall rank across all teams, meaning our system was the overall winner for both the perspectivist and soft tasks.

4.2 System Ablations

What was the effect of each component of our system? To answer this, we ablate 1) the continued model training via gradient descent, 2) the prompt components, and 3) the size of the underlying LLM. We ablate the components and report the raw scores along with the rankings of a hypothetical competition between the ablated systems. Results can be found in Table 4 and Figures 1.

As a note, the MP and CSC datasets were much larger (3.8k/704 test instances) than the Par/VEN datasets (50/50 test instances). This allows us to make more confident comparisons for the MP/CSC results and affects the size of the available training data for model training.

In-context rater examples were crucial. In the inference prompts, we included many demonstration ratings per annotator (average: 16/29/35/29 across MP/CSC/Par/VEN, c.f. Table 3). To ablate the effect of the examples, we experimented with only including a single rater demonstration. Across all dataset/task combinations, we saw a substantial performance degradation when restricting to only one example (statistically significant across 7/8 comparisons). This suggests that our system relies heavily upon the inclusion of these in-context demonstrations and the models' in-context learning ability.

Interestingly enough, this is true even for the Par dataset, where we include the annotator ID in the demographic description.[2] Even though the model theoretically should be able to connect the annotator instances from its training data to that annotator through the annotator ID, performance substantially dropped when omitting the in-context examples (perspectivist: .119→.216, soft: 1.20→2.27). In other words, in our case, the model is much better able to leverage rater examples when provided concretely in-context at inference time, as opposed to relying on its "soup" of model weights updated via gradient descent.

Demographics did not significantly help. Omitting the rater demographics, on the other hand, did not cause a significant drop in performance on CSC/Par/VEN, and caused only a slight drop in performance on MP. This suggests that the system was not able to significantly leverage sociodemographics in order to improve predictivity, in line with prior work (Orlikowski et al., 2025; Sorensen et al., 2025a).

Dataset-specific fine-tuning was important for the large datasets. For MP and CSC, omitting dataset-specific fine-tuning caused a significant drop in performance on both the perspectivist and soft tasks. We hypothesize that this dataset-specific fine-tuning helped mainly due to 1) (meta-)learning patterns of how to utilize in-context examples; 2) building better priors over how the average rater approaches the task; and 3) specializing to the instance data distribution.

Dataset-specific fine-tuning did not, however, make a significant difference on Par/VEN. We hypothesize that the difference in result is largely due to dataset size, with only 400/388 annotations for Par/VEN in the training data. We also used the same hyperparameters for all datasets, and did not particularly adapt them to squeeze more out of the smaller dataset. Further optimization may be able to extract more signal, but machine learning systems generally struggle more in this low-data regime.

Spectrum Tuning significantly helped on MP. Applying SpecT did significantly help on both MP tasks (perspectivist: .303→.289, soft: .450→.422), but did not significantly help or hurt on the other datasets. We are not sure why it significantly helped in some cases and did not others, but it is not due to any additional irony detection training data, as that was not included in the SpecT training mix (see Appendix D).

Performance improves with model size, but size alone does not compensate for dataset-specific

[2] Due to an oversight that was not realized until after the conclusion of the competition, annotator ID was not included in the prompt for the other datasets.

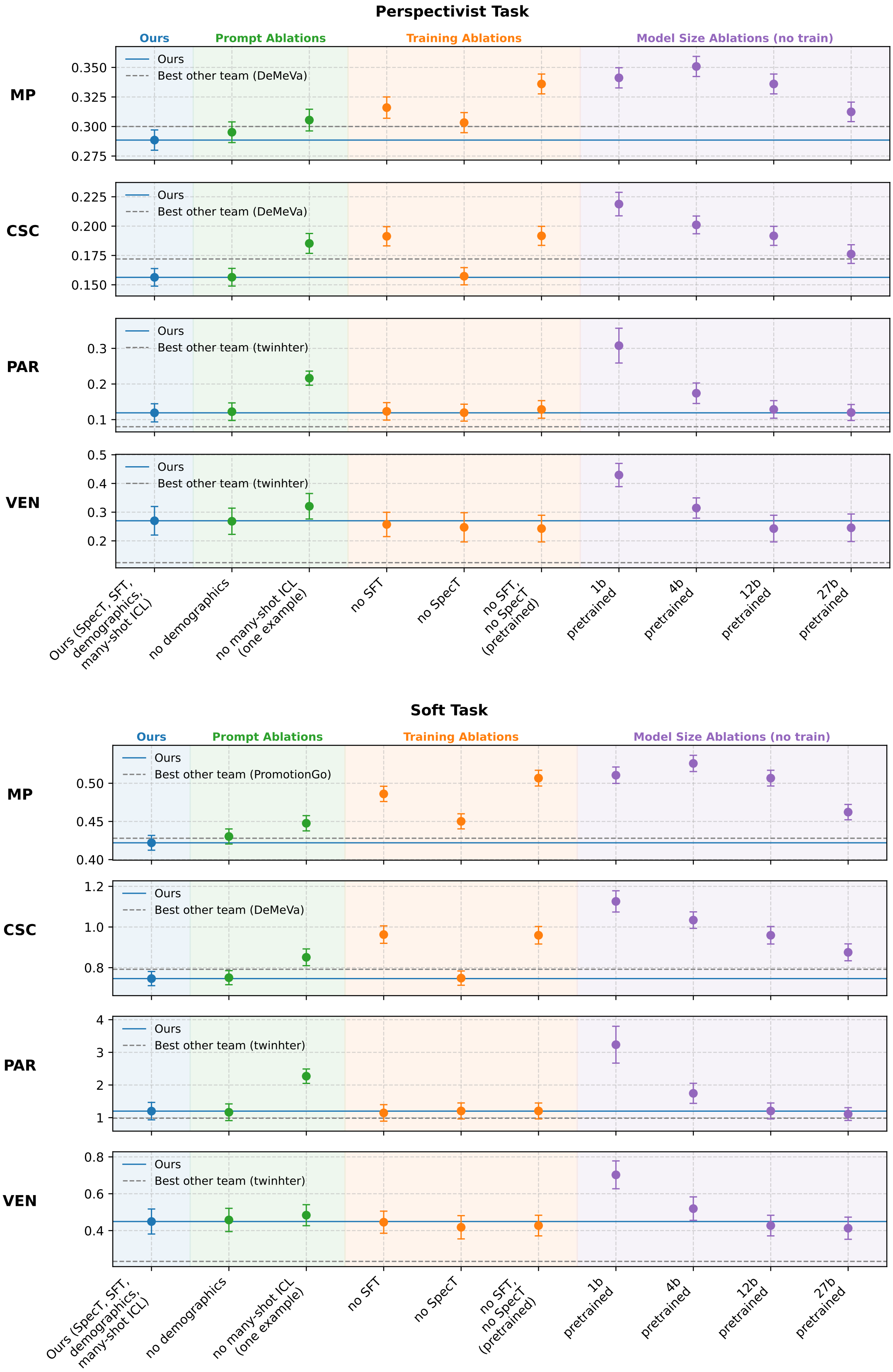


Figure 1: Ablation study results. Perspectivist Task: For MP/VEN, error rate is reported, and for CSC/Par, absolute distance is reported (lower is better for both). Soft Task: For MP/VEN, Manhattan distance is reported, and for CSC/Par, Wasserstein distance is reported (lower is better for both). Error bars indicate 95% confidence intervals, computed as $\pm$ 1.96 times the standard error of the mean of instance-level scores. Our system performance is shown as a solid line, and the best competing team performance is shown as a dashed line.

training. Due to computational constraints, we did not replicate our entire system (with SpecT/SFT) across multiple model sizes. However, we did evaluate the pretrained models of the `gemma-3` model family (1B/4B/12B/27B) on which our 12B system was based in order to get a feel for the importance of model size. In general, we observe the expected trend that bigger is better. However, there does seem to be a particular jump in performance from 1B to 4B. Additionally, on the larger datasets where dataset-specific SFT helped (MP/CSC), our 12B system outperforms the 27B system without SpecT/SFT.

5 Discussion and Conclusion

In summary, our system was able to perform strongly across the board and was **the overall winner on both tasks**. However, it did perform particularly well (1st) on MP and CSC, which had many unique annotators and larger training datasets, and performed less well on Par/VEN (perspectivist: 2nd on Par, 2nd on VEN; soft: 1st on Par, 3rd on VEN), which had only four annotators each and much smaller training sets.

Our approach has many advantages, including: 1) a single model for each dataset, 2) potential adaptation at test time to new raters; 3) strong performance even in the limited data regime; 4) no dataset-specific assumptions; 5) same system for perspectivist and soft tasks. However, some limitations include expensive inference (see App A.3),[3] as prompt lengths are quite long in order to contain in-context rater examples and that the method is unable to effectively leverage additional rater demonstrations that do not fit in the context window.

In our ablation study, we found that in-context demonstrations are crucial for performance, dataset-specific tuning helps given enough data, Spectrum Tuning helped on MP, and performance improves with model size (but scale alone does not make up for dataset-specific training).

Some interesting directions for future work include: 1) how performance scales with the number of in-context rater examples (including going beyond 3,000-token prompts), 2) whether selecting particular in-context examples at inference can outperform random selection, 3) the effect of including rater explanations on performance, 4) how well the approach generalizes to free-text / non-categorical tasks, and 5) methods to better extract dataset-specific signal from smaller datasets (e.g., Par/VEN).

Acknowledgments

This research was supported in part by DARPA under the ITM program (FA8650-23-C-7316).

[3]Although, this could be further optimized with techniques such as prompt caching (Gim et al., 2024).

References

Lora Aroyo, Alex S. Taylor, Mark Diaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-Garcia, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *Preprint*, arXiv:2306.11247.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

In Gim, Guojun Chen, Seung seob Lee, Nikhil Sarda, Anurag Khandelwal, and Lin Zhong. 2024. Prompt cache: Modular attention reuse for low-latency inference. *Preprint*, arXiv:2311.04934.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, page 1–19. ACM.

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.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Preprint*, arXiv:2404.16019.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Metaicl: Learning to learn in context. *Preprint*, arXiv:2110.15943.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. *Preprint*, arXiv:2502.20897.

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

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Taylor Sorensen, Benjamin Newman, Jared Moore, Chan Park, Jillian Fisher, Niloofar Mireshghallah, Liwei Jiang, and Yejin Choi. 2025b. Spectrum tuning: Post-training for distributional coverage and in-context steerability. `https://tsor13.github.io/files/spectrum.pdf`. Preprint.

Joaquin Vanschoren. 2018. Meta-learning: A survey. *Preprint*, arXiv:1810.03548.

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.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. *Preprint*, arXiv:2111.02080.

A Implementation details

A.1 General Details

- All experiments were carried out using 1-4 80GB A100s.
- For all experiments, since the gemma-3-pt models (Gemma Team et al., 2025) do not have a trained embedding for <start_of_turn>/<end_of_turn>, we copy over the (un/)embedding weights for these tokens from the gemma-3-it models, as in (Sorensen et al., 2025b).
- Our SpecT model is an early version of the model from (Sorensen et al., 2025b). For more details, see App. D.

A.2 Dataset-specific SFT hyperparameters

- Training hardware: 4 80GB A100s
- max_length: 1024
- per_device_train_batch_size: 1
- gradient_accumulation_steps: 4
- learning_rate: 1e-6

A.3 Inference Details

All inference was done on a single 80GB A100. MP needed a single forward pass per test rating: ($p(\{\texttt{"0"},\texttt{"1"}\})$), CSC also needed a single forward pass, ($p(\{\texttt{"1", "2", "3", "4", "5", "6"}\})$), Par required three forward passes, ($p(\{\texttt{" "," -"}\})$, $p(\{\texttt{"1", "2", "3", "4", "5"}\}|\texttt{" -"})$, $p(\{\texttt{"0", "1", "2", "3", "4", "5"}\}|\texttt{" "})$), and VEN required four forward passes $p(\{\texttt{"entailment","contradiction", "neutral"}\})$, ($p(\{\texttt{"entailment","contradiction", "neutral"}\})$, $p(\{\texttt{" neutral", " contradiction", "\}"}\}|\texttt{"entailment"})$, $p(\{\texttt{" entailment", " contradiction", "\}"}\}|\texttt{"neutral"})$, $p(\{\texttt{" entailment", " neutral", "\}"}\}|\texttt{"contradiction"})$).

The approximate run time for each inference pass on the entire test set was:

- MP: 23 hours, 30 minutes;
- CSC: 4 hours;
- Par: 11 minutes;
- VEN: 11 minutes;

This was not well optimized however, and could potentially be sped up with methods such as prompt caching (Gim et al., 2024) or vLLM (Kwon et al., 2023).

B Tie calculation

For calculating ties/significance, we used the competition organizer's code for the Wilcoxon signed-rank test to compare entries, as follows: "For each of the four datasets and tasks, to determine ranking, we compared each team to the leading system within a cluster using the Wilcoxon signed-rank test on item-level results from the test sets. Teams were compared sequentially to the leader, and as long as no statistically significant difference was observed, they were assigned the same rank. This process continued until a team showed statistically distinct performance, at which point a new rank was introduced." (quoted from the competition organizers (Leonardelli et al., 2025))

C Prompts

Here, we include example prompts for the four datasets.

lewidi_csc_sarcasm_detection_individual

```
Given a conversational context and response,
  ↪ rate how sarcastic the response is on a 1-6
  ↪  scale.
Annotator demographics: Gender: Female; Age: 26
{"context": "Steve has been going out non-stop
  ↪ for the past two months because he needs a
  ↪ distraction from his recent breakup.  You
  ↪ are worried that he might be becoming a bit
  ↪  too destructive.  Steve says, \"ugh, worst
  ↪  hangover yet. I feel like crap.\"", "
  ↪ response": "maybe try some selfcare", "lang
  ↪ ": "en"}
<start_of_turn>3<end_of_turn>
{"context": "You and Steve have long been
  ↪ planning to go to a new bar in town. But,
  ↪ he has canceled on you three times without
  ↪ telling you why. And just now, he calls you
  ↪  and says, \"I'm so sorry, but I'm gonna
  ↪ have to bail again. Next time?\"", "
  ↪ response": "yeah let me know when you've
  ↪ made the plans", "lang": "en"}
<start_of_turn>1<end_of_turn>
{"context": "Steve talks about the differences
  ↪ between two types of dinosaurs for an hour.
  ↪  You absolutely don't care about the topic
  ↪ .", "response": "anyways... next topic", "
  ↪ lang": "en"}
<start_of_turn>1<end_of_turn>
{"context": "Steve borrowed your spare phone
  ↪ charger two months ago. Then he took your
  ↪ toaster a month ago. He did not return any
  ↪ of them.  And now, Steve says, \"can I
  ↪ borrow your suitcase? I need one for my
  ↪ trip next week.\"", "response": "not really,
  ↪  i think I'm going to need it on the
  ↪ weekend", "lang": "en"}
<start_of_turn>3<end_of_turn>
{"context": "Steve bought a really expensive
  ↪ pair of shoes as a treat to himself for
  ↪ having finished a big project at work. The
  ↪ shoes go very well with his outfit today.",
  ↪  "response": "nice shoes!", "lang": "en"}
<start_of_turn>1<end_of_turn>
{"context": "Steve recently changed jobs. He is
  ↪ annoyed because he needs to deal with some
  ↪ bureaucracy regarding his health insurance.
  ↪  He says, \"I should have just stayed at my
  ↪  old job. If it hadn't been for this new
  ↪ job, I wouldn't have had to deal with so
  ↪ much crap.\"", "response": "maybe that's
  ↪ something you should've researched before
  ↪ but potentially ask for help or spend some
  ↪ time actually figuring this out.", "lang":
  ↪ "en"}
...
```

lewidi_mp_irony_detection_individual

```
Given a post-reply pair from social media (
  ↪ Twitter/Reddit), determine whether the
  ↪ reply is ironic given the post. Context
  ↪ includes platform source, reply depth level,
  ↪  language variety, and language code.
  ↪ Binary irony detection task.
Annotator demographics:
{"post": "My company have basically said we can
  ↪ work from home if we feel safer doing so...
  ↪  but only with our direct manager's
  ↪ approval.\nBut no one has the stones to
  ↪ make the first move on my team. Plenty of
  ↪ other teams have people at home now. But my
  ↪  team get the vibe our manager would be a
  ↪ bit shit if we started.\nHonestly I would
  ↪ definitely feel safer. I can work 100%
  ↪ remote, and my office is  giant open plan
  ↪ with nearly 1000 people who are constantly
  ↪ travelling for work, so if this actually
  ↪ kicks off it'll be a fair nightmare for
  ↪ spreading.", "reply": "Just bite the bullet
  ↪  and ask better safe than sorry worst they
  ↪ can say is no.", "source": "reddit", "level
  ↪ ": "1.0", "language_variety": "ie", "lang":
  ↪  "en"}
<start_of_turn>0<end_of_turn>
{"post": "I\u2019ve heard it all now. Albanese
  ↪ has described himself as being \
  ↪ u201cEconomically Literate\u201d.", "reply":
  ↪  "@USER Of course he is. Don't forget he
  ↪ said he was an economic adviser to Bob
  ↪ Hawke. Trouble is Bob didn't know that and
  ↪ neither did anybody else.", "source": "
  ↪ twitter", "level": "1.0", "language_variety
  ↪ ": "au", "lang": "en"}
<start_of_turn>1<end_of_turn>
{"post": "Bit worried about it actually. Work in
  ↪  health care and I have asthma.   If I do
  ↪ get it. I am going to be as sick as
  ↪ anything.", "reply": "Fingers crossed you
  ↪ don't! I work in retail and surrounded by
  ↪ people who decide that shopping is the best
  ↪  idea when suffering with colds and
  ↪ sickness bugs. A bit like the health care
  ↪ sector cos I worked there too!", "source":
  ↪ "reddit", "level": "1.0", "language_variety
  ↪ ": "gb", "lang": "en"}
<start_of_turn>0<end_of_turn>
{"post": "Can't get it without being anti-
  ↪ national.", "reply": "Nah , everyone will
  ↪ get it", "source": "reddit", "level": "1.0",
  ↪  "language_variety": "in", "lang": "en"}
...
```

lewidi_par_paraphrase_detection_individual

```
Given a pair of questions from Quora Question
  ↪ Pairs (QQP), assign a Likert scale score
  ↪ from -5 to 5 indicating how strongly the
  ↪ questions are paraphrases of one another,
  ↪ and provide a short explanation for your
  ↪ score.
Annotator demographics: annotator_id: Ann1;
  ↪ Gender: Male; Age: 26; Nationality: Chinese
  ↪ ; Education: master student
{"question1": "What are some things new
  ↪ employees should know going into their
  ↪ first day at Exact Sciences?", "question2":
  ↪  "What are some things new employees should
  ↪  know going into their first day at Garmin
  ↪ ?", "lang": "en"}
<start_of_turn>{"paraphrase_rating": -1,
  "explanation": "The companies are
  different."}<end_of_turn>
{"question1": "Who are the everyday heroes and
  ↪ heroines of life?", "question2": "What was
```

```
everyday life like under Nazi rule?", "lang": "en"}
<start_of_turn>{"paraphrase_rating": -5, "explanation": "Q1 asks about everyday heroes and heroines. Q2 is aobut everyday life under nazi rule"}<end_of_turn>
{"question1": "What does 'sandiaga' mean?", "question2": "What does \u064a\u0639\u0646\u064a mean?", "lang": "en"}
<start_of_turn>{"paraphrase_rating": -1, "explanation": "The words to be translated are different."}<end_of_turn>
{"question1": "What is the best way to become a voracious reader?", "question2": "What is a voracious reader?", "lang": "en"}
<start_of_turn>{"paraphrase_rating": -3, "explanation": "Q1 is about how to become a voracious reader. Q2 is about what is a voracious reader."}<end_of_turn>
{"question1": "How do I believe in myself more?", "question2": "How can I believe in myself ?", "lang": "en"}
<start_of_turn>{"paraphrase_rating": 3, "explanation": "The person in Q2 might not believe in themselves at all. The person in Q1 might have already believed in themselves."}<end_of_turn>
{"question1": "What is does \"get swoll\" mean and where does it stem from?", "question2": "Where did leafy get his name from?", "lang": "en"}
...
```

lewidi_variernli_nli_detection_individual

```
Given a premise and hypothesis from MNLI corpus, assign one or more labels from {Entailment, Neutral, Contradiction} indicating the logical relationship between them, and provide an explanation for your reasoning.
Annotator demographics: Gender: Female; Age: 22; Nationality: Chinese; Education: master student
{"context": "Even if auditors do not follow such other standards and methodologies, they may still serve as a useful source of guidance to auditors in planning their work under GAGAS.", "statement": "GAGAS requires strict compliance for auditors to follow.", "lang": "en"}
<start_of_turn>{"nli_label": "entailment", "explanation": "The context suggests that auditors need to plan their work under GAGAS."}<end_of_turn>
{"context": "In May 1967, Gallup found that the number of people who said they intensely disliked RFK--who was also probably more intensely liked than any other practicing politician--was twice as high as the number who intensely disliked Johnson, the architect of the increasingly unpopular war in Vietnam.", "statement": "Due to his attitudes on cheesecake, RFK was more disliked than Johnson.", "lang": "en"}
<start_of_turn>{"nli_label": "neutral", "explanation": "The reason why RFK was more disliked than Johnson is not mentioned in the context."}<end_of_turn>
{"context": "It was made up to look as much like an old-fashioned steam train as possible.", "statement": "It was built in the modern era to look like something built in the past.", "lang": "en"}
<start_of_turn>{"nli_label": "entailment", "explanation": "The context mentions the building of an old-fashioned train, the word old-fashioned would only be used in the modern era. So the statement is true."}<end_of_turn>
{"context": "Today it is possible to buy cheap papyrus printed with gaudy Egyptian scenes in almost every souvenir shop in the country, but some of the most authentic are sold at The Pharaonic Village in Cairo where the papyrus is grown, processed, and hand-painted on site.", "statement": "The Pharaonic Village in Cairo is the only place where one can buy authentic papyrus.", "lang": "en"}
...
```

D SpecT Implementation

The model used in our system was an early version of the model from Sorensen et al. (2025b). The differences between our submission version and the final model are 1) a slightly modified prompt structure (see examples for details), 2) a slightly smaller dataset mix (see App. D), and 3) an earlier hyperparameter set.

Hyperparameters:

- Training hardware: 4 80GB A100s
- `max_length: 1024`
- `per_device_train_batch_size: 1`
- `gradient_accumulation_steps: 512`
- `learning_rate: 3e-6`

Here is the subset of datasets from Sorensen et al. (2025b) that were used in training our system:

```
ambient_ambiguity_detection
ambient_disambiguation
ambient_interpretation_labels
ambient_linguist_annotations
ambient_premise_hypothesis
babynames
bare_enron
bare_gsm8k
bare_hotpot
bare_lcb
binomial
cards
categorical
changemyview_categories
changemyview_posts
chatbotarena_assistant
chatbotarena_individual_prefs
chatbotarena_prompts
coinflip
dices
diffuse_distribution
```

```
flight
generativesocialchoice_freetext
generativesocialchoice_validation
geometric
geometric_beta
globaloqa
gsm8k_answer_from_question
gsm8k_question
gsm8k_question_answer
gsm8k_question_from_answer
habermas_categorical
habermas_individual
habermas_individual_categorical
habermas_opinions
habermas_question
haikus
hatespeech_comment
hatespeech_individual
helpsteer
hypergeometric
imdb
issuebench
jeopardy_answer_prediction
jeopardy_question_generation
multinomial
negative_binomial
netflix_individual_ratings
netflix_individual_views
newsgroups
normal
novacomet_hypothesis
novacomet_premise
numbergame_individual
numbergame_perc
opinionqa_individual
opinionqa_questions
polis_comment
polis_vote
poisson
popquorn_individual
popquorn_og_categorical
prism_prompts
prism_prompts_individual
pubmed
titanic_all_variables
titanic_survival_prediction
valueconsistency
valueprism_misc
valueprism_situation
valueprism_vrd
valueprism_vrds_noncontextual
wvs_individual
zipfian
```

PromotionGo at LeWiDi-2025: Enhancing Multilingual Irony Detection with Data-Augmented Ensembles and L1 Loss

Ziyi Huang[1], **N. R. Abeynayake**[2], **Xia Cui**[2]

[1]Hubei University, Wuhan, China.
ziyihuang@hubu.edu.cn
[2]Manchester Metropolitan University, Manchester, UK.
{n.abeynayake, x.cui}@mmu.ac.uk

Abstract

This paper presents our system for the Learning with Disagreements (LeWiDi-2025) shared task (Leonardelli et al., 2025), which targets the challenges of interpretative variation in multilingual irony detection. We introduce a unified framework that models annotator disagreement through soft-label prediction, multilingual adaptation and robustness-oriented training. Our approach integrates tailored data augmentation strategies (i.e., lexical swaps, prompt-based reformulation and back-translation) with an ensemble learning scheme to enhance sensitivity to contextual and cultural nuances. To better align predictions with human-annotated probability distributions, we compare multiple loss functions, including cross-entropy, Kullback–Leibler divergence and L1 loss, the latter showing the strongest compatibility with the Average Manhattan Distance evaluation metric. Comprehensive ablation studies reveal that data augmentation and ensemble learning consistently improve performance across languages, with their combination delivering the largest gains. The results demonstrate the effectiveness of combining augmentation diversity, metric-compatible optimisation and ensemble aggregation for tackling interpretative variation in multilingual irony detection.

1 Introduction

Irony is a complex linguistic phenomenon in which the intended meaning of an utterance diverges from, or even contradicts, its literal expression. It often relies on contextual incongruity, implicit stance, or shared background knowledge, making it highly dependent on both linguistic and pragmatic cues. This complexity renders irony detection a particularly challenging task for computational systems, especially when extended to multilingual and multicultural contexts where the expression and interpretation of irony may vary substantially.

In such settings, human annotators frequently disagree on whether a given utterance is ironic. This disagreement stems not only from the inherent ambiguity of language but also from differences in cultural norms, humour styles, and pragmatic expectations. The *MultiPiCo* (Multilingual Perspectivist Irony Corpus, MP) (Casola et al., 2024), used in the LeWiDi-2025 shared task (Leonardelli et al., 2025), explicitly captures this variability by providing *soft labels* (i.e., empirical distributions over annotator judgments) rather than single hard labels. Modelling these distributions requires systems capable of representing annotation uncertainty and preserving distributional information in both training and inference phases.

We address this challenge by adopting a *perspectivist* framing of irony detection that emphasises multilingual generalisation and probabilistic supervision. Our system is built upon a multilingual transformer, such as `XLM-R` (Conneau et al., 2020), and incorporates several task-specific strategies: (1) a document representation pipeline that encodes post–reply pairs to preserve conversational context; (2) three targeted data augmentation methods to increase data diversity while maintaining semantic fidelity: *swap*, *prompt* and *translation*; (3) an ensemble training scheme to improve robustness and reduce variance; and (4) the use of an L1 loss function to directly optimise for the task's evaluation metric, *average Manhattan Distance*, which better reflects annotator agreement patterns than conventional cross-entropy loss.

Our contributions are as follows:

- A multilingual irony detection system that models soft labels using a transformer-based architecture aligned with human annotation distributions;
- Novel data augmentation techniques tailored to multilingual context-dependent irony detection;
- An ensemble-based training strategy that improves prediction stability under consistent modelling assumptions;

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 242–248
November 8, 2025 ©2025 Association for Computational Linguistics

- Empirical evidence supporting the use of Manhattan Distance as both an evaluation and optimisation target for soft-label learning.

Our system is designed to be robust, interpretable and adaptable across languages and cultural contexts, offering insights for future work on perspectivist approaches to subjectivity-driven language understanding tasks.

The source code for this paper is publicly available on GitHub[1].

2 Related Work

Irony and Sarcasm Detection. Detecting irony and sarcasm in text has been a long-standing challenge in computational linguistics due to its reliance on implicit meaning, context, and cultural cues. Early approaches relied on handcrafted features such as sentiment contrast or punctuation patterns (Davidov et al., 2010; Reyes et al., 2013). With the rise of deep learning, more robust methods using recurrent networks and attention mechanisms were introduced (Ghosh and Veale, 2016; Tay et al., 2018). Recent work has explored context-aware transformers, modelling not just the utterance but also conversational history or speaker intent (Bamman and Smith, 2021). While effective in monolingual settings, extending irony detection to multilingual and multicultural contexts remains an open problem, especially under limited annotated data.

Soft Labels and Annotator Disagreement. Standard supervised learning assumes a single ground truth label per instance, but tasks involving subjectivity, such as irony detection, frequently involve disagreement among annotators. This has motivated soft-label learning approaches that model the label distribution rather than a hard aggregated majority vote (Pavlick and Kwiatkowski, 2019). Soft supervision helps systems reflect uncertainty and align more closely with human perception. Galstyan and Cohen (2008) and Rizzi et al. (2024) provide comprehensive analyses of training objectives under soft labels, highlighting the inadequacy of cross-entropy loss and advocating for distance-based losses such as Manhattan Distance. These insights directly inform our use of L1 loss in both training and evaluation.

Multilingual Modelling and Data Augmentation. Multilingual pretraining has significantly advanced NLP systems' ability to generalise across languages. Models such as `mBERT` (multilingual BERT) and `XLM-R` (XLM-RoBERTa) have shown strong performance in zero-shot and few-shot cross-lingual transfer (Pires et al., 2019; Conneau et al., 2020). In tasks like irony detection, where training data may be imbalanced across languages, data augmentation becomes especially valuable. Prior work has applied back-translation (Sennrich et al., 2016), prompt-based reformulations (Bao et al., 2020) and contextual rewrites to enhance diversity. In line with these, we adopt a multilingual data augmentation framework that includes swapping discourse segments, prompt injection, and LLM-based translation to increase robustness across languages and cultural contexts.

3 System Overview

We consider the irony detection problem as a binary classification task. Given a set of dialogues D composed of post-reply pairs $(x^{(i)}_{\text{post}}, x^{(i)}_{\text{reply}})$, $\mathcal{D} = \{(x^{(i)}_{\text{post}}, x^{(i)}_{\text{reply}})\}_{i=1}^{N}$, N is the total number of instances in $\mathcal{D}$. Each instance is annotated with a probability distribution over labels $\mathbf{y}^{(i)} \in [0,1]^n$, where n is the number of annotators and $\sum_{j=1}^{n} y_j^{(i)} = 1$. The task is to train a model f_θ that maps each input pair to a predicted soft label distribution $\hat{\mathbf{y}}^{(i)} = f_\theta(x^{(i)})$, such that the average Manhattan Distance between predictions and target distributions is minimised.

3.1 Document Representation

To model the pragmatic and contextual signals that characterise irony, we use `xlm-roberta-base`[2], a multilingual transformer pretrained on 100+ languages. Each instance is represented as a concatenation of the post and its corresponding reply. Tokenisation is handled using the official Hugging Face tokeniser, preserving the consistency of subword units with the model's pretraining.

Let x = `[post] + [SEP] + [reply]` denote the tokenised input string. This is encoded into contextualised embeddings by the transformer encoder and passed through a linear projection followed by a softmax to produce the output distribution $\hat{\mathbf{y}}$.

3.2 Dataset Preprocessing

We use the official training and development splits provided by the shared task organisers. Each instance consists of a "*post*", a "*reply*", and a soft

[1] https://github.com/YhzyY/LeWiDi2025

[2] https://huggingface.co/FacebookAI/xlm-roberta-base

label distribution aggregated from multiple annotators. Instances are preprocessed and wrapped into a custom PyTorch dataset class, `MPDataset`, which encodes each post-reply pair jointly. This allows the model to account for discourse-level semantics essential for detecting irony.

3.3 Data Augmentation

To enhance robustness and reduce overfitting, we apply three task-specific data augmentation strategies, forming an augmented training set:

$$\mathcal{D}_{\text{train}} = \mathcal{D} \cup \mathcal{A}(\mathcal{D}) \tag{1}$$

where $\mathcal{A}$ denotes the augmentation pipeline. The following methods are used:

Swap. We reverse the order of the post and reply in each input sequence. This exposes the model to discourse variation and helps it focus on content and tone rather than fixed positional patterns.

Prompt-based Reformulation (Prompt). We prepend an instruction-style prompt to the input:

> “Given the following post: [post] and reply: [reply], determine whether irony can be detected.”

This method conditions the model on the task and improves generalisation, particularly in multilingual settings where implicit task signals vary.

Translation. Using `gpt-3.5-turbo`[3], we translate the original input into the nine target languages (Arabic, Dutch, English, French, German, Hindi, Italian, Portuguese and Spanish). Each translated version is treated as an augmented instance, inheriting the original soft label. This expands the dataset 9-fold and promotes cross-lingual robustness. The prompt ensures consistency and tone preservation:

> “Translate its ‘text’ part into 9 languages: Arabic, Dutch, English, French, German, Hindi, Italian, Portuguese and Spanish. Pay attention: the translation should preserve the ironic tone in the original dialogues.”

3.4 *K*-fold Ensemble Strategy

To further increase robustness, we use an ensemble training setup. The dataset is randomly shuffled and split into K equally sized subsets $\{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$. For each subset, we train an independent model $f_\theta^{(k)}$. The final prediction for an instance is the unweighted average of all K outputs:

$$\hat{\mathbf{y}} = \frac{1}{K}\sum_{k=1}^{K} f_\theta^{(k)}(x) \tag{2}$$

This approach reduces variance and helps the system better handle ambiguity and soft supervision (Lakshminarayanan et al., 2017).

3.5 Training and Optimisation

We use Hugging Face’s `Trainer` to train the model with a standard configuration (batch size, learning rate, epochs) tuned empirically. All models are fine-tuned on the task-specific data using the L1 loss.

Assume N denote the number of training instances and n for the number of classes, let $\hat{\mathbf{y}}^{(i)} = f_\theta(x^{(i)})$ be the predicted distribution and $\mathbf{y}^{(i)}$ the target distribution. Following Rizzi et al. (2024), we use an evaluation metric between these two distributions, *Average Manhattan Distance* (AvgMD), defined as:

$$\text{AvgMD} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n}\left|\hat{y}_j^{(i)} - y_j^{(i)}\right| \tag{3}$$

To align the optimisation with this metric, we train the model using the *L1 Loss*:

$$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n}\left|\hat{y}_j^{(i)} - y_j^{(i)}\right| \tag{4}$$

This loss provides a more faithful learning signal than cross-entropy in soft-label scenarios, especially for modelling disagreement among annotators (Rizzi et al., 2024).

4 Dataset and Experimental Setup

We use the *MultiPiCo* (MP) corpus (Casola et al., 2024), a multilingual dataset comprising short post–reply exchanges collected from Twitter and Reddit. Each reply is annotated in the context of the preceding post by approximately five crowd workers. Annotators label whether the reply is ironic, resulting in a binary classification task. Instead of collapsing annotations into hard labels, the corpus provides soft labels (i.e., distributions over the two classes) to preserve inter-annotator disagreement and enable models to learn from nuanced and perspectivist supervision.

[3] https://platform.openai.com/docs/models/gpt-3.5-turbo

The MP corpus spans nine languages: Arabic, Dutch, English, French, German, Hindi, Italian, Portuguese and Spanish. This makes it particularly well-suited for evaluating systems in multilingual and cross-cultural settings. The distribution of training, development, and test instances for each language is shown in Table 1.

Table 1: Number of training, development, and test instances per language in the MP corpus.

Language	**#Train**	**#Dev**	**#Test**
Arabic	1,399	363	419
Dutch	637	147	216
English	1,920	489	590
French	1,137	276	347
German	1,513	358	504
Hindi	505	132	149
Italian	646	159	195
Portuguese	1,286	325	383
Spanish	2,974	756	953

We strictly adhered to the official data splits provided by the shared task organisers for training, development and evaluation. No external resources or additional data were used at any stage. This ensures that our system is evaluated under the same constraints as other submissions, and that performance comparisons remain valid.

We use the hyperparameters provided in the `transformers` library for the `xlm-roberta-base` model. Training is performed using the Adam optimiser with a linear learning rate schedule and early stopping based on validation loss. All preprocessing, tokenisation and batching are handled using the Hugging Face framework.

5 Results

Our system achieved competitive performance in the LeWiDi-2025 shared task, highlighting the effectiveness of its multilingual architecture and soft-label modelling approach. In the following subsections, we present detailed evaluations, including augmentation and loss function ablations, ensemble comparison, as well as cross-lingual analyses. All the following experiments are using the same training arguments, the only differences between them are the data augmentation methods, loss function, and ensemble method. Due to the absence of publicly released gold labels for the test set, most validation results were obtained via the Codabench evaluation platform. Language-wise analyses (Section 5.4) could not be conducted on the hidden test set; for these, we report results on the development set instead.

5.1 Effect of Data Augmentation

We evaluated the impact of each proposed augmentation strategy (*swap*, *prompt* and *translation*), as well as their combinations. Table 2 reports the average Manhattan Distance (AvgMD; lower is better) for systems trained under different augmentation settings. All experiments in this section use the L1 loss function without ensemble methods, ensuring that the effects of data augmentation are measured in isolation. Results show that combining augmentation methods consistently outperforms individual ones, with the best performance obtained by using all three techniques or the *swap+translation* pairing. Among single strategies, *translation* yields the largest gain over the baseline, while *swap* and *prompt* produces only marginal improvements. This suggests that semantic-preserving transformations (e.g., translation) contribute more than structural manipulations when modelling irony across languages.

Table 2: AvgMD for different data augmentation configurations.

Augmentation	AvgMD
All Combined	**0.407**
Swap + Translation	**0.407**
Prompt + Translation	0.410
Translation	0.411
Swap + Prompt	0.428
No Augmentation	0.451
Prompt	0.464
Swap	0.473

5.2 Effect of Loss Function

We next compared three loss functions, cross-entropy (CE), L1 and KL divergence, on the baseline system without data augmentation and ensemble. (Figure 1). L1 loss achieved the lowest AvgMD, outperforming CE by 0.034 absolute points, confirming its suitability for aligning predictions with human-annotated distributions in the presence of label uncertainty. KL divergence performed substantially worse, likely due to over-sensitivity to distribution mismatches in low-resource or highly ambiguous cases. These results motivate our final system design, which integrates L1 loss with combined data augmentation to maximise robustness and cross-lingual generalisation.

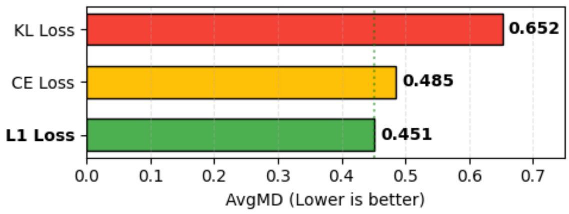


Figure 1: Performance using different loss functions (no augmentation).

5.3 Effect of Ensemble Training

We assess the impact of the ensemble approach using the L1 loss function in combination with different data augmentation settings. The ensemble is constructed by randomly shuffling the training set and partitioning it into five equally sized subsets ($K = 5$), each used to train an independent model. During inference, all models produce soft-label predictions on the test set, which are then averaged to form the final output. This averaging mitigates variance, reduces prediction noise, and improves robustness, particularly in the presence of noisy or ambiguous labels such as those found in irony detection.

Table 3 compares the ensemble results with their single-model counterparts under identical loss and augmentation settings. In all configurations, the ensemble achieves a lower AvgMD. The largest gain occurs with *swap* augmentation, which achieves the highest AvgMD. When no ensemble was applied, its AvgMD drops by 0.049 (from 0.473 to 0.424). Even the smallest gain, observed with *swap+prompt* augmentation, still yields a reduction of 0.008. These consistent improvements highlight the ensemble's ability to capture complementary decision patterns from models trained on different data partitions, leading to more stable and accurate soft-label estimations.

Table 3: AvgMD improvements from applying the ensemble method under different data augmentation settings.

Augmentation	w/o Ensemble	w/ Ensemble	ΔAvgMD
All Combined	**0.407**	0.396	-0.011
Swap + Translation	**0.407**	0.396	-0.011
Prompt + Translation	0.410	0.394	-0.016
Translation	0.411	**0.392**	-0.019
No Augmentation	0.451	0.417	-0.034
Swap + Prompt	0.428	0.420	-0.008
Prompt	0.464	0.428	-0.036
Swap	0.473	0.424	**-0.049**

5.4 Cross-lingual Performance

We assess the system's ability to generalise across the nine target languages using the best-performing configuration for single model(L1 loss with all combined data augmentations). Figure 2 reports AvgMD per language. Performance is better for English and Dutch (AvgMD < 0.4), which have relatively larger training sets and higher lexical similarity to other European languages in the corpus. Spanish and Arabic also perform well despite linguistic differences, suggesting the model effectively leverages cross-lingual transfer. In contrast, Portuguese, French, German and Italian exhibit higher AvgMD values, indicating reduced agreement with human annotations. These discrepancies may stem from smaller data sizes, domain-specific lexical variation, or cultural differences in the expression of irony. Overall, results highlight both the promise and the unevenness of cross-lingual generalisation in perspectivist irony detection.

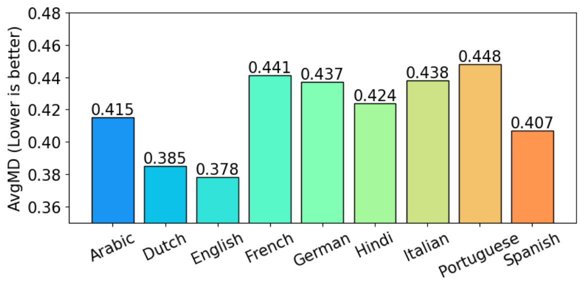


Figure 2: AvgMD per language (lower is better) using L1 loss and all augmentations.

5.5 Overall Result

Under the L1 loss setting without any data augmentation or ensemble, the baseline system achieves an AvgMD of 0.451. As shown in the augmentation ablation study, incorporating data augmentation yields consistent performance improvements, with the best-performing augmentation strategy being the one combined with swap, prompt and translation, which attains an AvgMD of 0.407. This confirms that certain augmentation combinations, particularly those leveraging complementary linguistic variations, can effectively reduce the divergence from human soft labels.

When further integrating the ensemble method, the results demonstrate an even more pronounced improvement. In all cases, the ensemble consistently reduces AvgMD compared to their non-ensemble counterparts, as discussed in the ensemble ablation analysis. The optimal configuration is achieved using the translation augmentation with ensemble, which delivers the lowest AvgMD of 0.392 across all experiments. This outcome aligns with our earlier explanation that ensembles, by aggregating predictions from models trained on diverse data subsets, capture richer and more complementary decision patterns, thereby achieving superior alignment with annotator distributions.

6 Conclusions

We presented a unified system for the LeWiDi-2025 shared task that addresses the challenges of annotator disagreement in multilingual irony detection. Our approach integrates complementary data augmentation strategies, loss functions tailored to the evaluation metric, and an ensemble framework to improve alignment with human-annotated soft labels. Experiments demonstrate that the combination of augmentation and ensemble learning yields substantial reductions in Average Manhattan Distance over strong baselines, with L1 loss proving particularly effective for soft-label prediction under this metric. These findings underscore the value of jointly leveraging data diversity, metric-compatible optimisation, and model aggregation to better capture interpretative variation in multilingual and culturally nuanced NLP tasks. Future work will explore more context-aware augmentation methods and adaptive ensemble schemes to further enhance cross-lingual robustness.

Limitations

While our system demonstrates strong performance in reducing divergence from annotator distributions, several limitations remain. First, our data augmentation strategies, though effective, are primarily heuristic and may not fully capture the full range of linguistic or cultural variability present in real-world irony. Second, the ensemble approach, while improving performance, increases computational cost during both training and inference, which may limit scalability in resource-constrained settings. Third, our experiments focus on multilingual but not truly cross-lingual transfer scenarios; future work should investigate whether the proposed framework generalizes effectively to unseen languages or domains. Finally, although L1 loss proved advantageous for the given metric, its effectiveness for other evaluation criteria remains to be systematically assessed.

Ethical Statements

This study builds upon publicly released datasets provided by the competition organizers, which include multilingual social media content originally collected from platforms such as X and Reddit. All data used are anonymised and intended for research purposes only. We do not introduce any additional user-generated content or external datasets beyond the official competition resources.

Given the subjective nature of irony and the perspectivist framing of this task, we acknowledge the potential for cultural and linguistic biases to influence both human annotations and model predictions. Our system is trained to align with aggregated soft labels that reflect annotator disagreement, however, it may still reflect dominant cultural interpretations embedded in the training data. Additionally, the use of multilingual large language models such as xlm-roberta-base and gpt-3.5-turbo may introduce biases inherited from their pretraining corpora. We encourage careful downstream use of such models and stress the importance of transparency, cultural sensitivity, and critical evaluation when deploying irony detection systems in real-world applications.

References

David Bamman and Noah Smith. 2021. Contextualized sarcasm detection on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):574–577.

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2020. Few-shot text classification with distributional signatures. In *International Conference on Learning Representations*.

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116, Uppsala, Sweden. Association for Computational Linguistics.

Aram Galstyan and Paul R. Cohen. 2008. Empirical comparison of “hard” and “soft” label propagation for relational classification. In *Inductive Logic Programming*, pages 98–111, Berlin, Heidelberg. Springer Berlin Heidelberg.

Aniruddha Ghosh and Tony Veale. 2016. Fracking sarcasm using neural network. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 161–169, San Diego, California. Association for Computational Linguistics.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, page 6405–6416, Red Hook, NY, USA. Curran Associates Inc.

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.

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

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Lang. Resour. Eval.*, 47(1):239–268.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1020, Melbourne, Australia. Association for Computational Linguistics.

twinhter at LeWiDi-2025: Integrating Annotator Perspectives into BERT for Learning with Disagreements

Nguyen Huu Dang Nguyen[1,2] **and Dang Van Thin**[1,2]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
23521045@gm.uit.edu.vn, thindv@uit.edu.vn

Abstract

Annotator-provided information during labeling can reflect differences in how texts are understood and interpreted, though such variation may also arise from inconsistencies or errors. To make use of this information, we build a BERT-based model that integrates annotator perspectives and evaluate it on four datasets from the third edition of the Learning With Disagreements (LeWiDi) shared task. For each original data point, we create a new (text, annotator) pair, optionally modifying the text to reflect the annotator's perspective when additional information is available. The text and annotator features are embedded separately and concatenated before classification, enabling the model to capture individual interpretations of the same input. Our model achieves first place on both tasks for the Par and VariErrNLI datasets. More broadly, it performs very well on datasets where annotators provide rich information and the number of annotators is relatively small, while still maintaining competitive results on datasets with limited annotator information and a larger annotator pool.

1 Introduction

Human language is often subjective and open to interpretation. In many NLP tasks, it's common for annotators to disagree sometimes for good reasons. But most traditional models ignore this variation and treat all labels as if there's only one correct answer. As a result, they may miss out on useful minority viewpoints and become less adaptable.

The third edition of the Learning With Disagreements (LeWiDi) shared task at EMNLP 2025 (Leonardelli et al., 2025) focuses on a critical challenge: building models that learn from disagreements rather than ignore them. The main objective of the task is to provide a unified evaluation framework for learning from disagreements. It introduces a benchmark including four datasets annotated with both soft labels and perspectivist annotations. Here, soft labels represent probability distributions over possible classes, capturing the degree of annotator disagreement, while perspectivist predictions aim to recover the individual label choices of each annotator. Participating teams are evaluated based on how accurately their models predict both types of outputs.

In our approach, we built a simple but effective BERT-based model (Devlin et al., 2019) that makes use of annotator perspectives during training. Instead of collapsing multiple labels into one, we create a separate training instance for each annotator's view and combine it with their background information. This way, the model learns to understand how different kinds of annotators might interpret the same input differently. Our approach performed well across all four shared task datasets. It was especially effective on tasks that involved a small set of annotators and provided natural language explanations alongside their labels.

2 Task Summary

2.1 Dataset

The LeWiDi 2025 shared task provides four diverse datasets across different NLP tasks. Each dataset is accompanied by annotator metadata, including basic demographic information about the annotators who provided the labels. See Table 1 for dataset statistics and Table 2 for available annotator metadata fields.

- **The Conversational Sarcasm Corpus (CSC)** (Jang and Frassinelli, 2024): A dataset of context–response pairs rated for sarcasm, with ratings from 1 to 6.
- **The MultiPico dataset (MP)** (Casola et al., 2024): A crowdsourced multilingual irony detection dataset. Annotators were tasked to detect whether a reply was ironic in the context of a brief post–reply exchange on social media.

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 249–255
November 8, 2025 ©2025 Association for Computational Linguistics

Table 1: Dataset statistics, including task type, instance counts for each split, and annotator information. Unseen annotators refer to annotators whose metadata is not provided.

Dataset	CSC	MP	Par	VariErrNLI
Task	Sarcasm Detection	Irony Detection	Paraphrase Detection	Natural Language Inference
No. of Instances				
Train	5628	12017	400	388
Dev	704	3005	50	50
Test	704	3756	50	50
Annotator Details				
Total annotators	840	506	4	4
Annotators / instance	4, 6	2 - 21	4	2, 3, 4
Unseen annotators	12	0	0	0

Field	Description	Datasets
Annotator ID	Unique identifier	All
Age	Annotator's age at the time of annotation	All
Gender	Self-identified gender	All
Nationality	Annotator's nationality	MP, Par, VariErrNLI
Education	Highest level of education completed	Par, VariErrNLI
Ethnicity (simplified)	The ethnicity of the annotator	MP
Country of birth	Annotator's country of birth	MP
Country of residence	Annotator's current country of residence	MP
Student status	Whether the annotator is a student	MP
Employment status	Annotator's employment status	MP

Table 2: Annotator metadata available across datasets.

Languages include Arabic, German, English, Spanish, French, Hindi, Italian, Dutch, and Portuguese.

- **The Paraphrase Detection dataset (Par)**: [1] A dataset of question pairs for which annotators rated whether the two questions are paraphrases of each other on a Likert scale. In addition to labels, annotators also provided short explanations for their choices.

- **The VariErr NLI dataset (VariErrNLI)** (Weber-Genzel et al., 2024): A dataset originally designed for automatic error detection, distinguishing between annotation errors and legitimate human label variation in Natural Language Inference. Annotators also included short textual explanations for their choices.

2.2 Tasks

The LeWiDi 2025 shared task defines two official evaluation settings. To ensure comparability with the leaderboard, we adopt the same metrics:

Task A (Soft Label Prediction): Given multiple annotator labels per instance, the goal is to predict a probability distribution over possible labels. Models are evaluated on how close the predicted label distribution is to the empirical human label distribution. Manhattan distance is used for binary label datasets (MP, VariErrNLI), and Wasserstein distance is used for ordinal label datasets (Par, CSC).

Task B (Perspectivist Prediction): This task focuses on predicting the individual labels assigned by each annotator. For binary label datasets (MP, VariErrNLI), performance is measured using error rate; for ordinal label datasets (Par, CSC), absolute distance is used.

3 Method

3.1 System Overview

Our model aims to capture how individual annotators see things differently. As shown in Figure 1, we convert each original sample into multiple training instances, each paired with information from a specific annotator. This lets the model pick up on patterns in how different people label the same text.

Dataset Construction: Instead of treating each sample as a single data point, we decompose it into multiple (text, annotator) pairs. Depending on the dataset, adjustments are applied to the input text (e.g., incorporating annotator explanations or source metadata) so that the model can capture how different annotators interpret the same input.

[1]The dataset is maintained by the MaiNLP lab and is not yet published.

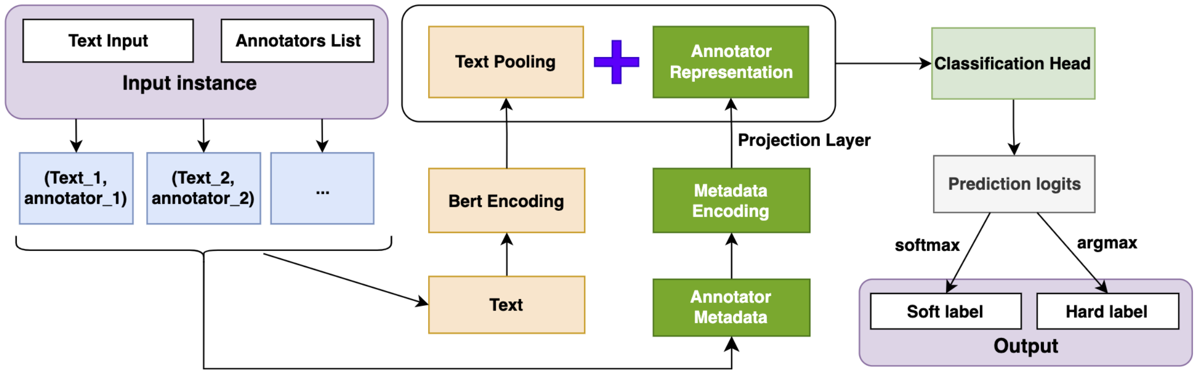


Figure 1: Representation of our BERT with Annotator Information

Detailed processing steps for each dataset are described in Section 3.2.

Input Representation: We encode the input text using a pretrained BERT model to obtain contextualized embeddings. In parallel, the annotator metadata is processed through a projection layer to produce a fixed-size feature vector. These two representations are then concatenated and passed to a classification layer.

Target Construction: Each (text, annotator) pair is treated as a distinct training sample with its corresponding label. This setup enables the model to learn from individual annotator perspectives.

Model Variants: We use MiniLM-L12-H384-uncased (Wang et al., 2020) for the CSC, Par, and VariErrNLI datasets, while DistilBERT-multilingual-cased (Sanh et al., 2019) is employed for MP, which contains multilingual samples.

Training Setup: We train the model using soft label supervision, comparing predictions to full label distributions. Optimization is performed using AdamW (Loshchilov and Hutter, 2019), with dropout, early stopping, and a learning rate scheduler to enhance training stability.

Baselines: We compare against three baselines: the official baseline from the organizers, a TF-IDF + Random Forest (Louppe, 2015) model, and a plain BERT model that doesn't use any annotator information. (furthur described in subsection **??**)

3.2 Text Processing

In all four datasets, each input sample is represented as a pair of textual fields, which we denote as `S1` and `S2`. The concrete meaning of these fields depends on the dataset:

- **CSC**: `S1` = context (the situation preceding the response), `S2` = response.
- **MP**: `S1` = post, `S2` = reply to the post.
- **Par**: `S1` = Question 1, `S2` = Questio n2.
- **VariErrNLI**: `S1` = context (premise), `S2` = statement (hypothesis).

These are concatenated using the `[SEP]` token:

```
S1 [SEP] S2
```

For the Par and VariErrNLI datasets, which include brief natural language explanations written by annotators, we append the explanation (`Exp`) of the corresponding annotator after a second `[SEP]` token:

```
S1 [SEP] S2 [SEP] Exp
```

For the MP dataset, which contains a source metadata field indicating the origin of the input (Reddit or Twitter), we prepend the source before the main text sequence to help the model disambiguate the context. This follows prior work on topic infusion (Sullivan et al., 2023):

```
Source [SEP] S1 [SEP] S2
```

Text Processing in Baseline Models: For the fine-tuned BERT baseline (which does not utilize annotator information), we concatenate all available annotator explanations (if present) and append them to the input sequence. This applies to datasets such as Par and VariErrNLI. For the TF-IDF + Random Forest baseline, we use the same input samples as in our main model, with tokenization performed using TF-IDF vectorization.

3.3 Annotator Metadata Encoding

Annotator metadata is encoded by combining one-hot encoding for categorical features and standard scaling for numerical ones. Missing or invalid values are imputed using the mode. The resulting feature vectors are concatenated into a single metadata representation for each annotator. For the MP dataset, we apply Principal Component Analysis (PCA) (Shlens, 2014), retaining 99.5% of the variance, to reduce the dimensionality from 91 to 31.

4 Experiment Setup

4.1 Comparison Models

We compare the Most Frequent baseline provided by the organizers with three approaches for modeling annotator disagreements and perspectives:

Organizer Baseline (Most Frequent): Two variants are provided by the organizers. (1) For Soft Label Evaluation, the mean label distribution over the training set is used as the prediction for all test items. (2) For Perspectivist Evaluation, each annotator's most frequent label is assigned across all items. Predictions are then evaluated using the respective metrics.

TF-IDF + Random Forest (TF-IDF + RF): For CSC, Par, and VariErrNLI, we extract TF-IDF features from the input text using TfidfVectorize and concatenate them with the annotator vectors. For Par and VariErrNLI, where the number of annotators is relatively small, we train an individual Random Forest regressor for each annotator to better reflect their subjective labeling tendencies. In contrast, for CSC, which includes over 800 annotators, we train a single model using soft labels aggregated across annotators. Due to the multilingual nature of MP, this model is not applicable there.

Fine-tuned BERT (No annotator Information): This baseline ignores annotator identity and treats each instance as a single aggregated sample. We fine-tune a BERT-based encoder using soft labels as targets. Specifically, we use MiniLM-L12-H384-uncased (Wang et al., 2020) for CSC, Par, and VariErrNLI; and DistilBERT-multilingual-cased (Sanh et al., 2019) for MP. This setup serves as a direct comparison point for evaluating the impact of annotator-aware modeling.

Fine-tuned BERT with Annotator Information (Main Model): The model described in subsection 3.1. It takes annotator information into account by treating each (text, annotator) pair as a distinct training sample. We encode the text using a BERT-based model and transform the annotator features via a projection layer. The two representations are then concatenated before classification.

For both **BERT-based models**, we use HuggingFace's `AutoTokenizer` (Wolf et al., 2020) associated with the respective pretrained encoder for text tokenization.

All models are trained using soft label supervision for Task A. For models that incorporate annotator information, we average predictions across annotators to obtain the final output. Predictions for Task B are then derived directly from the outputs of Task A. In contrast, models without annotator information generate a single output distribution, from which Task B labels are obtained via `argmax`.

4.2 Loss Function

For the CSC and Par datasets, which contain ordinal labels, we use Kullback-Leibler (KL) divergence loss for our model and the BERT baseline. The TF-IDF + Random Forest(RF) model is evaluated using the Wasserstein distance as a performance metric. For the MP and VariErrNLI datasets, which involve binary classification tasks, we use L1 loss for training. The TF-IDF + RF model for these datasets is evaluated using the Manhattan distance as a performance metric.

4.3 Evaluation Measures

Evaluation metrics are tailored to each dataset and task, and we follow the official definitions and evaluation scripts provided by the LeWiDi shared task organizers.

Soft Evaluation (Task A):

- **CSC, Par:** Average Wasserstein Distance
- **MP:** Average Manhattan Distance
- **VariErrNLI:** Average Multilabel Average Manhattan Distance

Perspectivist Evaluation (Task B):

- **CSC, Par:** Average Normalized Absolute Distance
- **MP:** Average Error Rate
- **VariErrNLI:** Average Multilabel Error Rate

Metric Summary: The evaluation metrics are designed to capture both aggregate performance (how well predicted distributions align with the overall human label distribution) and perspectivist performance (how well individual annotator perspectives are recovered). For brevity, we omit explicit formulas for commonly used metrics such as Manhattan Distance and Error Rate (and their multilabel variants). For readability, the metrics are presented in summarized form rather than with full mathematical expressions.

Wasserstein Distance (WD): Measures the effort required to transform one distribution into another, assuming ordinal classes:

$$WD(p,t) = \sum_{h=1}^{n}\sum_{k=1}^{n} \min(p_h, t_k) \cdot |h-k|$$

where p and t are discrete distributions over n ordinal categories.

Average Wasserstein Distance (AWD): Let $p^{(i)}$ and $t^{(i)}$ denote the predicted and target distributions for sample i, then:

$$AWD = \frac{1}{N}\sum_{i=1}^{N} WD(p^{(i)}, t^{(i)})$$

The average Wasserstein Distance is 0 in the case of a perfect match.

Normalized Absolute Distance (NAD): For a sample i, let $t_i = [t_{i,1}, \ldots, t_{i,a}]$ be the target labels and $p_i = [p_{i,1}, \ldots, p_{i,a}]$ the predictions for a annotators. The Normalized Absolute Distance is defined as:

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

where s is the Likert scale range. A value of 0 indicates perfect agreement.

Average Normalized Absolute Distance (ANAD): The final score is obtained by averaging NAD over all N samples:

$$\text{ANAD} = \frac{1}{N}\sum_{i=1}^{N} \text{NAD}(i)$$

5 Result

We report system performance across all datasets and evaluation tasks in Table 3, using the metrics described in Section 4.3. Our model consistently outperforms baseline methods on the VariErrNLI and Par datasets, and shows modest improvements over baselines on CSC and MP. According to the official LeWiDi 2025 leaderboard[2], our system ranks **top-5** on both tasks for the CSC and MP datasets, and achieves **1st place** on both tasks for the Par and VariErrNLI datasets. These rankings are consistent across both Task A (soft evaluation) and Task B (perspectivist evaluation).

6 Further Analysis

We conducted further analysis to understand how incorporating annotator information affects model performance. Overall, models that leverage annotator information tend to outperform those that do not.

For the **Par** and **VariErrNLI** datasets, both of our annotator-aware models (TF-IDF + RF and our proposed BERT-based model) consistently surpassed the organizers' baselines and the BERT-based models without annotator information. With a small and fixed set of annotators, the models can more easily capture individual behavior, helping them understand consistency in how samples are labeled. Additionally, the inclusion of textual explanations allows the models to learn multiple perspectives from each annotator, resulting in richer, more multi-dimensional instance representations and reducing ambiguity compared to using raw labels alone.

In contrast, for the **MP** and **CSC** datasets, using annotator information did not lead to much improvement. These datasets only provide basic metadata (e.g., age, gender), and the number of annotators is much larger, making it harder for the model to learn how each annotator behaves. Moreover, some annotator attributes were missing or not provided for certain examples, so we filled in missing values using the mode for each field. This imputation may have introduced noise and reduced the effectiveness of annotator-aware modeling. Still, in **MP**, the model with annotator features performs slightly better. In the case of **CSC**, our proposed model performed worse than the BERT model that does not use annotator information. A likely explanation is that many annotators in **CSC** lack associated metadata. As a result, we had to fill in missing values with default values, which may

[2] More information about the shared task and leaderboard is available at https://le-wi-di.github.io/.

Table 3: System performance across datasets. **Task A** is evaluated using **WD** (Wasserstein Distance) and **MD** (Manhattan Distance). **Task B** is evaluated using **NAD** (Normalized Absolute Distance) and **ER** (Error Rate). Arrows ↓ indicate lower values represent better performance. The *Baseline* model is provided by the organizers. Best performance for each dataset and metric is highlighted in bold.

Model	**CSC**		**MP**		**Par**		**VariErrNLI**	
	WD(↓)	NAD(↓)	MD(↓)	ER(↓)	WD(↓)	NAD(↓)	MD(↓)	ER(↓)
Baseline	1.169	0.238	0.518	**0.316**	3.23	0.36	0.59	0.34
TF-IDF + Random Forest	0.87	0.247	x	x	1.2	0.34	0.42	0.24
BERT - without annotator information	**0.835**	x	0.48	x	2.04	x	0.4	x
BERT - with annotator information	0.86	**0.228**	**0.45**	0.319	**0.98**	**0.08**	**0.23**	**0.12**

have introduced noise into the input representation and negatively affected the model's performance. These results highlight a general challenge: when the annotator pool is large and metadata is sparse or missing, modeling individual annotator behavior may become difficult.

Model rankings in Task B largely reflect those in Task A, indicating that understanding annotator behavior contributes to overall prediction quality. While annotator-aware modeling benefits datasets with small, information-rich annotator pools, generalizing to larger, sparse pools remains challenging. These results suggest that the approach is most effective when annotator numbers are limited and data is semantically rich, but its effectiveness may decrease as the pool grows and label distributions become sparse, highlighting an open question for future research.

7 Conclusion

In this work, we presented a model that predicts labels for each (text, annotator) pair, aiming to capture individual annotator perspectives rather than just aggregated labels. We evaluated our method on four datasets covering sarcasm detection, irony detection, paraphrase detection, and natural language inference. Our results show that including annotator information often leads to better performance, especially in datasets where annotator perspectives are clearly defined and consistent. However, for datasets with many annotators or missing metadata, the improvement is less clear, and in some cases, using annotator features may introduce noise.

Overall, our findings suggest that modelling individual perspectives is a promising direction for tasks involving subjective annotation. Future work may explore more advanced architectures or evaluate on additional datasets to further understand the benefits and limitations of this approach.

Limitations

Our model has several limitations. First, some datasets (e.g., CSC) lack annotator metadata, requiring us to use dummy or average values, which may negatively affect the model's accuracy. Second, our model does not scale well to datasets with a large number of annotators, since each (text, annotator) pair is treated as a separate input. Third, we use a simple architecture that concatenates text and annotator embeddings, without exploring more advanced approaches like attention or expert mixtures. Lastly, we did not compare our approach against some strong solutions such as multi-task learning (Fornaciari et al., 2021), which could provide useful insights.

References

Silvia Casola, Simona Frenda, Soda Marem Lo, Erhan Sezerer, Antonio Uva, Valerio Basile, Cristina Bosco, Alessandro Pedrani, Chiara Rubagotti, Viviana Patti, and Davide Bernardi. 2024. MultiPICo: Multilingual perspectivist irony corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16008–16021, Bangkok, Thailand. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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, 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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Gilles Louppe. 2015. Understanding random forests: From theory to practice. *Preprint*, arXiv:1407.7502.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Jonathon Shlens. 2014. A tutorial on principal component analysis. *Preprint*, arXiv:1404.1100.

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.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Appendix A: Hyperparameter Details

TF-IDF + Random Forest. For CSC, we use `TfidfVectorizer(max_features=4000)` for tokenization. For Par and VariErrNLI, we use `TfidfVectorizer(max_df=0.7, min_df=2, ngram_range=(1, 3))`. Random Forest hyperparameters (`n_estimators`, `max_depth`) are selected via grid search using the validation set.

BERT-based Models. All transformer-based models are optimized using AdamW (`weight decay=0.01`). Training is done with soft label regression.

CSC and MP: We train for 5 epochs with early stopping based on validation loss. We set dropout rate to 0.4, batch size to 32, and learning rate to 2e-5. We use `ReduceLROnPlateau (mode='min', factor=0.5, patience=1)`. Texts are tokenized with `max_len=128`.

Par and VariErrNLI: Models are trained for up to 30 epochs with early stopping. Batch size is 16, dropout rate is 0.3 and learning rate is 2e-5. Texts are tokenized with `max_len=128`.

Classification Head: We use a linear layer followed by a dropout layer and another linear projection to the output logits.

Annotator Projection Layer: Annotator metadata is passed through a linear layer followed by a ReLU activation to obtain a fixed-size embedding vector.

Annotator Projection Sizes: 5 (CSC), 32 (MP), 16 (Par), 16 (VariErrNLI). The size of the text representation corresponds to the encoder's hidden size (e.g., 384 for MiniLM).

Uncertain (Mis)Takes at LeWiDi-2025: Modeling Human Label Variation With Semantic Entropy

Ieva Raminta Staliūnaitė
University of Cambridge
irs38@cam.ac.uk

Andreas Vlachos
University of Cambridge
av308@cam.ac.uk

Abstract

The VariErrNLI task requires detecting the degree to which each Natural Language Inference (NLI) label is acceptable to a group of annotators. This paper presents an approach to VariErrNLI which incorporates measures of uncertainty, namely Semantic Entropy (SE), to model human label variation. Our method is based on the assumption that if two labels are plausible alternatives, then their explanations must be non-contradictory. We measure SE over Large Language Model (LLM)-generated explanations for a given NLI label, which represents the model uncertainty over the semantic space of possible explanations for that label. The system employs SE scores combined with an encoding of the inputs and generated explanations, and reaches a 0.31 Manhattan distance score on the test set, ranking joint first in the soft evaluation of VariErrNLI.[1]

1 Introduction

Annotator disagreement has recently received more attention in NLP research (Fornaciari et al., 2021; Leonardelli et al., 2021; Sandri et al., 2023). Human label variation has consequences for the data, modeling, and evaluation in ML tasks (Plank, 2022). The question of data quality is related to distinguishing legitimate human label variation, which stems from different interpretations or opinions, from errors. In the context where a single label is correct, the problem of determining annotation reliability has been addressed by Hovy et al. (2013), who propose to evaluate the trustworthiness of each annotator in predicting the correct label. Allowing human label variation adds a layer of difficulty to determining whether annotations are valid, since every combination of labels may be correct. Some work has used the difference between annotator entropy and model entropy to predict which instances may require more annotations in an active learning setup (Baumler et al., 2023).

In this work we propose to solve the VariErrNLI task with Uncertainty Quantification (UQ), specifically Semantic Entropy (SE) (Farquhar et al., 2024). This approach expands on the work by Baumler et al. (2023) by including the semantics of the input as well as sampled Large Language Model (LLM)-generated explanations, and applies it to predicting the soft labels themselves rather than quantifying additional annotation needs. SE has mostly been employed to detect hallucinations (Farquhar et al., 2024), where a prediction with a high SE is interpreted as likely to have been hallucinated, given that the model is uncertain over the semantic space of the output. This is in line with prior work on UQ, which focuses on model calibration (Gupta et al., 2006) and detecting noisy training data (Northcutt et al., 2021). Staliūnaitė et al. (2025) propose to use uncertainty metrics such as similarity-sensitive entropy (Cheng and Vlachos, 2024) for detecting bias in machine translation, by leveraging the fact that uncertainty can also arise from ambiguity (Baan et al., 2024). Models should be uncertain not only when they are not apt, but also when the input is ambiguous, where uncertainty is caused by more than one output being acceptable.

2 Task Summary

Weber-Genzel et al. (2024) introduced the Natural Language Inference (NLI)-inspired task VariErrNLI, which contains both 1) valid annotator disagreement and 2) annotation errors. The dataset builds on the ChaosNLI (Nie et al., 2020) dataset, which is composed of NLI items with soft labels. A subset of ChaosNLI instances is annotated from scratch in two rounds by Weber-Genzel et al. (2024), with four annotators providing initial NLI labels, and then returning to evaluate their own as well as their peers' judgments in a second round.

[1] The code is available at https://github.com/ieva-raminta/uncertain-mis-takes

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 256–262
November 8, 2025

Annotations that are self-corrected are interpreted as errors and are not included in the gold label sets.

VariErrNLI is one of the tasks in the LeWiDi shared task (Leonardelli et al., 2025). In this paper we discuss a system that solves VariErrNLI with soft label prediction. That is, for an instance of VariErrNLI, we predict the acceptance rate of each label after the second round of annotation. This creates a multilabel binary classification setup with soft targets, where the score for each label reflects the proportion of annotators who accepted it. The example below illustrates an instance where after the second round of annotations, half the annotators believe that the entailment label is appropriate, three quarters of the annotators accept Neutral as a valid label, and none support the Contradiction label:

Context: "The next year, he built himself a palace, Iolani, which can still be toured in Honolulu."
Statement: "Lolani was built in only 1 year."
Labels: Entailment: 0.5, Neutral: 0.75, Contradiction: 0.0

In the shared task, systems are evaluated with soft labels, measuring how well the predicted label distribution matches the acceptability ratings of the different possible interpretations for each instance, as introduced by Uma et al. (2022). Specifically, Manhattan distance is used to measure the difference between the predicted and target distributions, which has been shown to be particularly reliable for binary classification (Rizzi et al., 2024).

3 System Overview

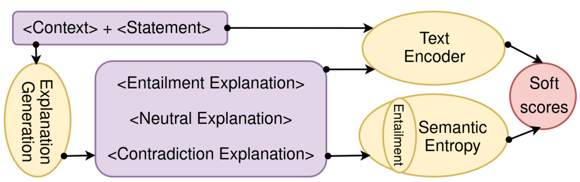


Figure 1: System pipeline: 1. An Explanation Generation (llama3-8B) model generates explanations for each combination of <Context>, <Statement>, and one of Entailment, Neutral, Contradiction labels; 2. Semantic Entropy is calculated for each set of explanations for a given instance using an Entailment model (finetuned bart-large-nli); 3. A Text Encoder (bart-large-nli) is used to embed the combination of <Context>, <Statement> and explanations for each label; 4. Soft scores are predicted from the SE and Text Encoder outputs.

The goal of our system is to be able to quantify the ambiguity in the NLI instances. We postulate that if an instance is ambiguous, the explanations for different labels are likely to not entail one another. For instance, in the example from Section 2, an explanation for the Entailment label could read *"The context states that he built himself a palace next year, which means that he finished it within the year"*, whereas the Neutral label could be explained with *"He may have started to build the palace the next year, but we do not know when he finished it"*. These explanations do not entail each other, which is indicative of ambiguity in the instance. In contrast, explanations for an instance which has only one valid interpretation should only have explanations which entail one another.

Thus, we build a pipeline that uses SE over the explanations for different labels, with the goal of representing the ambiguity of an instance. Predictive Entropy (PE) for an input x is calculated by taking the Shannon entropy of the model's predicted probability distribution over labels. SE is an extension of PE, which is calculated by sampling multiple model outputs, clustering them into sets of sequences of tokens that express the same meaning, and measuring the entropy between the clusters c (Farquhar et al., 2024):

$$\mathrm{SE}(x) = -\sum_{c} P(c \mid x) \log P(c \mid x) \quad (1)$$

Clustering is performed in such a way that any two samples are attributed to the same cluster if and only if they entail one another. The SE of model predictions is higher for instances where more than one interpretation is valid, as more contradictory generated explanations are likely to appear.

We combine the outputs of SE with an embedding of the input and generated explanations for each label. Figure 1 illustrates the full pipeline.

4 Experimental Setup

4.1 Models

This section describes the models used in each component of the pipeline.

Explanation Generation. First, to generate the explanations for each NLI label, we use llama3-8B (AI@Meta, 2024), chosen for its balance of efficiency and reasoning capabilities. The instructions for the model are as follows: *"You are an NLI* *assistant. Given a statement, context, and a judgment label (Entailment, Neutral, or Contradiction), explain why the label is appropriate.*`\n\n` *<Examples>*`\n\n` *Now consider the following example:*`\n` *Statement: <Statement>*`\n` *Context: <Context>*`\n` *Judgment: Contradiction*`\n` *Explanation:".* <Examples> contains a 6-shot list of instances with explanations, two for each label.[2]

Text Encoder. Second, for embedding the inputs along with the generated explanations, we use bart-large-nli (Lewis et al., 2019). bart-large-nli is finetuned on the NLI task, which is highly relevant to the task we are solving, namely predicting soft scores for each NLI label.

Entailment. Third, the calculation of Semantic Entropy over the LLM-generated explanations requires an entailment model for the clustering step. We use bart-large-nli for this step as well. However, we further finetune bart-large-nli on the gold explanations in the VariErrNLI dataset. The data preprocessing for this step is described in Section 4.2. The LLM-generated explanations in our system pipeline are clustered by the finetuned bart-large-nli model to calculate SE.

Semantic Entropy. We follow the implementation of Semantic Entropy by Farquhar et al. (2024). We sample 128 generated explanations for all three NLI labels, cluster them together if and only if they entail each other, and calculate SE over the clusters obtained. We calculate seven SE scores, corresponding to each member of the powerset of NLI labels: ((Entailment), (Neutral), (Contradiction), (Entailment, Neutral), (Entailment, Contradiction), (Neutral, Contradiction), and (Entailment, Neutral, Contradiction)). This formulation allows us to isolate the contribution of each label to the total semantic uncertainty by comparing entropy values across subsets.

4.2 Data

For training the Text Encoder we use the full ChaosNLI dataset (Nie et al., 2020) and generate explanations for each label with an LLM.

For training the Entailment model for clustering explanations in SE calculation, we use the gold explanations from VariErrNLI dataset. Each data point is a set of two explanations from a single instance. We assume that two explanations have the Entailment relation if they explain the same label, and that two explanations have a Neutral relationship if they explain different labels but the annotators accept each others' judgments, and finally that two explanations are Contradictory if the annotators reject each others' judgments.

4.3 Configurations

Table 1 presents the different configuration values that we experiment with. We model the task as either a classification or regression task. In the regression setup we directly predict the probability of a given label being accepted, whereas in the classification task we either predict one of seven real values for each label: (0.0, 0.25, 0.33, 0.5, 0.66, 0.75, 1.0) or predict one of 20 combinations of real values which sum to one: ((0.0, 0.0, 1.0), (0.0, 0.25, 0.75), (0.0, 1.0, 0.0), (0.25, 0.0, 0.75), etc).[3] The classes cover the observed soft label distributions.

For the real-valued prediction setup we use either KL divergence or MSE loss, while for classification we use Cross Entropy loss, and we also experiment with a cross label loss function that incorporates dependencies between labels in multi-label classification (Ferreira and Vlachos, 2019).

Hyperparameter	**Values**
Learning Objective	classification, multilabel classification, regression
Dropout	0.1, 0.3, 0.5
Loss Function	Cross Entropy, Cross Label, KL Divergence, MSE
Learning Rate	1e-1 to 1e-5
Weight Decay	1e-2 to 1e-6
Unfrozen Layers	0, 1, 2, 3
Scheduler	step LR, cosine, linear, reduce on plateau
SE embedding size	8, 16
fusion layer size	128, 256
feature combination method	concatenation, fusion, fusion MLP
Entropy Penalty (β)	0, 0.05, 0.1
Temperature Annealing	1.0, 1.5, 2.0
Regularise Against Mean (λ)	0, 0.05, 0.1
Sum < 1 Penalty (γ)	0, 0.05, 0.1

Table 1: Search space for hyperparameter values, regularisation terms, and other model specifications.

We explore several regularisation methods in order to ensure that the predicted scores do not diverge from the targets. To begin with, in initial runs we observed rather similar predictions for instances where they should differ, and thus

[2] Please see Appendix A for the full list of examples.

[3] Please see Appendix B for the full list of the most common combinations of the binary soft labels.

experiment with (1) entropy penalty, which encourages the model to generate more diverse outputs by penalising low entropy (Grandvalet and Bengio, 2004) and (2) temperature annealing (Kirkpatrick et al., 1983; Hinton et al., 2015). Similarly, with many scores appearing close to the mean distribution of the target values, we add a (3) regularisation against the mean (Szegedy et al., 2016; Pereyra et al., 2017). Finally, in order to ensure that the sum of the predicted scores is no lower than one, we add a (4) penalty to the loss whenever the sum of the three scores is below one. All the penalties are applied to the loss, except for the temperature annealing, which is directly applied to the logits.

$$\mathcal{L}_{\text{entropy}} = \beta \cdot \sum_{i=1}^{N} \sum_{j=1}^{C} p_{ij} \log p_{ij} \tag{2}$$

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{3}$$

$$\mathcal{L}_{\text{mean}} = \lambda \cdot \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{p}_i - \bar{\mathbf{p}}\|_2 \tag{4}$$

$$\mathcal{L}_{\text{sum}} = \gamma \cdot \sum_{i=1}^{N} \max(0, 1 - \sum_{j=1}^{C} p_{ij}) \tag{5}$$

We use three different methods to combine the text embeddings with SE information. The first one is straightforward concatenation. The second one is a fusion model, where both representations are projected onto the same shape and summed with weights that are learned during training. The third one is a fusion Multilayer Perceptron (MLP), where the representations are first concatenated, followed by an MLP layer that learns non-linear interactions between the text and entropy modalities.

5 Results

Our best score on the test set is 0.31 Manhattan distance (lower is better), which is ranked number one in the LeWiDi VariErrNLI task (soft evaluation). It is substantially better than the most frequent baseline score of 0.59, and is only surpassed by a system that reaches 0.23, however the difference is not statistically significant. The configuration that led to the best score of our system is described in Appendix C.

We assess the contribution of each component of our pipeline by running an ablation study and excluding one of: the Semantic Entropy features altogether, the clustering step for SE, the finetuning of the Entailment model, the complete Text Encoder output or the generated explanations. The results on the development set provide us with additional insights into the workings of the system (see Figure 2). We find that the SE component contributes the most to the performance of the model, as the performance drops from 0.37 to 0.54 in Manhattan distance without it. Furthermore, we find that the generated explanations as well as finetuning of the Entailment model are also instrumental in our pipeline. In addition, we find that the methods for incorporating different types of input do not significantly impact the outcomes. We further discover that the best result on the development set is achieved by completely excluding Text Encoding features. However, this SE-only model does not yield the best score on the test set, which we interpret as an indication that the model overfits.

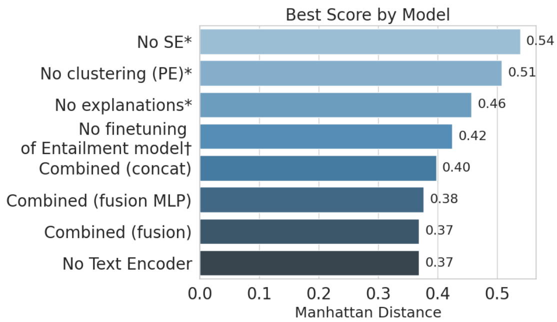


Figure 2: Ablation study results on the development set. Statistical significance between each model and the next best score is marked with * ($p < 0.05$ for all three labels), and † ($p < 0.05$ for at least one of the labels).

6 Conclusion

This work presents an approach to soft label NLI, which proves to yield competitive results. The ablation study shows that SE is the most important module of the system, highlighting its versatility beyond hallucination detection and signal for further annotation needs. In future work this approach could be more specifically applied to detecting annotation errors by learning the different Semantic Entropy patterns associated with annotations that are incompatible with valid interpretations. The proposed method can further be applied to other tasks that include generation and ambiguity.

Limitations

The main limitation of this study is the requirement of an LLM for the explanation generation step. First, generating multiple explanations and calculating SE involves sampling and clustering steps that are computationally expensive, which may limit scalability or real-time applicability in practical settings. Second, our method relies on the quality of the explanations generated by the LLMs or alternatively on human generated explanations, which is labour-intensive.

Acknowledgments

Ieva Raminta Staliūnaitė is supported by Huawei. Andreas Vlachos is funded by the European Union's Horizon 2020 Research and Innovation programme grant AVeriTeC (Grant agreement No. 865958) and a grant from Translated.

References

AI@Meta. 2024. Llama 3 model card.

Joris Baan, Raquel Fernández, Barbara Plank, and Wilker Aziz. 2024. Interpreting predictive probabilities: Model confidence or human label variation? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 268–277, St. Julian's, Malta. Association for Computational Linguistics.

Connor Baumler, Anna Sotnikova, and Hal Daumé III. 2023. Which examples should be multiply annotated? active learning when annotators may disagree. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10352–10371, Toronto, Canada. Association for Computational Linguistics.

Julius Cheng and Andreas Vlachos. 2024. Measuring uncertainty in neural machine translation with similarity-sensitive entropy. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2115–2128, St. Julian's, Malta. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

William Ferreira and Andreas Vlachos. 2019. Incorporating label dependencies in multilabel stance detection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6350–6354, Hong Kong, China. 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.

Yves Grandvalet and Yoshua Bengio. 2004. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

Hoshin V Gupta, Keith J Beven, and Thorsten Wagener. 2006. Model calibration and uncertainty estimation. *Encyclopedia of hydrological sciences*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science*, 220(4598):671–680.

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.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

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.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

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.

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.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don’t you do it right? analysing annotators’ disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Ieva Raminta Staliūnaitė, Julius Cheng, and Andreas Vlachos. 2025. Uncertainty quantification for evaluating machine translation bias. *Preprint*, arXiv:2507.18338.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470.

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.

A Examples for 6-shot Setup

The following is a list of six examples, two for each label, and their corresponding explanations:

"Statement: Everything can be found inside a shopping mall." "Context: Enter the realm of shopping malls, where everything you’re looking for is available without moving your car." "Judgment: Entailment" "Explanation: The context implies that the shopping mall has everything one might look for, as it can be found without moving your car."

"Statement: The matter of whether or not the Mass is a sacrifice for the remission of sins is controversial." "Context: As for the divisive issue of whether the Mass is a sacrifice for the remission of sins, the statement affirms that Christ’s death upon the cross ..." "Judgment: Entailment" "Explanation: The context states that the Mass being a sacrifice for the remission of sins is divisive, which can be interpreted as a synonym for controversial."

"Statement: Most rock concerts take place in the Sultan’s Pool amphitheatre." "Context: In the summer, the Sultan’s Pool, a vast outdoor amphitheatre, stages rock concerts or other big-name events." "Judgment: Neutral" "Explanation: The context does not specify whether it is most or only some rock concerts that are staged in the Sultan’s Pool."

"Statement: This information was developed thanks to extra federal funding." "Context: Additional information is provided to help managers incorporate the standards into their daily operations." "Judgment: Neutral" "Explanation: The context does not indicate where the information came from, which may or may not be federal funding."

"Statement: He had recently seen pictures depicting those things." "Context: He hadn’t seen even pictures of such things since the few silent movies run in some of the little art theaters." "Judgment: Contradiction" "Explanation: If the pronoun ’he’ and the object ’those things’ refer to the same things in the statement and the context, then the statement negates the context. "

"Statement: Octavius Decatur Gass refers to four people. " "Context: One opportunist who stayed was Octavius Decatur Gass. " "Judgment: Contradiction" "Explanation: The context names one person as Octavius Decatur Gass, and does not mention additional referrents. "

B Score Combinations

Table 2 presents the most common combinations of binary soft labels.

Combination	Count
((1.0,0.0),(1.0,0.0),(0.0,1.0))	80
((1.0,0.0),(0.75,0.25),(0.25,0.75))	40
((1.0,0.0),(0.0,1.0),(1.0,0.0))	36
((0.75,0.25),(1.0,0.0),(0.25,0.75))	33
((1.0,0.0),(0.5,0.5),(0.5,0.5))	32
((0.0,1.0),(1.0,0.0),(1.0,0.0))	26
((0.5,0.5),(1.0,0.0),(0.5,0.5))	24
((1.0,0.0),(0.25,0.75),(0.75,0.25))	17
((0.25,0.75),(1.0,0.0),(0.75,0.25))	15
((1.0,0.0),(0.33,0.67),(0.67,0.33))	12
((0.75,0.25),(0.75,0.25),(0.5,0.5))	10
((0.67,0.33),(1.0,0.0),(0.33,0.67))	7
((0.75,0.25),(0.25,0.75),(1.0,0.0))	7
((1.0,0.0),(0.67,0.33),(0.33,0.67))	6
((0.75,0.25),(0.5,0.5),(0.75,0.25))	6
((0.5,0.5),(0.75,0.25),(0.75,0.25))	6
((0.25,0.75),(0.75,0.25),(1.0,0.0))	5
((0.5,0.5),(0.5,0.5),(1.0,0.0))	5
((0.33,0.67),(1.0,0.0),(0.67,0.33))	3
((0.33,0.67),(0.67,0.33),(1.0,0.0))	2

Table 2: Frequency of label distribution combinations

C Best Configuration

The best performing variant of our system had the following configuration: a multilabel classification task with seven classes for each label, cross-label loss, embedding dimension of 16 for the entropy module, using fusion MLP to combine the text and entropy features in a layer of size 256, dropout of 0.1, learning rate of 1e-5, weight decay of 1e-6, step LR scheduler, all embedding layers frozen, no regularisation against mean penalty, entropy penalty ($\beta = 0.05$), no temperature annealing, no sum < 1 penalty.

Author Index