LREC 2026

NLPerspectives @ LREC 2026

Workshop Proceedings

Editors
Gavin Abercrombie, Valerio Basile, Shiran Dudy,
Simona Frenda, Elisa Leonardelli, and Davide Bernardi

12 May 2026

ISBN 978-2-493814-69-2

Preface

A recent research space in Natural Language Processing (NLP) has emerged to tackle the challenge of what has been called a perspectivist turn. Researchers have increasingly moved away from traditional data curation protocols such as majority vote aggregation in favor of considering multiple perspectives as legitimate ground truths. Perspectivist models leverage human label variation to better account for user diversity and adopt evaluation strategies capable of embracing disagreement.

Perspectivist NLP is related to personalisation, i.e., a family of methods that aim at adapting NLP models towards individual users. Personalisation has an impact on NLP applications where the user is specific and known, such as virtual assistants. However, applications intended for the broad public, such as automatic moderation and large-scale language analysis, do not benefit from this level of specificity, either because data about all possible users is not available, or because the intended output abstracts away from single users and towards more encompassing audiences.

The most common domain where perspectivist approaches are found to be relevant is the analysis of pragmatic phenomena in language, especially hate speech or other kinds of undesirable language. However, even tasks less commonly associated with differences in human perception, such as grammatical and multi-modal tasks, are sensitive to systematic label variation.

There are still many open challenges in perspectivist approaches, but research is proliferating, as demonstrated, for example, by the ongoing success of this workshop as well as other ones as Context and Meaning: Navigating Disagreements in NLP Annotation and Cross Cultural Considerations in NLP (C3NLP).

This volume presents the proceedings of the Fifth Workshop on Perspectivist Approaches to NLP (NLPerspectives 2026), co-located with LREC 2026 in Palma de Mallorca. The workshop had a total of 22 submissions, with 13 archival papers presented in this volume and 6 non-archival papers. The workshop continues to serve as a forum for researchers exploring how to model, preserve, and learn from the diversity of human perspectives in language annotation and natural language processing systems. This year, we are honored to feature an invited talk by Federico Cabitza (University of Milano-Bicocca), author of fundamental work in perspectivist machine learning, whose contributions have been instrumental in shaping the theoretical foundations of this field.

Workshop Committees

Organizing Committee

Gavin Abercrombie, Heriot Watt University
Valerio Basile, University of Turin
Davide Bernardi Amazon Research
Shiran Dudy, Northeastern University
Elisa Leonardelli Fondazione Bruno Kessler
Simona Frenda, Heriot-Watt University

Program Committee

Amanda Cercas Curry, CENTAI Institute
Samuele D'Avenia, University of Turin
Eliana Di Palma, University of Turin
Esra Dönmez, University of Stuttgart
Teddy Ferdinan, Wrocław University of Science and Technology
Aiqi Jiang, Heriot-Watt University
Anna Koufakou, Florida Gulf Coast University
Sofie Labat, Ghent University
Soda Marem Lo, University of Turin
Sergio Lopez Sancio, Amazon
Marta Marchiori Manerba, University of Turin
Michele Mastromattei, University of Rome Tor Vergata
Julia Romberg, GESIS – Leibniz Institute for the Social Sciences
Pratik Sachdeva, University of California, Berkeley
Erhan Sezerer, Amazon Research
Tiago Timponi Torrent, Federal University of Juiz de Fora
Giovanni Valer, Fondazione Bruno Kessler
Tharindu Cyril Weerasooriya, Accenture

Table of Contents

Conference Program

Tuesday, May 12, 2026

09:00–9:10 **Opening Session**
Room: Room 4

09:10–10:10 **Keynote Session by Dr. Federico Cabitza**
Room: Room 4
Chair: Valerio Basile

10:10–10:30 **One-Minute Lightning Talks**
Room: Room 4
Chair: Gavin Abercrombie

10:30–11:00 **Coffee Break**

11:00–12:30 **Poster Session**
Room: Poster Area

11:00–12:30 *What is Truth in NLP? Reflecting on Progress, Lessons, and Open Challenges as NLPerspectives turns Five*
Gavin Abercrombie

11:00–12:30 *The Rashomon Wikipedia: A Data-Perspectivist Analysis of Divergent Historical Narratives*
Claudiu Creanga, Liviu P. Dinu and Anca Dinu

11:00–12:30 *GSI:detect - A Perspectivist Approach to Gender Stereotypes Identification in Italian*
Davide Testa, Sofia Brenna, Manuela Speranza, Gloria Comandini, Stefania Cavagnoli and Bernardo Magnini

11:00–12:30 *Quantifying and Predicting Disagreement in Graded Human Ratings*
Leixin Zhang and Çağrı Çöltekin

11:00–12:30 *HurtLens: A Perspectivist Corpus Analysis of Hurtful Language*
Samuele D'Avenia, Eliana Di Palma, Marta Marchiori Manerba and Valerio Basile

11:00–12:30 *A Measure of Systematic Disagreement*
Valerio Basile

11:00–12:30 *Fine-Grained Perspectives: Modeling Explanations with Annotator-Specific Rationales*
Olufunke O. Sarumi, Charles Welch and Daniel Braun

Tuesday, May 12, 2026 (continued)

11:00–12:30 *Structured Disagreement in Health-Literacy Annotation: Epistemic Stability, Conceptual Difficulty, and Agreement-Stratified Inference*
Olga Kellert, Sriya Kondury, Candice Koo, Nemika Tyagi and Steffen Eikenberry

11:00–12:30 *SubData: Bridging Heterogeneous Datasets to Enable Theory-Driven Evaluation of Political and Demographic Perspectives in LLMs*
Pietro Bernardelle, Leon Froehling, Stefano Civelli and Gianluca Demartini

11:00–12:30 *ChatGPT, why can't anyone afford a house? On the Effects of LLM pre-annotation on Annotator Subjectivity*
Emilie Francis, Céline Leuzinger, Ricardo Muñoz Sánchez and Lee D. Gauthier

11:00–12:30 *An Overview of Current Practices and Recommendations for Working with Stereotypes in NLP*
Alessandra Teresa Cignarella and Matteo Pellegrini

11:00–12:30 *Modeling Perspectives in NLP: Parameter-Efficient Perspective Conditioning for Span Extraction and Summarization*
Harikrishnan Gurushankar Saisudha and Sabine Bergler

11:00–12:30 *A Pilot Study Investigating Stakeholder Subjectivity in Collaborative Dialog Analysis*
Ananya Ganesh, Martha Palmer and Katharina von der Wense

13:00–14:00 Lunch Break

14:00–15:30 Participatory Session
Room: Room 4
Chairs: Valerio Basile, Gavin Abercrombie

Tuesday, May 12, 2026 (continued)

16:00–16:30 **Coffee Break**

16:30–17:30 **Panel Discussion**
Room: Room 4
Chair: Gavin Abercrombie

17:30–17:45 **Closing Session**
Room: Room 4
Chairs: Gavin Abercrombie, Valerio Basile

What is Truth in NLP? Reflecting on Progress, Lessons, and Open Challenges as NLPerspectives turns Five

Gavin Abercrombie
The Interaction Lab
School of Mathematical & Computer Sciences
Heriot-Watt University, Edinburgh, Scotland
g.abercrombie@hw.ac.uk.org

Abstract

This paper reflects on five years of the Workshop on Perspectivist Approaches to NLP (NLPerspectives) and examines how this research community has helped to reconceptualise the notion of ground truth in human-labelled data and its classification. As NLP research has increasingly engaged with social and affective tasks, traditional assumptions about annotation reliability–centred on inter-annotator agreement and single 'gold standard' labels–have proven insufficient for capturing the genuine diversity of human perspectives. I review the developments that have driven the 'Perspectivist Turn,' assess its influence on mainstream NLP practice, and highlight the methodological challenges that arise when modelling disagreement, subjectivity, and annotator variation. In particular, I consider unresolved questions around evaluation paradigms, task formulation, population representation, community norms, and the implications of using pre-trained generative models as classifiers. By synthesising discussions from five years of workshops, keynotes, and related publications, I outline open challenges and propose directions for future work aimed at more rigorous perspectivist NLP. I argue that we should focus on more realistic task formulations and the centering minoritised standpoints, and caution against viewing potentially harmful interpretations as equally legitimate reactions to 'subjective' phenomena.

Keywords: Perspectivism, Variation, Disagreement

1. Introduction

'*What is truth*?' asked Pontius Pilate (John 18:38), but famously, he did not provide or wait for an answer. In recent years a growing section of the Natural language processing (NLP) research community has been asking a similar question. In this paper marking the fifth edition of the Workshop on Perspectivist Approaches to NLP (NLPerspectives),[1] I examine what we mean by *(ground) truth* when it has become (relatively) more common to collect, model, and attempt to represent multiple diverging responses in human label data. Here, I reflect on both progress made, and attempt to take a snapshot of the field, including open questions, challenges, and limitations of current approaches.

2. Background

As a field that grew out of the related discipline of computational linguistics (CL), from the early 2000s, NLP saw a shift away from more well defined linguistic tasks such as part-of-speech tagging and dependency parsing, which had been the focus of research up to and including the turn of the Millenium, to begin considering social and affective tasks like sentiment analysis (e.g. Pang et al., 2002; Wiebe et al., 2005) and emotion recognition (e.g. Mohammad and Turney, 2013).

Throughout this period, methodologies and standards were developed for establishing the reliability of human labels applied to text data. These were believed to provide an indication of the *reliability* of these annotations as markers of the *ground truth* categories to which individual data points belonged.

Chief among these was the measurement of chance-adjusted inter-annotator agreement (IAA). Borrowed from Behavioral Science, this was introduced to CL by Carletta (1996), who pointed out that the quality of previous work had been purely 'judged according to whether or not the reader found the explanation plausible.' In comparison, the application of a rigorous statistical measurement was a common sense way to improve the robustness of data collection methods. An iterative NLP corpus creation methodology was established in which: (1) guidelines were drawn up explaining the phenomena of interest, (2) data was annotated following these instructions, (3) IAA was measured, (4) reasons for disagreement were interrogated, attempts made to iron them out, and guidelines adjusted to avoid such deviance in the future. We then returned to step (1), and repeated until the Cohen's *kappa*, Krippendorf's *alpha*, or another chosen statistic was deemed to be satisfactory, usually by meeting some quite arbitrary threshold (Warrens, 2015). Manuals were written with step-by-step explanations of how to follow this procedure (Artstein, 2017; Pustejovsky and Stubbs, 2012).

One effect of this standardisation was that, as

[1] https://nlperspectives.di.unito.it/

Proceedings of NLPerspectives @ LREC 2026, pages 1–10
12 May 2026.

high IAA scores became synonymous with reliability, and therefore the quality of data collection, it seemed to become almost impossible to publish work that showed even moderate levels of variation in annotator behaviour (as many researchers struggling to achieve high IAA at this time would probably attest). Where disagreement was found, this became reason for devising methods to weed out supposedly unreliable annotators (Hovy et al., 2013) and 'noisy' labels (e.g., aggregation by majority vote), or, as Aroyo and Welty (2013) pointed out, to try to force consensus through over-specified and overly ungeneralisable development.

The 'Perspectivist Turn' There are, in fact, a number of examples of CL and NLP work acknowledging the potential for finding 'signal' in the 'noise' of human label variation going back to the 2010s, and beyond. Aroyo and Welty (2013) proposed to collect the '*Crowd Truth*' distribution of annotator responses believing (dis)agreement levels provided information about the relative 'clarity', 'vagueness', or 'ambiguity' of a labelled item. Further examples are Jurgens (2014), who considered annotator disagreement as a proxy for item difficulty in a word sense labelling task, and Plank et al. (2014), who showed that such disagreements were systematic for part-of-speech labelling. Arguing that the notion of *acceptability* should replace that of 'ground truth', Alm (2011) pointed to work on subjectivity in CL and linguistics going as far back as the 1930s.

However, mainstream NLP methodology continued to prioritise the collection and modelling of single 'gold standard' class labels until the field saw the beginnings of a 'Perspectivist turn' in the early 2020s. Signs of this included talk of the 'end of the gold standard' (Basile, 2020) and the 'need to talk about disagreement' (Basile et al., 2021), the launch of a Perspectivist Data Manifesto urging researchers to follow disaggregated data practices,[2] a growing interest in 'learning with disagreement' (Uma et al., 2021b), including the launch of the Le-Wi-Di shared task (Uma et al., 2021a; Leonardelli et al., 2023, 2025), publication of a prominent survey of relevant resources (Plank, 2022), and an emerging interest in modelling of individual annotators (Cercas Curry et al., 2021; Davani et al., 2022; Vitsakis et al., 2023).

NLPerspectives at Five Which brings us to the launch of the present workshop series. Conceived of in 2021, and with its first edition taking place at LREC in May 2022, NLPerspectives is now celebrating its 5th edition. During this time, it has seen the publication of 68 research papers (up to edition 4, see Figure 1), the presentations of several other non-archival works and research communications, and hosted five keynote talks on topics relevant to and overlapping with perspectivist data practices: Su Lin Blodgett on participatory design (2022 LREC, Marseille), Przemysław Kazienko on personalised NLP (2023 ECAI, Krakow), Barbara Plank on human label variation and model uncertainty (2024 LREC-COLING, Turin), Jose Camacho Collados on cultural factors in multilingual models (2025 EMNLP, Suzhou), and now at LREC 2026 in Palma de Mallorca, Federico Cabitza, author of 'the Perspectivist Turn'.[3]

Each workshop has ended with a panel discussion featuring the organisers, invited speakers, and other researchers from the community reflecting on progress, challenges, the state of the field. In this paper, I attempt to synthesise some of the talking points from these conversations.

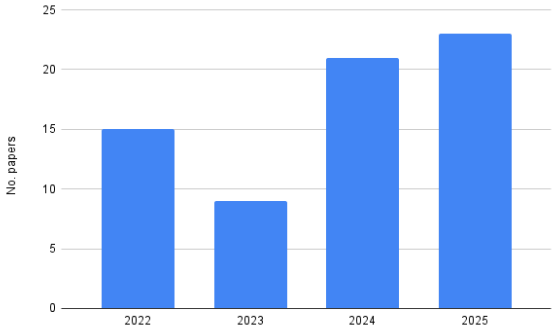


Figure 1: Published papers over time at NLPerspectives. With the exception of 2023, when the workshop was presented outwith the *ACL community (at ECAI), the workshop has seen steady growth in submissions and accepted papers.

Influence on the wider NLP field The idea of collecting, preserving, and modelling multiple labels appears to have become considerably more mainstream than it was five years ago. In a reversal of the situation described previously, anecdotally at least, peer reviewers sometimes now consider a lack of consideration for legitimate annotator disagreement to be a methodological weakness.

There have also been a number of developments that indicate that the field may have been influenced by the workshop and the research of those working in this community. Other workshops and events have sprung up with similar themes, such as Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation (Roth and Schlechtweg, 2025) and the choice of 'subjectivity and disagreement in abusive language data' as special theme for the 7th Workshop on Online Ab-

[2] https://pdai.info/

[3] Information about the workshop is archived at https://nlperspectives.di.unito.it/.

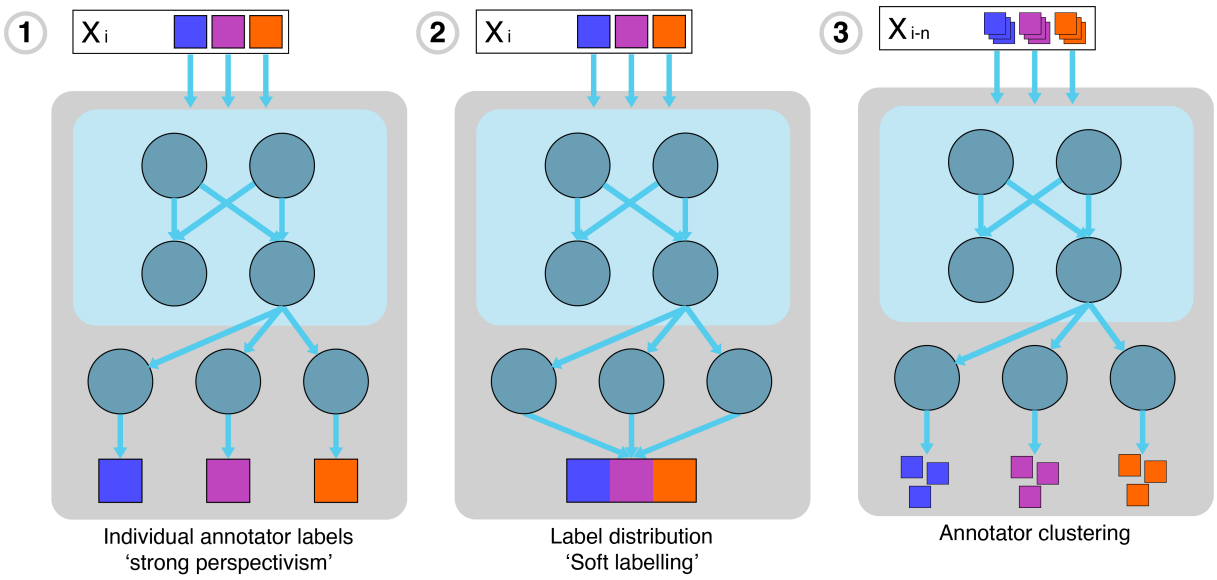


Figure 2: Three perspectivist paradigms. For a given data point X_i, and set of labels (here represented by differently coloured squares), models (1) predict individual annotator labels; 92) output a prediction of the distribution of labels; or (3) focus on grouping annotators according to their labelling behaviour. Figure after Davani et al. (2022).

use & Harms (WOAH) (Chung et al., 2023). Most notably, the ACL 2025 keynote speech titled 'Whose Gold?' was given by founding member of the workshop's organising committee, Verena Rieser, and expounded on perspectivist themes (Che et al., 2025).

3. Open questions and challenges

Although the developments described above indicate a shift in the field towards more mainstream acceptance of the tenets of perspectivist NLP, there remain many challenges. In many ways, the embrace of perspectivism opens more questions than it resolves. While (Frenda et al., 2025) reviewed work conducted under the perspectivist banner, in the rest of this paper, I consider these questions, and make suggestions for approaches.

3.1. Evaluation

From the beginnings of the workshop, perhaps the most frequently posed and least satisfactorily answered question has been 'how (or what) should we evaluate?' Here, three main paradigms have emerged (see Figure 2):

1. The individual annotators responsible for providing the labels are modelled, and a label is output for each that represents a prediction of what they *would* label the item in question (e.g. Cercas Curry et al., 2021; Davani et al., 2022; Lo et al., 2025; Orlikowski et al., 2023). This represents a form of descriptive relativism, in which all individual truths are seen as equally valid.

2. Models are designed to predict the distribution of labels, i.e., in the example illustrated in Figure 2, the soft label is $[0.33, 0.33, 0.33]$. This has been used as the main evaluation metric for the Le-Wi-Di shared task, where it is referred to as 'soft labelling' (Leonardelli et al., 2023). Other examples of this approach include Madeddu et al. (2023), Parappan and Henao (2025), Weerasooriya et al. (2023a), and Weerasooriya et al. (2023b).

3. Some work is concerned primarily with modelling the relationship between the labelling behaviours of different people, often with the aim of finding like-minded groups within the cohort of annotators (Lo and Basile, 2023; Liu et al., 2019; Prabhakaran et al., 2024; Vitsakis et al., 2024) or with finding individuals with particular attitides or beliefs (Chulvi et al., 2023; Jiang et al., 2024).

Each of these formulations has some drawbacks and theoretical weaknesses. For (1), beyond the technical issues of it being difficult to scale up and of problems arising when annotators are represented sparsely in datasets (Leonardelli et al., 2023), it is a little difficult to see many use cases for this approach. One struggles, for instance, to see how this paradigm might fit into any scenario that requires a decision to be taken, such as a content moderation pipeline.

We might make a similar criticism of paradigm (2), which may need some kind of threshold to be established for decision-making applications—in which case we are not far from the majority vote paradigm that perspectivism seeks to avoid. In fact, some work uses the distribution of soft labels primarily as a signal to improve prediction of aggregated labels (e.g., Uma et al., 2020). However, it is easier to see how this output could be informative for downstream tasks, particularly if we have some information about *who* is represented in the distribution (see subsection 3.3).

Paradigm (3) encompasses work that uses demographic information (e.g. Gordon et al., 2022; Prabhakaran et al., 2024), annotator behaviour (e.g. Lo and Basile, 2023; Vitsakis et al., 2024), or underlying beliefs and attitudes of annotators (e.g. Chulvi et al., 2023; Jiang et al., 2024). A weakness here, when compared to similar work in other fields, is that NLP practitioners tend not to consider the populations that they model, limiting both the rigour of the research and the utility of the models (see subsection 3.3).

3.2. Unclear Task Formulation

As discussed in subsection 3.1, there is something of a lack of clarity in the relationship between much of the research published in this area and the real-world tasks that motivate it. For all the faults with traditional 'gold label' modelling that perspectivist NLP has attempted to address, it is far easier to see how systems that output single label predictions might be applied in practice. Perspectivist researchers have so far failed to demonstrate how their approaches—which currently seem more suited to exploratory data analysis—might fit into such decision-making or predictive systems.

One way to do this might be to consider collecting evidence from extrinsic evaluation practices, which have so far been lacking in the field (Reiter, 2025).

3.3. Increasing but Non-rigorous Complexity

With the rise of perspectivism, researchers have been gradually de-simplifying supervised classification tasks, by adding further levels of complexity, realism, and information that needs to be accounted for.

Accepting that some annotation items are subjective or ambiguous, or that readings of them can legitimately differ, researchers became interested in *who* holds *which* perspectives, seeking to harness demographic and other information about them. However, unlike social scientists, we have all but ignored the concept of representing a target population,[4] leaving the research open to accusations of lack of rigour. Extreme examples include contending that three individual annotators can each represent *conservative*, *moderate*, and *liberal* points of view (Almanea and Poesio, 2022).

Once it was accepted that there might be valid variation in annotator responses, we began to look at the causes of these differences, seeking to tease apart fixed opinions, ambiguous and difficult data items, and noisy and erroneous annotation work. For example, in a series of longitudinal experiments, Abercrombie et al. (2023a, 2025) found that annotators are internally inconsistent on repeated annotation items around 75% of the time, which they put down largely to ambiguity in the data. In a field that collects labels primarily from anonymous crowdworkers working in completely uncontrolled environments, there are a number of factors that might make one doubt that these represent any kind of truth, even a subjective one.

As we acknowledge an increasing number of factors that cause label diversity, we should be careful to apply rigour in modelling them.

3.4. The Problem of Noise

A side effect of accepting that label variation can be valid for a wide variety of reasons is that we have lost the tools that we previously believed provided evidence of the reliability of our data and collection practices. In addition to IAA scores becoming variation analysis tools rather quality indicators, it has become very difficult to apply previously common methods such as attention check items. After all, if many reactions are valid, how can we say that annotators should label a particular item a certain way?

In situations where we are *particularly* interested in minoritised perspectives, such as hate speech detection, in which only those with lived experience may be capable of recognising the phenomenon of interest, a valid approach is to seek those people as annotators (Abercrombie et al., 2023b; Fleisig et al., 2024). While one method is to establish a pool of tried and tested annotators (e.g. Jiang et al., 2024), there is a danger of a lack of rigour and that this may be done mainly on 'vibes'. With the recognition that the 'crisis of reproducibility' is firmly embedded in NLP data practices (Belz et al., 2023; Dinkar et al., 2024), and observed specifically in human labelling for supervised classification (Sasidharan Nair et al., 2024), we need to establish new methods for validating the quality of the labels we collect.[5]

[4] Exceptions include Pei and Jurgens (2023), who draw representative population samples, and Eckman et al. (2025), who weight annotations by populations.

[5] See Fleisig et al. (2025) for an exploration of heuristics for this purpose.

3.5. Subjectivity vs Community Norms

Undoubtedly the most researched topic in perspectivist NLP has been that of hate speech and other toxic language. While this is probably due to its inherently contested nature, it has led to often repeated claims that hate speech is a subjective phenomenon (e.g., Akhtar et al., 2021; Almanea and Poesio, 2022; Basile, 2020). This is not only untrue from a philosophical (as well as legal) point of view (Barendt, 2019), but creates the danger of *bothsidesing* the points of view of perpetrators and targets of hate speech. This may be particularly likely in the individual annotator modelling paradigm (subsection 3.1), in which each annotator may be given precisely equal weight (at least in the absence of a well designed task formulation). In fact, hate speech should be defined at the community level, preferably according to the norms of those it impacts (Cercas Curry et al., 2024).

Researchers should be careful to define 'subjectivity' precisely in relation to other phenomena that influence label variation. When working on hate speech phenomena (e.g., racism, sexism), we should consider how to model the community norms of those affected.

3.6. Taking a stand(point)

To do this, we may need to accept that objectivity in research design is neither possible or desirable. As, following *standpoint theory*, only people with relevant lived experience are capable of recognising the phenomena of interest, NLP researchers may need to actively seek to foreground those voices. As one of the theory's principal proponents, Sandra Harding, put it 'a standpoint is not the same as a viewpoint or a perspective, for it requires both science and a political struggle' (Harding, 1998, p.150).

One approach to this is through participatory/co-design with stakeholders. This can take many forms, including focus groups, workshops, and Delphi studies (Wilson et al., 2025), but should aim to involve participants beyond simple consultation and validation of technical goals, and ideally hand over a level of ownership of the research agenda (Caselli et al., 2021; Delgado et al., 2023).

While it may not be easy to fully achieve these aims (due to e.g. conflicting motivations and funding issues (Wilson et al., 2025)), another strand of research investigates how to scale such work up to select specific crowdworkers that share the values of these communities. This is important, as there is growing evidence that demographic information is not a reliable predictor of annotation behaviour (see e.g. Orlikowski et al., 2023, 2025). One approach is to collect information about annotators' underlying attitudes using validated surveys (Jiang et al., 2024).

3.7. Generative Models as Classifiers

As Star and Bowker (1999, p.1) contend, 'to classify is human'. But is it also LLM? The vast majority of work in this area has focused on the type of tasks traditionally solved by discriminative machine learning models. However, since 2022–coincidentally both the year of the first workshop and the release of ChatGPT–we are increasingly undertaking classification tasks with pre-trained generative language models (see e.g. Balestrucci et al., 2025; Plaza-del Arco et al., 2024; Pavlovic and Poesio, 2024), often referred to as 'LLM-as-a-judge'.

While discriminative models were designed to output discrete labels, the latter, if left to their own devices, will emit long screeds of text, expounding on the topic in a vaguely knowledgeable style and marked by formulaic rhetorical devices and bullet points. In many cases this output has been guided through reinforcement learning to appear as helpful and engaging as possible, and models are known to engage in seemingly obsequious behaviours, changing their responses in order to affirm the viewpoints of users (Ranaldi and Pucci, 2025). In short, 'truth' may be a somewhat secondary concern for such models, optimised to satisfy users, and referred to by Hicks et al. (2024) as 'bullshitters' .

As text generation models are now pervasive, perspectivist research needs to take into account these behaviours and consider what it means to use such models for classification tasks, when, in the real-world, users must contend with non-determinism, affirmation bias (Sharma et al., 2023), refusal behaviour (Ouyang et al., 2022), and generation of false information.

At the same time, we should broaden our focus from classification tasks to the human preference ranking data that to a large extent underlies the success of these models. Despite the many sources of disagreement in such data (Dsouza and Kovatchev, 2025), and indications that inclusion of disagreements can lead to better performance (Gooding and Mansoor, 2023), there is currently little evidence that minoritised perspectives are actually maintained in RLHF training data for generative models.

4. Related Work

In addition to the historical work discussed in section 2, two recent surveys take critical looks at perspectivist NLP research. Frenda et al. (2025) provide a systematic review of work published at the workshop and beyond. They focus primarily on thematic analysis of publications up until the writing of the review (2024), and particularly highlight the lack of a clear direction on how to evaluate perspectivist modelling.

Fleisig et al. (2024) take a broader look at the shift to perspectivist methods, highlighting several

issues that overlap with the points we make here. They particularly highlight the need to make normative decisions highlighting annotators with expert knowledge of the phenomenon of interest, including relevant lived experience, and suggest particpatory practices as a means of doing so, as I advocate in subsection 3.5.

Combining the approaches of these two works, in this position paper, I have tried to provide a snapshot of the field on the occasion of the 5th NLPerspectives workshop, and set out a personal view of what is currently lacking in this area.

5. Conclusion

As NLPerspectives marks its fifth edition, we are able to reflect on this research area's growth from a niche and fringe community challenging the orthodoxy of 'ground truth' in NLP to a situation approaching mainstream adoption in the field. At the same time, a number of difficult questions and unsolved challenges have come under discussion at the workshop. In this paper, I have attempted to set out some of these issues, and argued that we need to be more rigorous and deliberate in defining the truths that we seek to model.

Acknowledgements

Thanks to the members of the NLPerspectives Programme Committee for their helpful and insightful comments and suggestions, which I have tried to incorporate in this version of this paper.

Thanks also to the research group at IMS Stuttgart, who invited me to give a talk in March 2026, on which this paper was based.

This work was supported by the EPSRC project 'Equally Safe Online' (EP/W025493/1).

References

Gavin Abercrombie, Tanvi Dinkar, Amanda Cercas Curry, Verena Rieser, and Dirk Hovy. 2025. Consistency is key: Disentangling label variation in natural language processing with intra-annotator agreement. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 63–74, Suzhou, China. Association for Computational Linguistics.

Gavin Abercrombie, Dirk Hovy, and Vinodkumar Prabhakaran. 2023a. 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.

Gavin Abercrombie, Aiqi Jiang, Poppy Gerrard-Abbott, Ioannis Konstas, and Verena Rieser. 2023b. Resources for automated identification of online gender-based violence: A systematic review. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 170–186, Toronto, Canada. Association for Computational Linguistics.

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*.

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.

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. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *WebSci '13*.

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

Pier Felice Balestrucci, Michael Oliverio, Elisa Chierchiello, Eliana Di Palma, Luca Anselma, Valerio Basile, Cristina Bosco, Alessandro Mazzei, and Viviana Patti. 2025. Towards a perspectivist understanding of irony through rhetorical figures. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 27–36, Suzhou, China. Association for Computational Linguistics.

Eric Barendt. 2019. What is the harm of hate speech? *Ethical Theory and Moral Practice*.

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.

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.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Krahmer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood, Margot Mieskes, Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. 2021. Guiding principles for participatory design-inspired natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35, Online. 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.

Amanda Cercas Curry, Gavin Abercrombie, and Zeerak Talat. 2024. Subjective isms? on the danger of conflating hate and offence in abusive language detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 275–282, Mexico City, Mexico. Association for Computational Linguistics.

Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors. 2025. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria.

Berta Chulvi, Lara Fontanella, Roberto Labadie-Tamayo, Paolo Rosso, et al. 2023. Social or individual disagreement? perspectivism in the annotation of sexist jokes. In *2nd Workshop on Perspectivist Approaches to NLP (NLPerspectives)*, volume 3494. CEUR Workshop Proceedings.

Yi-Ling Chung, Paul Röttger, Debora Nozza, Zeerak Talat, and Aida Mostafazadeh Davani, editors. 2023. *Proceedings of the 7th Workshop on Online Abuse & Harms (WOAH)*. Association for Computational Linguistics, Toronto, Canada.

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.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA. Association for Computing Machinery.

Tanvi Dinkar, Gavin Abercrombie, and Verena Rieser. 2024. ReproHum #0927-03: DExpert evaluation? reproducing human judgements of the fluency of generated text. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 145–152, Torino, Italia. ELRA and ICCL.

Russel Dsouza and Venelin Kovatchev. 2025. Sources of disagreement in data for LLM instruction tuning. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 20–32, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Stephanie Eckman, Bolei Ma, Christoph Kern, Rob Chew, Barbara Plank, and Frauke Kreuter. 2025. Aligning NLP models with target population perspectives using PAIR: Population-aligned instance replication. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 100–110, Suzhou, China. 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.

Eve Fleisig, Matthias Orlikowski, Philipp Cimiano, and Dan Klein. 2025. Balancing quality and variation: Spam filtering distorts data label distributions. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 47–62, Suzhou, China. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, , Raffaella Panizzon, Alessandra Teresa Cignarella, and Davide Marco, Cristina Bernardi. 2025. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*.

Sian Gooding and Hassan Mansoor. 2023. The impact of preference agreement in reinforcement learning from human feedback: A case study in summarization.

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 *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Sandra Harding. 1998. *Is science multicultural?: Postcolonialisms, feminisms, and epistemologies*. Indiana University Press.

Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. Chatgpt is bullshit. *Ethics and Information Technology*.

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.

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.

David Jurgens. 2014. An analysis of ambiguity in word sense annotations. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3006–3012, Reykjavik, Iceland. European Language Resources Association (ELRA).

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 The 4th Workshop on Perspectivist Approaches to NLP*, pages 182–195, Suzhou, China. Association for Computational Linguistics.

Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher Homan. 2019. 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.

Soda Marem Lo and Valerio Basile. 2023. Hierarchical clustering of label-based annotator representations for mining perspectives.

Soda Marem Lo, Silvia Casola, Erhan Sezerer, Valerio Basile, Franco Sansonetti, Antonio Uva, and Davide Bernardi. 2025. PERSEVAL: A framework for perspectivist classification evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22334–22359, Suzhou, China. Association for Computational Linguistics.

Marco Madeddu, Simona Frenda, Mirko Lai, Viviana Patti, and Valerio Basile. 2023. DisaggregHate it corpus: A disaggregated Italian dataset of hate speech. In *Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 243–250, Venice, Italy. CEUR Workshop Proceedings.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

Mohammed Fayiz Parappan and Ricardo Henao. 2025. Learning subjective label distributions via sociocultural descriptors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20322–20338, Suzhou, China. Association for Computational Linguistics.

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.

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.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. 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.

Flor Miriam Plaza-del Arco, Debora Nozza, and Dirk Hovy. 2024. Wisdom of instruction-tuned language model crowds. exploring model label variation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 19–30, Torino, Italia. ELRA and ICCL.

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.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O' Reilly.

Leonardo Ranaldi and Giulia Pucci. 2025. When large language models contradict humans? large language models' sycophantic behaviour.

Ehud Reiter. 2025. We should evaluate real-world impact. *Computational Linguistics*, 51(4):1419–1431.

Michael Roth and Dominik Schlechtweg, editors. 2025. *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*. International Committee on Computational Linguistics, Abu Dhabi, UAE.

Sachin Sasidharan Nair, Tanvi Dinkar, and Gavin Abercrombie. 2024. Exploring reproducibility of human-labelled data for code-mixed sentiment analysis. In *Proceedings of the Fourth Workshop*

on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024, pages 114–124, Torino, Italia. ELRA and ICCL.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. In *Proceedings of the International Conference on Learning Representations*.

Susan Leigh Star and Geoffrey Bowker. 1999. Sorting things out. *Classification an its consequences The MIT Press, Cambridge, Massachusetts, London, England*.

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 AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 173–177.

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.

Nikolas Vitsakis, Amit Parekh, Tanvi Dinkar, Gavin Abercrombie, Ioannis Konstas, and Verena Rieser. 2023. iLab at SemEval-2023 task 11 le-wi-di: Modelling disagreement or modelling perspectives? In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1660–1669, Toronto, Canada. Association for Computational Linguistics.

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.

Matthijs J Warrens. 2015. Five ways to look at Cohen's kappa. *Journal of Psychology & Psychotherapy*, 5.

Tharindu Cyril Weerasooriya, Sarah Luger, Saloni Poddar, Ashiqur KhudaBukhsh, and Christopher Homan. 2023a. Subjective crowd disagreements for subjective data: Uncovering meaningful CrowdOpinion with population-level learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 950–966, Toronto, Canada. 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.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*.

Marianne Wilson, David M. Howcroft, Ioannis Konstas, Dimitra Gkatzia, and Gavin Abercrombie. 2025. Participatory design for positive impact: Behind the scenes of three NLP projects. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 252–263, Vienna, Austria. Association for Computational Linguistics.

The Rashomon Wikipedia: A Data-Perspectivist Analysis of Divergent Historical Narratives

Claudiu Creanga[2,3], **Liviu P. Dinu**[1,3], **Anca Dinu**[3,4]
[1] Faculty of Mathematics and Computer Science,
[2] Interdisciplinary School of Doctoral Studies,
[3] HLT Research Center,
[4] Faculty of Foreign Languages and Literatures,
University of Bucharest, Romania
claudiu.creanga@fmi.unibuc.ro, ldinu@fmi.unibuc.ro, anca.dinu@lls.unibuc.ro

Abstract

Wikipedia aims to provide a unified, neutral record of history, yet its independent language editions often function as distinct epistemic communities, creating divergent narratives around contested events. This paper investigates cross-lingual historiographical bias by analyzing Wikipedia articles across five languages (Romanian, Hungarian, Russian, Turkish, and English) focusing on three contentious events in Romanian history: the Battle of Posada (1330), the Soviet occupation of Bessarabia (1940), and the Night Attack at Târgoviște (1462). Using human annotators and Large Language Models (LLMs) to classify citation stance and quantify narrative evolution from 2005 to 2024, we identify a phenomenon of "citation isolation". In the case of the Battle of Posada, only 2 out of 119 citations were shared between language editions, with the Romanian edition exhibiting a 91% pro-national bias compared to the balanced Hungarian edition. Longitudinal analysis reveals that these narratives are volatile and responsive to contemporary geopolitics, evidenced by a significant shift in the Russian framing of Bessarabia in 2024. Finally, we propose a "Peace-Maker" pipeline to automate conflict reconciliation. We demonstrate that while standard prompting leads models to hallucinate consensus, "adversarial" prompting, which explicitly instructs the model to preserve and attribute disagreement, achieves near-perfect neutrality scores.

> History is that certainty produced at the point where the imperfections of memory meet the inadequacies of documentation.
>
> Julian Barnes

1 Introduction

Wikipedia is the first place many people turn to for historical information. It aims to provide a "Neutral Point of View" (NPOV), representing conflicting perspectives fairly and without bias. However, Wikipedia is not a single unified record. It is a federation of independent language editions, each written by a distinct community of editors who often rely on their own national historiography. When history is contested, such as in wars or territorial disputes, these communities can produce articles that describe the same event in fundamentally different ways.

This problem goes beyond simple translation differences. A reader of the Romanian Wikipedia might learn about a heroic defense of independence, while a reader of the Turkish or Russian Wikipedia reads about a failed rebellion or a justified annexation. While the NPOV policy exists in all editions, its application varies. Local editors often cite local sources, creating "citation universes" that rarely overlap. As a result, the bias is not just in the text, but in the choice of which facts to include and which authorities to trust.

In this paper, we study how these conflicting narratives form and evolve. We focus on three events that are central to Romanian history but contested by its neighbors: the Battle of Posada (1330) against Hungary, the Soviet occupation of Bessarabia (1940) involving Russia, and the Night Attack at Târgoviște (1462) against the Ottoman Empire. We analyze articles from five language editions (Romanian, Hungarian, Russian, Turkish, and English) to answer three questions:

1. How do citation choices reflect national bias? We find that editors almost exclusively cite authors from their own country. In our analysis of the Battle of Posada, only 2 out of 119 citations were shared between the Romanian, Hungarian, and English editions.

2. How do narratives evolve over time? By analyzing article versions from 2005 to 2024, we show that these narratives are not static. While the Romanian account of the Night Attack has moderated in recent years, the Russian account of Bessara-

Proceedings of NLPerspectives @ LREC 2026, pages 11–20
12 May 2026.

bia has been volatile, shifting significantly in 2024.

3. Can AI help reconcile these conflicts? We test whether LLMs can act as neutral arbiters. We find that standard prompting fails because models try to "fix" the conflict by choosing one side. However, an "adversarial" prompting strategy, which explicitly instructs the model to preserve the disagreement, can generate summaries that are rated as perfectly neutral.

It is important to note that "neutrality" in history is a contested concept. Often, there is no single objective truth between two national myths. Therefore, we do not define neutrality as finding a "middle ground" or a "correct" version of events. Instead, we define it operationally as *perspectival balance*: a neutral summary is one that accurately represents the existence and nature of the conflict itself, attributing claims to their respective traditions without endorsing one over the other.

2 Related Work

The challenge of addressing conflicting historical narratives on Wikipedia sits at the intersection of computational social science, natural language processing, and digital history. While early work focused on metadata-driven analysis of "edit wars", recent advances in LLMs have shifted attention toward semantic analysis of bias and automated conflict resolution.

2.1 Epistemic Communities & Cross-Lingual Divergence

Wikipedia is ostensibly a global project, but empirical research suggests it functions more as a federation of distinct "epistemic communities" (Samoilenko et al., 2016). Hecht and Gergle (2009) first quantified the "self-focus bias" inherent in these communities, showing that language editions disproportionately cover topics related to their own geography. Miquel-Ribé and Laniado (2018) expanded this to 40 languages, revealing a persistent "culture gap" where shared knowledge is the exception rather than the rule.

In the domain of history, this divergence often manifests as "citation isolation". Taylor et al. (2025) found that non-English Wikipedias cite millions of unique sources not found in English Wikipedia, effectively creating parallel knowledge bases. Baigutanova and Others (2023) further showed that sources deemed unreliable in one language often persist in others, highlighting the lack of a unified standard for verifiability. Our work provides a granular case study of this phenomenon: we show that for the Battle of Posada, the "Romanian epistemic community" and the "Hungarian epistemic community" rely on entirely disjoint sets of historical authorities, with only 2 shared sources out of 119.

2.2 Bias Detection: From Syntax to Semantics

Traditional approaches to bias detection in NLP have focused on linguistic cues. Recasens et al. (2013) created the foundational "NPOV corpus", identifying subjective intensifiers (e.g., "famous", "outrageous") as markers of bias. Pryzant et al. (2020) advanced this by using BERT-based models to automatically rewrite such sentences into neutral forms.

However, historical bias is often implicit and structural rather than purely stylistic. Rogers and Sendijarevic (2012) demonstrated this in their manual analysis of the Srebrenica massacre articles, where the bias lay in **which** facts were selected rather than **how** they were phrased. Recent work by Ghanbari Haez and Dragoni (2025) confirms that modern LLMs still struggle with this "narrative bias", often reproducing intersectional identity biases even when explicitly instructed to be neutral. Our methodology addresses this by moving beyond sentence-level style transfer to document-level narrative analysis, using LLMs to quantify the "stance" of citations and the evolution of historical framing over decades.

2.3 Perspectivism and Automated Reconciliation

The emerging "Perspectivist Data" paradigm (Perspectivist Data Manifesto, 2020) argues that disagreement in data should not always be aggregated away or treated as noise. Instead, valid conflicting perspectives should be preserved. This is particularly relevant for LLMs, which tend to hallucinate a single "consensus" reality when none exists. Köksal et al. (2023) and Li et al. (2023) have shown that LLMs exhibit strong "nationality bias", often aligning with the geopolitical views of their training data's dominant language.

To mitigate this, recent frameworks like "MoDS" (Moderating a Mixture of Document Speakers) (Balepur et al., 2025) and "Multi-Perspective Fusion" (Guan et al., 2025) have proposed treating documents as debating agents. Our "Peace-Maker" pipeline extends this line of inquiry. We show that

standard "summarization" prompts fail because they encourage the model to resolve conflict (a form of **hallucinated consensus**), whereas "adversarial" prompts that enforce the preservation of conflict, aligning with the perspectivist approach, achieve significantly higher neutrality scores.

3 Methodology

Our approach combines historical data mining with LLM-based analysis to quantify bias and generate neutral narratives. The pipeline consists of three stages: (1) multi-lingual data collection, (2) granular stance classification, and (3) adversarial narrative generation.

3.1 Data Collection

We targeted three historical events chosen for their contentious nature in Eastern European historiography:

1. **Battle of Posada (1330):** A foundational conflict between Wallachia and Hungary.
2. **Soviet Occupation of Bessarabia (1940):** A territorial dispute between Romania and the USSR/Russia.
3. **Night Attack at Târgoviște (1462):** A military encounter between Vlad the Impaler (Wallachia) and Mehmed II (Ottoman Empire).

For the **citation analysis** (Posada), we extracted the full content of the Romanian (ro), Hungarian (hu), and English (en) Wikipedia articles as of February 2026. We used a standard github library (Kurtovic, 2023) to parse the MediaWiki markup and extract all citations, including those in '<ref>' tags and bibliography sections.

For the **temporal analysis** (Bessarabia and Târgoviște), we used the Wikipedia Action API (Wikimedia Foundation, 2024) to fetch historical snapshots of the articles from January 1st of 2005, 2010, 2015, 2020, and 2024. This resulted in a dataset of 29 article versions across Romanian, Russian, Turkish, and English editions.

3.2 Stance Classification & Metrics

To quantify bias, we used LLMs and human annotators.

3.2.1 Citation Stance (Posada)

We classified 119 citations into four categories: *Pro-Romanian*, *Pro-Hungarian*, *Neutral*, and *Contested*. A major challenge was that many citations in the bibliography lacked context (e.g., just "Djuvara, p. 180"). To address this, we developed a **Context Enhancer** module. For each citation, the module searched the full article text for mentions of the author or title, extracting a ± 300 character window around the mention. This "enhanced context" was then fed to Google's Gemini 3.0 Pro Preview model (Gemini Team, Google, 2025) with the following prompt structure:

> "You are an expert historian analyzing the Battle of Posada (1330) between Wallachia and Hungary. Analyze the following text snippet from a Wikipedia article to determine the STANCE of the citation. The citation being analyzed is: "$citation - text$". The surrounding context is: "$context$". Determine if the citation is used to support a specific narrative:: Pro-RO (heroic defense), Pro-HU (treacherous trap), Neutral, or Contested".

All API calls were executed with a temperature of 0 to ensure deterministic reproducibility.

3.2.2 Human validation

To validate the LLM's classifications, we used two human annotators. One annotator reviewed a random 50% sample of the Romanian and English citations, while a second annotator reviewed the complete set of Hungarian citations. Both annotators assessed the same "enhanced context" provided to the LLM using the identical 4-category scheme (Pro-RO, Pro-HU, Neutral, Contested), blinded to the LLM's ratings. The human-LLM agreement was substantial, yielding Cohen's Kappa scores of 0.80 for Romanian, 0.75 for English, and 0.85 for Hungarian.

3.2.3 Granular Narrative Metrics (Bessarabia/Târgoviște)

For the temporal analysis, we defined event-specific metrics on a 0-100 scale. For example, for the Night Attack, we measured:

- **Vlad Heroism:** Extent to which Vlad is portrayed as a brave defender.
- **Ottoman Threat:** Emphasis on the magnitude of the invading force.
- **Cruelty:** Emphasis on impalement and brutal tactics.

We processed each article version through the LLM to score these metrics, allowing us to track the evolution of the narrative over 20 years. The full prompt used for this granular scoring is provided in Appendix A.

While the prompt requested categorical stances and confidence scores (0.0-1.0), we aggregated these into 0-100 scales for visualization by mapping the confidence of a "Pro-X" classification to a positive score and "Pro-Y" to a negative or lower score, normalized to a 0-100 range where 100 represents the maximum intensity of the national narrative.

3.3 The "Peace-Maker" Pipeline

To address the challenge of representing conflicting narratives, we developed the "Peace-Maker" pipeline.

1. **Claim Extraction:** We first extract key factual claims from each source text (e.g., "Romanian source claims 15,000 Ottoman casualties").

2. **Conflict Matching:** We use an LLM to identify pairs of conflicting claims (e.g., "RO: Attack was a victory" vs. "TR: Attack was a failed assassination").

3. **Adversarial Generation:** We tested four prompting strategies to generate a summary:
   - *Standard:* "Summarize these sources".
   - *Academic:* "Write as a peer-reviewed historian".
   - *Mediator:* "Write a diplomatic report acceptable to both sides".
   - *Adversarial:* "Preserve the conflict. Explicitly state 'Source A says X while Source B says Y'. Do not resolve the dispute".

4. **LLM-as-a-Judge:** We evaluated the generated summaries using a separate LLM instance to score them on *Neutrality* (0-1), *Conflict Preservation* (0-1), and *Source Attribution* (0-1).

4 Experiments & Results

We present results from three experiments: (1) citation stance analysis of the Battle of Posada, (2) temporal analysis of narratives surrounding Bessarabia and Târgoviște, and (3) the evaluation of the Peace-Maker pipeline.

4.1 Experiment 1: Citation Isolation (Posada)

We analyzed 119 citations across the Romanian (RO), Hungarian (HU), and English (EN) Wikipedia articles on the Battle of Posada (1330).

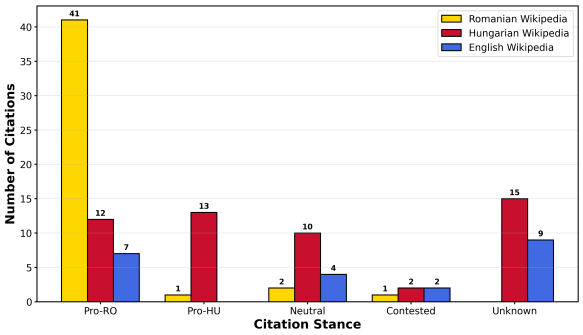


Figure 1: Raw count of citations by stance. Note the overwhelming volume of Pro-Romanian citations in the Romanian edition compared to the balanced distribution in Hungarian.

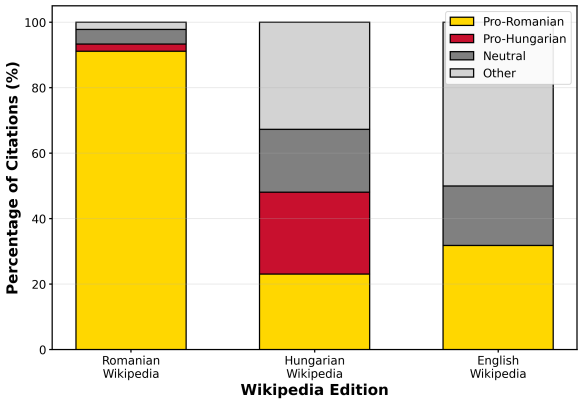


Figure 2: Stance distribution of citations in the Battle of Posada articles. Romanian Wikipedia is 91% Pro-Romanian, while Hungarian Wikipedia is balanced.

Bias Scores: The Romanian edition exhibited a strong national bias, with 91.1% of citations classified as *Pro-Romanian* and only 2.2% as *Pro-Hungarian* (Figure 2). In contrast, the Hungarian edition was remarkably balanced, with 23.1% *Pro-Romanian* and 25.0% *Pro-Hungarian* citations. The English edition, often assumed to be neutral, showed a moderate Pro-Romanian lean (31.8% Pro-RO vs 0% Pro-HU).

Figure 1 highlights the sheer volume disparity. It is important to note that raw citation volume is not a proxy for quality; a higher count could simply indicate a more detailed article. However, in this context, the volume reinforces the echo chamber. The Romanian article builds an "illusion of consensus" through the repetition of a large, monoperspectival corpus, whereas the Hungarian article achieves a balanced narrative with a smaller but more diverse set of references.

Citation Isolation: The most striking finding was the lack of overlap. Out of 119 unique ci-

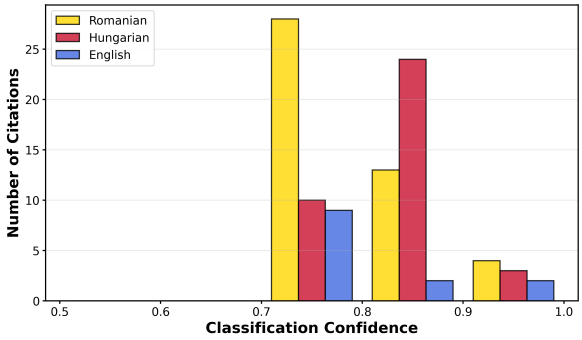


Figure 3: LLM Confidence Distribution. The high confidence scores across all stances indicate that the "Citation Isolation" is not due to model uncertainty but reflects clear, distinct narrative framing in the sources.

tations, only **2 sources** were shared across more than one language edition. This confirms that these Wikipedia communities operate in distinct "citation universes", constructing their narratives from entirely disjoint sets of historical authorities. It is important to note that this isolation is partly driven by linguistic accessibility: Romanian scholars naturally cite Romanian historians writing in Romanian, while Hungarian scholars cite Hungarian sources. However, the result is the same: readers of different language editions are presented with fundamentally different evidentiary bases. As shown in Figure 3, the LLM's classification confidence remained high across all categories, validating that these sources express clear, unambiguous stances rather than vague or subtle biases.

4.2 Experiment 2: Temporal Evolution

We tracked narrative metrics across 29 article versions from 2005 to 2024.

Bessarabia (1940): The Romanian article for "Basarabia" showed significant moderation, with its Pro-Romanian score dropping from a peak of 76 (2010) to 55 (2024) (Figure 4, Left). Conversely, the Russian article was highly volatile. After fluctuating between Pro-Soviet (30) and balanced (65) stances, the 2024 version of the "Accession" article spiked to a Pro-Romanian score of 72. We speculate that this may reflect a post-2022 shift where Russian editors are becoming more critical of Soviet expansionism, though further qualitative research is needed to confirm this causal link.

Figure 5 reveals the specific dimensions of this divergence. While both editions now acknowledge the Molotov-Ribbentrop Pact, they differ fundamentally on the consequences. The Romanian edition emphasizes "Deportations" (Score: 40) and "Victimhood" (Score: 80), whereas the Russian edition emphasizes "Soviet Justification" (Score: 40). Figure 6 tracks this "Deportation Gap" over time, showing that while Romanian mentions of atrocities have remained high, Russian mentions have been near-zero for two decades, only appearing slightly in 2024.

Night Attack (1462): The Romanian article peaked in nationalism in 2020 (Score: 85), portraying Vlad the Impaler as a "valiant defender", before moderating to 72 in 2024 (Figure 4, Right). The Turkish edition showed a gradual decline in Pro-Romanian sentiment (70 $\rightarrow$ 65) and consistently emphasized Vlad's cruelty (Score: 80) far more than the Romanian edition (Score: 60). The English edition remained remarkably stable at a score of 70 throughout the 14-year period.

The shape of these narratives is visualized in Figure 7. The Turkish narrative is defined by high "Cruelty" and "Ottoman Threat" scores but lower "Heroism". The Romanian narrative is the inverse. Figure 8 isolates the "Cruelty Gap", showing a persistent 15-20 point difference between Turkish and Romanian portrayals of Vlad's tactics.

4.3 Experiment 3: Peace-Maker LLM

We evaluated four prompting strategies for generating neutral summaries of the Night Attack, using conflicting claims extracted from Romanian and Turkish articles.

Prompt Strategy	Neutrality	Conflict Pres.
Standard	0.47	0.60
Mediator	0.95	1.00
Academic	0.97	1.00
Adversarial	**1.00**	**1.00**

Table 1: Performance of prompting strategies. Neutrality and Conflict Preservation are scored 0-1.

As shown in Table 1 and Figure 9, the *Standard* prompt failed (Neutrality: 0.47) because the LLM attempted to resolve the conflict, often choosing one side's version of events (e.g., regarding the attack's success). The *Adversarial* prompt, which explicitly instructed the model to *preserve* conflict, achieved perfect scores. Figure 10 breaks this down further, showing that the Adversarial prompt's success stems from its high "Source Attribution" score—it explicitly attributes disputed claims ("Romanian sources state X.."), avoiding the trap of hallucinated consensus.

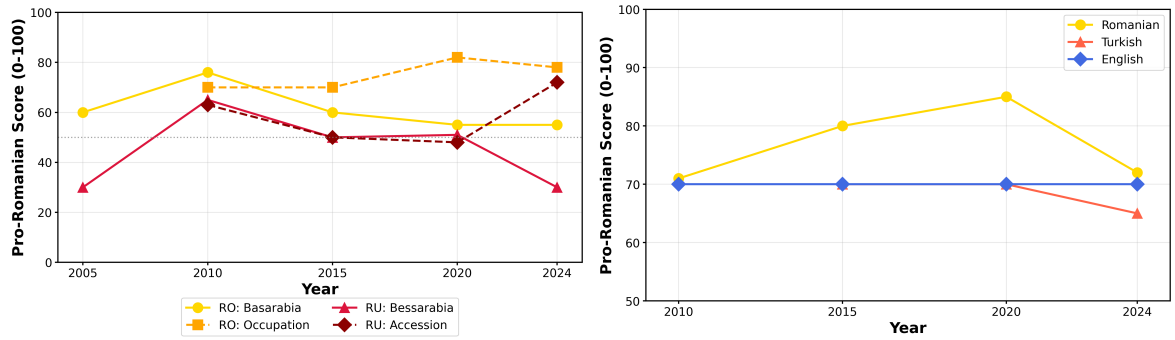


Figure 4: Evolution of the "Pro-National" score over time. Left: Bessarabia (RO vs RU). Right: Night Attack at Târgoviște (RO vs TR vs EN).

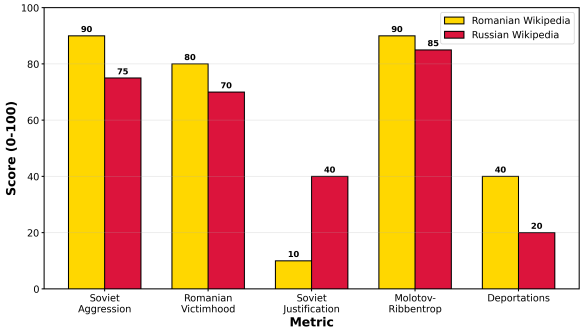


Figure 5: Detailed metric comparison for Bessarabia (2024). Note the divergence in "Soviet Justification" and "Deportations".

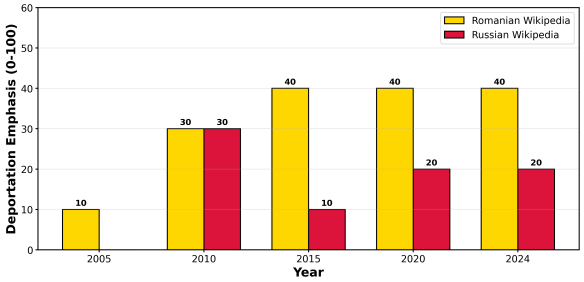


Figure 6: Deportation mentions over time. Romanian Wikipedia consistently highlights Soviet atrocities, while Russian Wikipedia largely omits them.

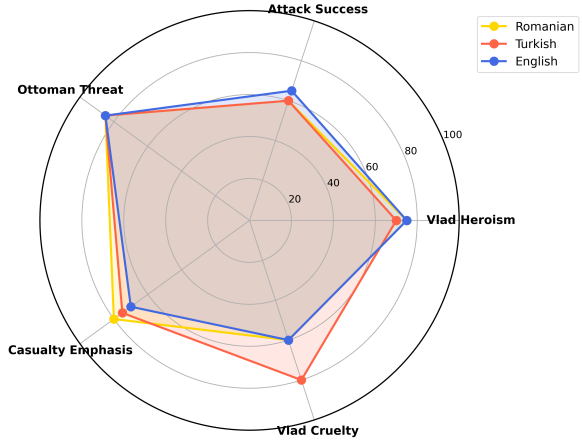


Figure 7: Radar chart of narrative metrics for the Night Attack (2024). The shapes illustrate distinct narrative priorities: Turkish (Cruelty-focused), Romanian (Heroism-focused), and English (Balanced).

5 Discussion

Our findings challenge the assumption that Wikipedia functions as a unified global encyclopedia. Instead, it operates as a federation of distinct epistemic communities, each validating knowledge through its own national lens.

5.1 The Illusion of Neutrality

The "Citation Isolation" we observed in the Posada case study shows the challenge of achieving neutrality. An article can adhere perfectly to NPOV guidelines, using neutral language and avoiding editorializing, while still presenting a heavily biased narrative simply by selecting sources from a single national tradition. This "bibliography bias" is invisible to standard NLP tools that focus on sentiment analysis, but it is the primary driver of narrative divergence in historical topics.

5.2 Narrative Evolution and Geopolitics

Our temporal analysis shows that these narratives are not static. The moderation of the Romanian "Night Attack" article (from 85 to 72) suggests that community maturity and international scrutiny can temper nationalism over time. However, the volatility of the Russian "Bessarabia" article demonstrates that Wikipedia is not immune to external geopolitical shocks. The sudden 2024 shift toward a Pro-Romanian stance likely reflects a broader realignment of Russian opposition discourse following the invasion of Ukraine, where Soviet imperial

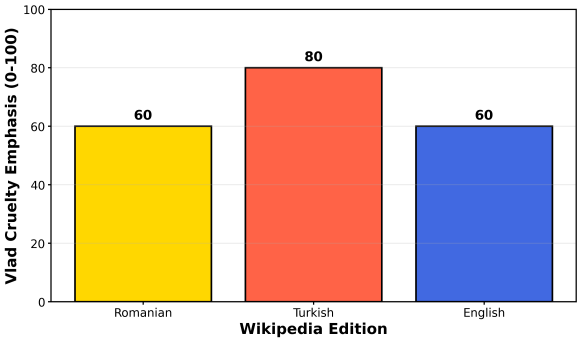


Figure 8: Comparison of "Vlad Cruelty" scores. Turkish Wikipedia consistently emphasizes Vlad's brutality much more than the Romanian edition.

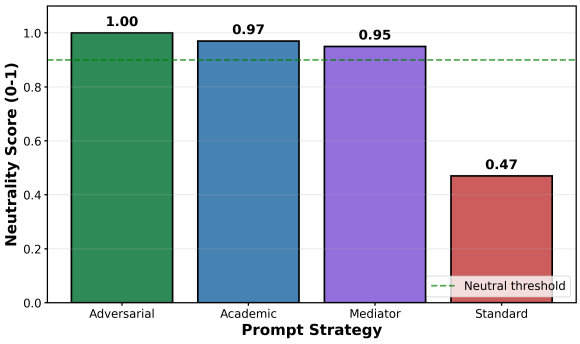


Figure 9: Neutrality scores (0-1) for different prompting strategies. The Adversarial prompt achieves perfect neutrality.

history is being critically re-examined.

Bias often manifests as silence. As Figure 6 illustrates, the Russian Wikipedia's historical omission of Soviet deportations is a form of narrative control. The sudden appearance of these mentions in 2024 signals a potential "thaw" in this historiographical freeze. Similarly, the "Cruelty Gap" in the Targoviste narrative (Figure 8) shows how national identity is constructed not just by what is celebrated (Heroism), but by what is minimized (Cruelty).

5.3 AI as a Mediator, Not a Judge

Standard LLM summarization fails because it mimics the human tendency to seek a single, coherent truth, effectively hallucinating a consensus where none exists. By explicitly prompting the model to be "adversarial" and preserve conflict, we force it to act as a mediator rather than a judge.

However, a distinction must be drawn between interpretive disagreements (e.g., whether a war was "justified") and factual contradictions (e.g., casualty counts). For the latter, "preserving conflict" does not imply that all claims are equally valid, but

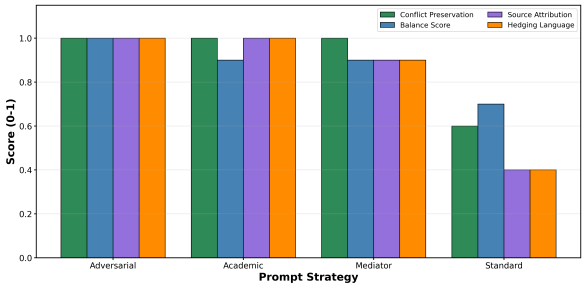


Figure 10: Detailed evaluation of generated summaries. The Adversarial prompt excels not just in Neutrality, but in Source Attribution and Conflict Preservation.

rather that the *disagreement itself* is a historical fact worth reporting. When Source A claims 15,000 casualties and Source B claims 1,500, a LLM that averages these numbers fails both. A LLM that reports the discrepancy preserves the epistemic integrity of the conflict. Future systems must balance this perspectivism with rigorous fact-checking to avoid "false balance" when one perspective is demonstrably pseudohistorical. This suggests that the future of AI in education and historiography lies not in generating "objective" answers, but in synthesizing and attributing diverse perspectives.

6 Conclusion and Future Work

This study demonstrates that Wikipedia functions not as a unified global archive, but as a federation of distinct epistemic communities defined by "citation isolation". Our analysis of the Battle of Posada revealed that the Romanian and Hungarian editions shared only 2 out of 119 citations, constructing mutually exclusive historical realities based on non-overlapping authorities. Longitudinally, we found these narratives to be volatile and responsive to geopolitics, evidenced by the 2024 shift in the Russian framing of Bessarabia.

Methodologically, we show that standard LLM prompting fails to address this by hallucinating consensus. Instead, we propose a "Peace-Maker" pipeline using "adversarial" prompting. By explicitly instructing models to preserve conflict rather than resolve it, we achieve "perspectival balance", generating summaries that accurately attribute disagreement without enforcing a false middle ground.

We propose several directions for future research:

- **Baseline Comparison:** Compare these contested events to uncontested historical topics to establish a baseline for citation overlap

and narrative divergence. This would verify whether the "citation isolation" we observed is specific to conflict or a general feature of Wikipedia's language editions.

- **Geographic and Domain Expansion:** While our case studies are well-chosen for their contentiousness, they are limited to Eastern European history. Future research should investigate if "Citation Isolation" holds true for global topics (e.g., WWII in the Pacific or Colonialism in the Americas) or scientific controversies.

Limitations

Our study relies on human annotators and LLMs. LLM-based classification, while calibrated, may still contain inherent biases. We focused on three specific events in Eastern European history; results may vary for other regions. Our validation methodology presents a potential circularity: annotators assigned by language (Romanian/English vs. Hungarian) were native speakers whose historiographical perspectives may reflect the national biases we sought to measure, rather than providing an independent benchmark. Annotator neutrality is hard to verify.

Ethics Statement

This study utilizes publicly available data under the Creative Commons Attribution-ShareAlike 4.0 license. To validate our LLM-based classifications on sensitive historical topics, we used two human annotators. Strict privacy protocols were maintained to preserve their anonymity. We emphasize that our computational metrics quantify "perspectival balance" and are not intended to adjudicate historical truth or resolve geopolitical disputes.

Acknowledgments

This research is supported by:

- the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416;

- a grant of the Ministry of Research, Innovation and Digitization, CNCS - UEFISCDI, project SIROLA, number PN-IV-P1-PCE-2023-1701, within PNCDI IV;

A Granular Narrative Scoring Prompt

The following prompt was used for temporal analysis of the Bessarabia and Târgoviște articles:

```
Analyze the following Wikipedia article excerpt about Bessarabia/Soviet occupation.
ARTICLE: {article}
LANGUAGE: {lang} (Romanian or Russian)
YEAR: {year}
TEXT:
{text}
—
Analyze the NARRATIVE STANCE of this article. Consider:
1. **Overall Stance**: Is the article Pro-Romanian, Pro-Russian, Neutral, or Mixed?
- Pro-Romanian: Portrays Soviet actions as occupation, theft, aggression
- Pro-Russian: Portrays Soviet actions as liberation, reunification, protection
- Neutral: Balanced presentation of facts without emotional framing
- Mixed: Contains elements of both perspectives
2. **Key Themes**: What are the main themes/topics discussed? (list 3-5)
3. **Emotional Tone**: What is the emotional framing?
- Accusatory (blaming the other side)
- Neutral (factual, academic)
- Defensive (justifying actions)
- Victimization (emphasizing suffering)
4. **Key Terminology**: What loaded terms are used? For example:
- "occupation" vs "liberation" vs "annexation" vs "reunification"
- "ultimatum" vs "agreement" vs "request"
- How is the USSR/Romania described?
Respond in this EXACT JSON format:
{
"overall_stance": "Pro-RO" or "Pro-RU" or "Neutral" or "Mixed",
"stance_confidence": 0.0 to 1.0,
"key_themes": ["theme1", "theme2", "theme3"],
"emotional_tone": "Accusatory" or "Neutral" or "Defensive" or "Victimization",
"terminology": {
"event_name": "what the event is called",
"soviet_actions": "how Soviet actions are described",
"key_loaded_terms": "any loaded/biased terms used"
},
"reasoning": "Brief explanation of why you classified it this way"
}
```

References

A. Baigutanova and Others. 2023. Reference reliability divergence in multilingual wikipedia. *arXiv preprint arXiv:2309.00196*.

Nishant Balepur, Alexa Siu, Nedim Lipka, Franck Dernoncourt, Tong Sun, Jordan Boyd-Graber, and Puneet Mathur. 2025. Mods: Moderating a mixture of document speakers to summarize debatable queries in document collections. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Gemini Team, Google. 2025. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. Updated to include Gemini 3, November 2025.

Saba Ghanbari Haez and Mauro Dragoni. 2025. Neutral is not unbiased: Evaluating implicit and intersectional identity bias in llms through structured narrative scenarios. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Xin Guan, Pei-Hsin Lin, Zekun Wu, Ze Wang, Ruibo Zhang, Emre Kazim, and Adriano Koshiyama. 2025. Mpf: Aligning and debiasing language models post deployment via multi-perspective fusion. *Findings of IJCNLP 2025*.

Brent Hecht and Darren Gergle. 2009. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 11–20.

Abdullatif Köksal, Omer Yalcin, Ahmet Akbiyik, M. Kilavuz, Anna Korhonen, and Hinrich Schuetze. 2023. Language-agnostic bias detection in language models with bias probing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12735–12747.

Ben Kurtovic. 2023. mwparserfromhell: A parser for mediawiki wikicode. https://github.com/earwig/mwparserfromhell.

Bryan Li, Samar Haider, and Chris Callison-Burch. 2023. Geopolitical bias in large language models. *arXiv preprint arXiv:2305.14610*.

Marc Miquel-Ribé and David Laniado. 2018. Wikipedia culture gap: quantifying content imbalances across 40 language editions. *Frontiers in Physics*, 6:54.

Perspectivist Data Manifesto. 2020. The perspectivist data manifesto. https://pdai.info/. Accessed: 2026-02-09.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjectivity in text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8669–8676.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659.

Richard Rogers and Emina Sendijarevic. 2012. Neutral or national point of view? a comparison of srebrenica articles across wikipedia's language versions. In *Wikipedia Academy: Research and Free Knowledge*, Berlin. Wikipedia Academy.

Anna Samoilenko, Fariba Karimi, Daniel Edler, Martin Rosvall, and Markus Strohmaier. 2016. Linguistic neighbourhoods: explaining cultural borders on wikipedia through multilingual co-editing activity. *EPJ Data Science*, 5:1–21.

Michael Taylor, Roisi Proven, and Carlos Areia. 2025. Evaluating the diversity of scientific discourse on twenty-one multilingual wikipedias using citation analysis. *arXiv preprint arXiv:2501.09666*.

Wikimedia Foundation. 2024. Mediawiki action api. https://www.mediawiki.org/wiki/API:Main_page. Accessed: 2025-02-09.

GSI:detect - A Perspectivist Approach to Gender Stereotypes Identification in Italian

Davide Testa[1,2], Sofia Brenna[1,3], Manuela Speranza[1], Gloria Comandini[4], Stefania Cavagnoli[5], Bernardo Magnini[1]

[1]Fondazione Bruno Kessler (FBK), [2]University of Rome La Sapienza, [3]Free University of Bozen-Bolzano
[4]Istituto Italiano di Studi Germanici (IISG), [5]University of Trento
{dtesta, sbrenna, manspera, magnini}@fbk.eu
comandini@studigermanici.it, stefania.cavagnoli@unitn.it

Abstract

The deconstruction of gender stereotypes is essential to prevent discrimination, marginalization and gender-based violence. Despite the increasing attention to this issue, research in this field often focuses on explicitly sexist or hateful communication, leaving out all the cases where stereotypes are produced unconsciously or even with apparently positive intentions. Moreover, the identification and analysis of gender stereotypes is often a very subjective task, heavily influenced by the researcher's background, beliefs and personal sensitivity. In this context *GSI:detect*, a dataset for gender stereotypes identification in Italian, has been annotated following a perspectivist approach that gives value to the different points of view of four annotators. It has been designed to address (i) the lack of resources focusing on naturally occurring and non-hateful language conveying implicit or ambiguous forms of gender stereotypes, and (ii) the scarcity of datasets that can capture multiple interpretations as well as the inherent variation and disagreement in human perception. Baseline experiments with several LLMs confirm the challenging nature and value of such a linguistic resource, revealing both apparent differences and limitations in performance among the evaluated models, and raising questions about the extent to which current LLMs are suitable for detection and classification tasks in this field.

Content warning: Examples taken from the GSI:detect dataset may contain sensitive or potentially distressing content.

Keywords: gender stereotypes, learning from disagreement, dataset, linguistic resources, perspectivism

1. Introduction

In the context of Italian language and culture, stereotypes are studied by many areas of expertise, such as psychology, sociology and, more importantly for this work, linguistics (Arcuri and Cadinu, 1998; Peruzzi et al., 2019b; Cavagnoli and Dragotto, 2021). We can define stereotypes as pre-constituted, generalised and simplistic opinions about people, events, and situations. Stereotypical opinions are not based on personal evaluation of individual cases, but are mechanically repeated, creating what has been described as a "culturally constructed cage" (Biemmi, 2020). Therefore, stereotypes hinder critical thinking (Biemmi, 2020) and elicit the insurgence of prejudice, as well as discrimination (Jackson, 2011).

Gender stereotypes (GSs) are stereotypes based on socially constructed beliefs regarding the "appropriate" roles, behaviours, and appearances that people should have according to their assigned gender. Therefore, GSs are a cultural conformist force that pushes individuals to shape their personalities based on specific social expectations, which in the Western world are polarized into two groups: men and women (Peruzzi et al., 2019a)[1]. From this perspective, the two poles are in a hierarchical relationship, with the female pole subordinate to the male pole (Biemmi, 2020). Moreover, deviation from the stereotypical perceived "normalcy" of femininity and masculinity, including the "contaminations" between the two genders (i.e. women who adopt features culturally perceived as masculine: playing soccer, having short hair, not wearing makeup, etc. - and the other way round for men), may result in social marginalization and stigmatization (Peruzzi et al., 2019a).

This situation makes the deconstruction of GSs necessary to prevent discrimination and gender-based violence. Consequently, in recent years the rise of Natural Language Processing (NLP) has made it possible to automatically identify and analyse the linguistic expressions through which GSs are conveyed. However, despite the increasing attention to this issue, research in this field often focuses on explicitly sexist or hateful communication, leaving out all the ambiguous cases where stereotypes are produced unconsciously or even with apparently positive intentions. Moreover, the subjectivity involved in identifying such phenomena has often been simplified through binary annotation

[1]Although our present study focuses on stereotypes targeting women and men, this choice is purely methodological and does not assume in any way a binary view of gender. The analysis of GSs applied to non-binary people might be the object of future works.

Proceedings of NLPerspectives @ LREC 2026, pages 21–32

schemes, obscuring the diversity of human point of views and calling for evaluation frameworks that explicitly account for annotator variability.

In this context, we introduce the GSI:detect dataset, a new Italian linguistic resource for the study of GSs in short texts. The dataset is designed to address two main gaps in current research: (i) the lack of resources focusing on naturally occurring, non-hateful language containing implicit or ambiguous forms of GSs, while also contributing to a more comprehensive understanding of the phenomenon by including stereotypes directed at both women and men; and (ii) the scarcity of datasets that adopt a perspectivist approach to annotation, useful for capturing multiple interpretations as well as the inherent diversity and disagreement in human perception of stereotypes. With this purpose in mind, GSI:detect offers a novel resource for both linguistic and computational research. From a linguistic perspective, it allows the exploration of heterogeneous ways in which stereotypes manifest in everyday Italian discourse. From a computational point of view, it provides a challenging test-bed to assess how well Large Language Models (LLMs) are able to identify, interpret, classify, and reason about GSs in Italian, especially in situations where subjectivity plays a central role.

The paper is organized as follows. Section 2 presents some related work, focusing both on GSs in NLP and on the perspectivist approach. In Section 3, we describe the GSI:detect dataset, outlining the data collection procedure, the manual annotation process, and related statistics. Finally, Section 4 details the experimental setup, including task design and evaluation metrics, and Section 5 presents and discusses the results, illustrating the effectiveness of GSI:detect for assessing model performance on these tasks.

2. Related Work

2.1. Gender Stereotypes in NLP

GSs have been studied extensively in NLP; however, for the purposes of this work, we focus on two primary research directions.

The first one is the automatic recognition of GSs in texts produced by humans and, more recently, also produced by LLMs. These investigations are generally applied in the context of hateful sexist communication, and therefore are often associated with hate speech (HS) detection. This is due to the fact that most cases of misogynistic hate speech are also imbued with GSs (Fersini et al., 2018). For example, in Kirk et al. (2023) GSs are a subcategory of sexist animosity, while in Plaza et al. (2023) they are one of the many categories of sexism. Similarly, in the context of the Italian language, stereotypes detection has been mostly a subject in automatic recognition of misogynistic hate speech (Fersini et al., 2018).

The second field is the investigation of GSs encoded in LLMs, due to the presence of prejudiced material in their training datasets. In fact, it has been widely assessed that LLMs are prone to spread stereotypes, prejudices and, more generally, social biases regarding, for example, race, ethnicity, religion, age, sexual orientation and, of course, gender identity (Cao et al., 2022; Ovalle et al., 2023), due to the presence of these biases in the human-produced texts used to train them (Talat et al., 2022).

Taking into account the results obtained from these two fields of research, our investigation aims to go one step further: to discover how well LLMs - given their inclination to spread social biases - perform in recognizing GSs. A similar study has been recently conducted by Mitchell et al. (2025), who focused on the recognition of multilingual social stereotypes by LLMs. The dataset used by Mitchell et al. (2025) is, however, quite different from ours, as it consists of a set of short stereotypical texts [2], paired with non-stereotypical counterparts generated through a template-based process. On the other hand, as will be better explained in Section 3.1, our dataset (in addition to focusing solely on GSs) consists of texts that were not produced specifically for our experiment, but rather texts that were naturally written in uncontrolled online environments. In this way, we wanted to collect not only the gender stereotypes that a group of experts might imagine, but also and above all the unexpected and ambiguous cases that arise in spontaneous writing on the web.

Furthermore, we did not want to investigate GSs only in hate speech. In fact, while they can be found in misogynistic hate speech (Kirk et al., 2023; Fersini et al., 2018), they can appear in non-hateful communication as well, as unconscious stereotypes can also be used with well-meaning intentions by both men and women (e.g., *a woman's intuition is never wrong!*). From this perspective, GSs can be intertwined with the phenomenon of micro-aggressions (Sue, 2010), whose nature of implicit indignity makes them particularly prone to being produced even by their own victims (e.g., from: *I didn't do well [in a science exam], but, oh well, girls aren't supposed to be good at science anyway, ha-ha.* (Harrison and Tanner, 2018).

[2]The texts from Mitchell et al. (2025) were produced by a group of data creators who spoke different native languages and regarding not only GSs, but also prejudices about age, race, physical appearance etc.

2.2. The Perspectivist Approach

Recognizing stereotypes is often a very subjective task, as seen in Sanguinetti et al. (2018) where the inter-annotator agreement between expert annotators for racist stereotypes is 0.41 (Cohen's *k*). In fact, due to their pervasiveness and sometimes implicit nature, highly subjective tasks performed by humans (such as GS or HS recognition[3].) can be heavily influenced by interiorized biases (Basile et al., 2023; Muscato et al., 2024), level of expertise in the task, affinity or belonging to the group victim of prejudice (Wojatzki et al., 2018), and even more generic personal opinions (Klenner et al., 2020).

Therefore, disagreement in annotation is not always caused by lack of attention or other kinds of genuine mistakes, but might be a useful cue for exploring the complexity of human experience with respect to subjective themes. This methodology has been defined a Perspectivist Approach (Basile, 2020; Basile et al., 2023; Rizos and Schuller, 2020; Muscato et al., 2024), as it revolves around the idea that NLP should not rely on the classic gold-standard corpora with a majority-aggregated annotation, but should adopt new strategies in order to integrate the added value of opinion's diversity in annotation. In fact, several studies, see for example Klenner et al. (2020), underline that majority voting may suppress rightful points of view that add new information about a topic, and that they should not be classified as errors just because they are a minority vote.

Given that GS annotation can be a very subjective task, we decided to adopt a Perspectivist approach in the creation of our dataset, aiming to take advantage of the inherent subjectivity of the task to better understand the reasons behind different annotations and to study the ambiguous gray area in the continuum between explicit GSs and non-stereotyped communication. While prior work models annotator disagreement as a full probability distribution (Madeddu et al., 2023), more recent approaches estimate or elicit such distributions directly from models, either via token-level probabilities over a closed label space (Santurkar et al., 2023) or by prompting models to output probability distributions (Pavlovic and Poesio, 2024). In contrast, we adopt a simpler scalar formulation based on the mean of binary judgments across annotators, which preserves disagreement in a compact form while ensuring a model-agnostic evaluation setting, particularly for closed decoder-based models where access to probability distributions is limited.

[3]However, it is important to underline that there are also different opinions on the subjectivity of HS, such as (Cercas Curry et al., 2024). Moreover, while there is a level of subjectivity in the annotation of phenomena such as GSs and HS, this subjectivity does not deny the harm done by HS and GSs, nor the fact that there are certain kinds of hateful or stereotyped texts that are recognised as such with an extremely high agreement. This is the case of, for example, extremely sexist HS (e.g. *women like to be raped*), which is recognized as such by both male and female annotators, despite their disagreement on more subtle forms of sexism (e.g. *female quotas are useless*) (Wojatzki et al., 2018).

3. The GSI:detect Dataset

GSI:detect is a new Italian resource developed for the detection, analysis and classification of gender stereotypes in written texts; it has been used as the reference dataset for a shared task with the same name organized within the Evalita 2026 Evaluation Campaign[4] (Comandini et al., 2026) and is distributed under a *Creative Commons NonCommmercial-ShareAlike License*. The dataset is publicly available on its official GitHub repository.

It consists of 1,010 short written Italian texts (for a total of 52,118 tokens), collected to capture authentic, naturally occurring language and to represent a wide range of communicative contexts in which gender stereotypes may appear at different levels of prototypicality, or not at all.

3.1. Data Collection

The texts included in GSI:detect have been manually collected from both social media and informative websites, in order to provide a balanced representation of both formal and informal written Italian.

We collected comments from discussion threads or articles related to different topics from the following social media:

- social media pages discussing gender issues from diverse ideological perspectives[5];
- Facebook and Instagram pages of major Italian newspapers[6] and sports newspapers (*La Gazzetta dello Sport* and *Eurosport Italia*);
- public Facebook groups related to chess and mathematics;

[4]https://gsi-d-evalita.fbk.eu/

[5]Such as the Instagram page of the feminist influencer Chiara Becchi Manzi, the pick-up artists agency *Playlover Academy* and the "mom influencer" *amoree__di_mamma*, as well as from the Facebook pages of ironic and parodic groups such as *Alpha Woman* and *La società femminista*.

[6]*Domani* and *Il Post* from Instagram, and *ANSA*, *Il Corriere della Sera*, *La Repubblica*, *La Stampa*, *La Verità*, *Open* and *SkyTG24*.

Labels	GS value	Example
no-no-no-no	0	Non comprendo come si possano paragonare due fenomeni, gravissimi entrambi e concordo, come femminicidi e morti sul lavoro. *(I don't understand how one can compare two phenomena, both very serious, and I agree on that, such as femicides and workplace deaths.)*
yes-no-no-no	0.25	Tenete duro ancora qualche giorno e i vostri fidanzati partiranno in vacanza con le loro mogli. *(Hold on for a few more days and your boyfriends will be going on vacation with their wives.)*
yes-yes-no-no	0.50	Io rimango dell'idea che un figlio ha sempre bisogno della sua mamma, anche per dire buongiorno e buona notte. E la mamma idem. Soprattutto la mamma *(I still think an [adult] child always needs his/her mother, even to say good morning and good night. And the mother too. Especially the mother*
yes-yes-yes-no	0.75	[Commento ad articolo di giornale dal titolo "Negli Usa quasi un manager su due è donna. In Italia meno di 1 su 3"] Infatti il Made usa va' peggio del Made italy *([Comment on a newspaper article titled "In the US, almost one in two managers is a woman. In Italy, less than one in three"] In fact, Made in USA is doing worse than Made in Italy)*
yes-yes-yes-yes	1	[Rivolto a una utente donna] fatevi voi una doccia e copritevi. Le donne vere si coprono. Gli animali vanno in giro nudi. *([Addressed to a female user] Take a shower and cover up. Real women cover up. Animals go around naked.)*

Table 1: GS values corresponding to the five possible combinations of labels assigned by the four annotators (non aggregated labels).

- more generic sources on Facebook such as gossip pages (*Cosa?*) and pseudo-scientific speculations (*Ghiandola pineale - Il terzo occhio*)
- Reddit pages focusing on dating and relationships (such as *dating_advice*).

As far as informative websites are concerned, we collected excerpts from minor websites spanning from women-oriented spaces (e.g. *Femal*) to local Christian websites (e.g. *Amici del Timone*).

This variety allows us to explore GSs in different contexts (e.g., family, sport, politics, etc), as presented in detail in Table 2.

The resulting dataset includes texts that may or may not display stereotypical content within a linguistic structure that distinguishes between two types of items:

- NO CONTEXT texts, which can be understood without additional contextual information (Examples 1 and 3);
- WITH CONTEXT texts, which are not self contained and are therefore enriched with standardized human-generated metadata, which usually contains contextual information such as the headline of a newspaper article (Example 2).

Furthermore, the dataset includes not only GSs regarding women (Example 1), but also GSs regarding men (Example 2), or both (Example 3).

1. Vabbè oggettivamente le femmine su alcune cose non sono in grado. XD Fagli cambiare una ruota di scorta XD

   (Come on, females are objectively incapable of some things. XD Have them change a tire XD)

2. (Commento ad articolo di giornale dal titolo "Il corpo di ballo di Marco Mengoni balla al ristorante sulle note di 'Mi fiderò'") Li vedo bene in guerra contro i russi. XDXDXD

   ([Comment to a newspaper article titled "Marco Mengoni's dance troupe dances at the restaurant to the notes of 'Mi fiderò'"] I can see them doing well at war against the Russians XDXDXD)

3. Paga l'uomo... Se paga lei è lei l'uomo della coppia.

   (The man should pay... If she pays, she is the man of the couple.)

Topic	N. of texts	Percentage
Family	95	9%
Gender	167	17%
Gossip	115	11%
Politics	88	9%
Romance	68	7%
Sport	138	14%
Violence	67	7%
Work	95	9%
Other	177	17%
Total	1010	100%

Table 2: Distribution of the texts by topic.

3.2. Manual Annotation

The whole dataset has been annotated manually by four trained annotators[7], who have spent around three weeks training on the subject and discussing

[7]Following the recommendations of Basile et al. (2023), we provide more in-depth information about the annotators; we do this as disaggregated data to preserve their anonymity. All four are Italian native speakers, cisgender, and sensitive to gender-related issues due to either their studies, or their affinity to feminist movements and/or to the queer community, or their personal experience with gender-based discrimination. The annotation

Category	Example
ROLE	Cento uomini possono creare un accampamento, ma serve una donna per fare una casa. ENG: *A hundred men can build a camp, but it takes a woman to make a home.*
PERSONALITY	Sentivo qualcosa di speciale e sai, una donna non sbaglia mai le sensazioni. ENG: *I felt something special and you know, a woman never gets her feelings wrong.*
COMPETENCE	[Commento ad articolo con titolo "La pilota della British Airways ubriaca in volo: cacciata dall'aereo, aggredisce pure i poliziotti"] Come si possono affidare le sorti di un aereo ad una donna? ....scherzo, naturalmente... ENG: *[Comment on an article titled "British Airways pilot drunk on flight: kicked off plane, she even attacks police"] How can you trust a plane's fate to a woman? ....just kidding, of course...*
PHYSICAL	Oppure c'hanno le 5^, vanno in giro scollate come i manifesti messi d'inverno, e poi se rimani ""attirato"" dalle loro protuberanze ci rimangono male Povere cucciole. ENG: *Or they are a size D, they walk around with low-cut tops like winter posters, and then if you get ""attracted"" by their protuberances, they get upset. Poor little things.* [In the Italian text, there's a pun in the word *scollate*, which can mean both *wearing a low-cut top* and *coming unstuck because the glue has worn out*.]
SEXUAL	[Rivolto a una utente donna] fatevi voi una doccia e copritevi. Le donne vere si coprono. Gli animali vanno in giro nudi. ENG: *[Addressed to a female user] Take a shower and cover up. Real women cover up. Animals go around naked.*
RELATIONAL	[Commento a meme con testo "Aspettavo che mi mandassi tu un messaggio" e sotto l'immagine di un uomo vestito da principessa] Tipico post da zitella ENG: *[Comment on a meme with the text "I was waiting for you to text me" and underneath a picture of a man dressed as a princess] Typical spinster post*

Table 3: Examples of texts belonging to the six GS categories.

and defining the annotation guidelines.[8] Both the extensive preparation phase[9] and the involvement of multiple trained annotators ensured a shared understanding of the task and consistency in the application of the criteria, thereby contributing to the overall quality and reliability of the dataset.

Furthermore, as one of the key contributions of this work, we propose a new taxonomy for the semantic classification of gender stereotypes, with each category representing a different dimension of this phenomenon. This classification, which is outlined in the annotation guidelines mentioned above, was developed to capture the variety of ways in which stereotypes manifest in language and to support both linguistic analysis and automatic detection tasks.

For each text, the following information is provided:

- **GS value**: a number in the interval $[0 - 1]$ indicating the degree to which the text reflects or refers to a gender stereotype (where 1 is the maximum and 0 is the minimum GS degree);
- **GS category**: the category to which the gender stereotype (if present) belongs to.

GS Value Annotation. The overall annotation procedure consists of two steps:

1. Each annotator manually assigns, for each short text, a binary label *yes*/*no* indicating whether or not the text reflects or refers to a GS;
2. The GS value is computed by combining the four individual annotations.

Although the dataset was annotated by four trained annotators, the inherent subjectivity of the task inevitably introduced a certain level of disagreement. Following the perspectivist approach we opted for merging all annotations into a numerical GS value, rather than selecting a binary label obtained through annotation aggregation on the basis of majority voting [10]. This choice aligns with recent findings which indicate that leveraging disagreement is more convenient than reliably trying to eliminate it (Basile et al., 2023; Muscato et al., 2024).

As shown in Table 1, the underlying assumption is that full inter-annotator agreement (IAA) corresponds to the endpoints of the continuum: if all four annotators agree that there is no GS, the resulting GS value is 0; if all four annotators agree that there is a GS indeed, then the resulting GS value is 1. On the other hand, disagreement between the annotators in the selection of the binary label is supposed to indicate intermediate GS values, such as 0.25 (three *no* labels and one *yes* label), 0.5 (two *yes* labels and two *no* labels), and 0.75 (three *yes* labels and one *no* label).

GS Value annotation and the strictly dependent Gender Stereotype Detection task was the main focus of the above-mentioned GSI:detect shared task.

GS Category Annotation. If a text is marked by any annotator as containing or referring to a GS, it is also assigned to one of the six GS categories,

team consists of three women and one man: two are between 20 and 30 years old, one between 30 and 40, and one above 40. In terms of education, one holds a PhD in linguistics and three hold a master's degree in either linguistics, foreign language studies or computational linguistics.

[8]The annotation guidelines are available for download at this link.

[9]Which includes the study of existing literature on GSs (see 2.1), preliminary annotation of a subset of the GSI:detect dataset and the discussion of disagreement.

[10]The distributed dataset contains both the non-aggregated annotations by the four annotators, and the merged numerical GS value.

according to the classification described in the annotation guidelines and summarised as follows (examples are provided in Table. 3):

- ROLE: social and cultural expectations about what women and men should do and about how they should be;
- PERSONALITY: emotional and behavioural traits assigned to men and women based on their gender;
- COMPETENCE: generalized judgments of a person's abilities based on their gender;
- PHYSICAL: expectations about the physical appearance of men and (especially) women, and all aspects of personal care in general;
- SEXUAL: attitude and behaviour that men and women should have regarding sexuality;
- RELATIONAL: the way in which women and men should behave in interpersonal/sentimental relations.

To the best of our knowledge, this is the first attempt to provide a systematic taxonomy to classify gender stereotypes into semantic categories; it is not intended to be fully exhaustive, however, as it is derived from an abstraction over the direct observation of the examples contained in our dataset.

Given the more explorative nature of this level of annotation, its derived Gender Stereotype Classification task was presented at Evalita 2006 as a pilot subtask. Accordingly, we opted for a more traditional approach where a category is selected based on majority voting in case of disagreement between the annotators, resorting to a super-judge in case of ties (this however amounted to only 60 out of 1,010 entries, approximately 6%). For the same reason, we also adopted some simplifications in the annotation guidelines, such as to always select only one category.

Inter-Annotator Agreement (IAA) The total IAA between the four annotators on the choice of the *yes* or *no* label is 0.613 (Fleiss' k), which is a moderate agreement. Figure 1 provides an overview of the inter-annotator agreement among the four annotators, visualized through pairwise Cohen's k values for both the GS value and GS category annotations. The two heatmaps respectively illustrate the agreement patterns for the numerical and categorical scoring schemes.

As shown in Figure 1a, A2 and A3 have the highest agreement (0.679), A1 and A4 have the lowest agreement (0.486). A1 has the lowest average pairwise IAA (0.579), followed by A4 (0.583), A3 (0.631) and A2 (0.659). Regarding the GS category annotation (Figure 1b), the four experts scored another moderate agreement, with a IAA of 0.611 (Fleiss' k). Inspecting again the pairwise IAA values (Cohen's k), A2 and A3 have again the highest agreement (0.684), while A1 and A4 have the lowest agreement (0.509). In this case, however, A4 is the one with the lowest average pairwise IAA (0.573), followed by A1 (0.587), A3 (0.622) and A2 (0.657).

	Dev set	Test set	Total
WITH CONTEXT texts	82	323	405
NO CONTEXT texts	118	487	605
Total	200	810	1,010

Table 4: Dataset's size and split.

	Tokens	Items	Av. length
Texts only	33,673	1,010	33.3 tok.
Contexts only	18,445	405	45.5 tok.
Whole dataset	52,118	1,010	51.6 tok.

Table 5: Dataset's size in details.

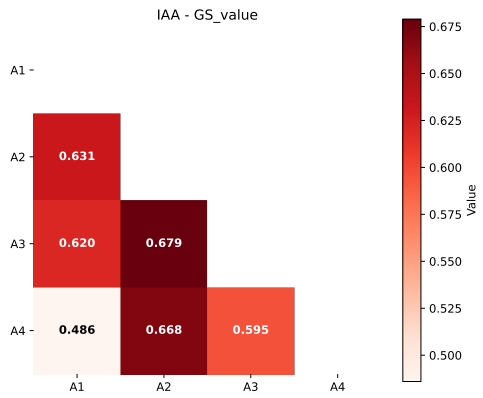
3.3. Statistics

As the dataset has been used in the context of a shared task, its 1,010 texts have been split as follows (Table 4): 20% of the dataset is used as development data (*dev set*), while the remaining 80% of the dataset is used for the official evaluation and ranking of participant systems (*test set*). The rationale behind this proportion is to balance the need for sufficient data for model tuning with the availability of a larger and more representative test set for evaluation purposes. Both the dev and the test set have a ratio of around 58% texts with context and 41% texts without context.

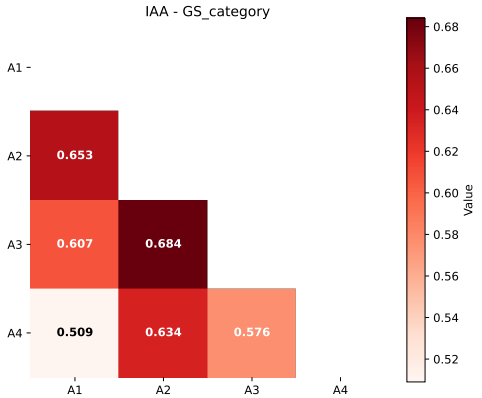
Table 5 reports detailed information about the size of the dataset in terms of tokens. The token count was computed using the Italian rule-based tokenizer included in the *spaCy* library[11] (version 3.8.7) as part of the *it_core_news_sm* linguistic model. The average length of the GSI:detect texts is 33.3 tokens if we consider pure original text and 51.6 in we include the added contextual information. The context metadata (added to 405 texts) consist on average of 45.5 tokens.

When defining the development and test splits, particular attention was paid to maintaining a balanced distribution of examples across both sets. As shown in Table 6, the overall proportion (i.e., the percentage) of items assigned to each GS value in the *Total %* column is approximately reflected in the relative composition of both the dev set and test set, when considering the ratio between the number of items per GS value and the total size of

[11] https://spacy.io



(a) IAA – GS_value



(b) IAA – GS_category

Figure 1: Comparison of IAA across the two annotation tasks, i.e. (a) GS value and (b) GS category annotation. Each heatmap visualizes pairwise agreement between annotators.

GS value	**Dev set**	**Test set**	**Total**	**Total%**
0	60	242	302	29.90%
0.25	25	84	109	10.79%
0.50	27	85	112	11.09%
0.75	25	105	130	12.87%
1	63	294	357	35.35%
	200	810	1,010	

Table 6: Dataset distribution by GS value.

Category	**Dev set**	**Test set**	**Total**	**Total%**
Role	30	107	137	13.56 %
Personality	29	108	137	13.56%
Competence	34	120	154	15.25%
Physical	20	90	110	10.89%
Sexual	14	72	86	8.52%
Relational	13	71	84	8.32%
GS value = 0	60	242	302	29.90%
	200	810	1,010	

Table 7: Dataset distribution by GS category.

the respective subset. This indicates that the split preserves the original distribution of GS values, ensuring a consistent representation of different degrees of stereotypicality across both subsets. A similar balance is maintained with the GS category (see Table 7), where the relative proportions of the six stereotype dimensions remain comparable between the dev and test sets.

This careful selection confirms that both subsets are representative of the overall dataset, providing a reliable basis for model tuning and evaluation, while avoiding unwanted biases in category distribution.

The current release of the dataset maintains the split between dev and test set. It also preserves the non-aggregated judgments of all annotators, thus allowing system to both learn from disagreement and test their predictions against the GS value emerging from all judgments.

4. Experiments

In order to provide reference baselines for future research, we conduct a series of experiments on the GSI:detect dataset. The aim of these experiments is not to achieve state-of-the-art performance, but rather to establish an initial benchmark showing how a predefined set of LLMs perform on the two tasks for which the dataset was designed: (i) the GS value detection and (ii) the GS category prediction.

Experimental setting. All models are evaluated in a zero-shot setting, without any fine-tuning, hyperparameter optimisation, or even task-specific adaptations. To ensure a fair and consistent evaluation, a minimal prompt engineering phase was carried out on the dev set: four researchers independently proposed one prompt each for both tasks (i.e., GS value detection and GS category prediction), and the best-performing prompt was selected based on preliminary experiments conducted with the *GPT5-nano model* (OpenAI, 2025). The final evaluation was then performed on the test set using the three selected models.

This experimental design allows us to provide transparent and reproducible baselines that can serve as a point of comparison for future studies exploring more advanced or fine-tuned approaches.

Tasks. We prompted the models instructing them to output a GS value between or equal to 0 and 1 (*GS value detection* task) and a GS category label for each instance in the dataset (*GS category prediction* task).

4.1. Models

To establish reference baselines on the GSI:detect dataset, we evaluate three LLMs differing in architecture, size, and accessibility, thereby including both closed-source models and open-source one.

GPT-5. (OpenAI, 2025) We use the *gpt-5-nano-2025-08-07* variant, a lightweight OpenAI model optimized for classification and reasoning tasks, employed in its text-only configuration.
GPT-4o. (OpenAI, 2024b,a) A proprietary multimodal model from the GPT-4 family, widely recognized for its strong (multimodal) reasoning and language understanding skills and used exclusively with its linguistic component.
Qwen3-14B. (Qwen-Team, 2025) An open-source transformer model available on the Hugging Face Hub. We employ the 14B-parameter variant.

Note that for all the models we tested both a *Split Prompt* configuration – where each task was addressed separately[12] – and a *Unified Prompt* configuration, where both tasks were handled within a single prompt.

4.2. Metrics

We adopted task-specific metrics to ensure a fair and accurate evaluation of model performance.

GS value Detection Task. The comparison of the models' performance on the GS value detection task is based on a normalized score derived from the Mean Squared Error (MSE), as reported in Table 8. MSE measures the average squared difference between the predictions and gold values, with larger errors more penalised, and is normalized by the variance of the data distribution to obtain NMSE. The final metric is defined as $\frac{1}{1+\text{NMSE}}$, bounded in $[0, 1]$, and is adopted for model comparison and ranking due to its improved interpretability. We compute also the Concordance Correlation Coefficient (CCC) score, that shows the agreement between predicted values and gold values, and how consistently the predictions align with gold values. CCC scores range from -1 (perfect disagreement) to $+1$ (perfect agreement), with higher CCC values indicating better model performance.
GS category Prediction Task. We assess the models' performance on the GS category prediction task using Macro F1 (not accounting for class imbalance) and Micro F1 scores (accounting for class imbalance), as seen in Table 9. A breakdown of models' per-category performance is provided in Figure 2, where higher F1 indicates better performance.

[12] i.e. with different prompts and different API calls.

For both tasks, models were also evaluated against several baselines.

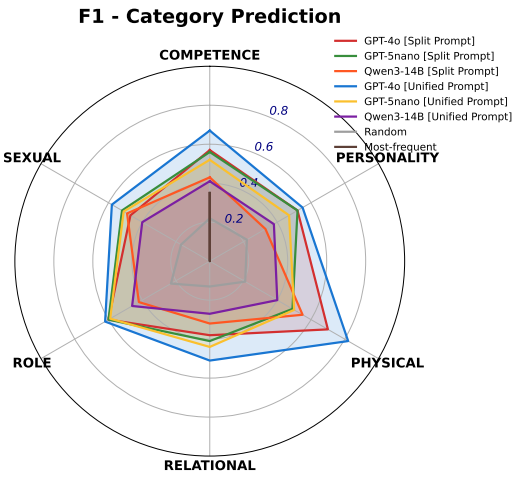


Figure 2: F1 score per Category

5. Results and Discussion

Table 8 presents results for the GS value detection task, clearly showing GPT-5-nano achieving the best overall performance in the split-prompt setting, with the highest normalized score (0.61) and CCC (0.60). GPT-4o exhibits a comparable behavior in the unified prompt setting with respect to the error-based metric, although with lower agreement. While the unified prompt setting does not bring substantial improvements for GPT-5-nano , it appears to benefit GPT-4o, exhibiting an increase of 0.10 points. Qwen3-14B does not achieve competitive performance overall. While in the unified prompt setting it reaches 0.54, slightly surpassing GPT-4o in the split configuration, it remains the weakest model in the split-prompt setting (0.49), where it achieves the lowest score among all evaluated models. Overall, all evaluated models consistently outperform both random and worst-case baseline. The same holds when considering the *0.5 baseline*[13], which assigns the midpoint value to all instances. However, Qwen3-14B in the split-prompt configuration does not surpass this baseline (0.49), making it the weakest model overall across both prompt configurations.

GS category prediction task results are described in Table 9. GPT-4o achieves the best Macro F1 score in the split-prompt configuration (0.54), and a clear improvement in Micro F1 (0.63) within the unified one. This indicates better handling of class imbalance. By contrast, Qwen3-14B consistently performs poorly also in this task,

[13] Results of this baseline perfectly correspond to 0.5 due to the balanced distribution of the dataset and the fact that GS values lie in the interval $[0, 1]$.

Setting	Model	1/(1+NMSE) ↑	CCC ↑
Baselines	Random	0.40	0.00
	Worst-case	0.18	-0.87
	0.5 value	0.50	0.00
Split Prompt	GPT-4o	0.51	0.43
	GPT-5nano	**0.61**	**0.60**
	Qwen3-14B	0.49	0.38
Unified Prompt	GPT-4o	0.61	0.55
	GPT-5nano	0.59	0.57
	Qwen3-14B	0.54	0.46

Table 8: GS value detection results. Best performance in bold.

Setting	Model	Macro F1 ↑	Micro F1 ↑
Baselines	Random	0.19	0.20
	Most-frequent	0.06	0.21
Split Prompt	GPT-4o	**0.54**	0.54
	GPT-5nano	0.52	0.53
	Qwen3-14B	0.38	0.40
Unified Prompt	GPT-4o	0.54	**0.63**
	GPT-5nano	0.50	0.52
	Qwen3-14B	0.39	0.40

Table 9: GS category prediction results for the evaluated models in a zero-shot setting. Best performance in bold.

achieving the lowest results among all evaluated models in both configurations (0.38 Macro F1 and 0.40 Micro F1 in the split setting), with only marginal differences between split and unified prompts (0.38 vs 0.39 Macro F1). Notably, all models remain substantially above both baselines, namely the random assignment and the most-frequent strategy (which always predicts the most common class, i.e., *competence*). Finally, Figure 2 (F1 scores per categories) shows that most models perform similarly across categories. However, the GPT-4o model not only outperforms the others, especially in the Unified Prompt setting, but shows a particular expertise in the PHYSICAL GS dimension.

Thus, overall, GPT-based models outperform the open model, with GPT-5-nano and GPT-4o respectively excelling in task 1 and 2. Such results highlight several interesting trends across the two tasks: GPT-5-nano is particularly aligned to humans capturing the gradient and continuous nature of stereotypicality in language, in contrast, GPT-4o performs better in the categorical classification of stereotypes. Moreover, the poor performance of Qwen3-14B across both tasks may be attributed to its smaller scale or to a training set of data that lacks the right amount of exposure to Italian data and gender-related social phenomena, suggesting that such performance may be also influenced by the ability of the model to capture language- and culture-specific patterns.

Finally, while almost all models outperform the random, worst-case and 0.5 baselines, the improvements over them remain relatively limited. In particular, when considering the 0.5 baseline, even the best-performing model (GPT-5-nano in the split setting) exceeds it by only 0.11 points, with GPT-4o showing a marginal gain of 0.01 points in the same configuration and a larger improvement (0.11) in the unified setting, and Qwen3-14B improving by only 0.04 points in its best configuration. Combined with the modest improvement over the random reference (~20 points for GPT-based models and ~10 for Qwen3-14B), this suggests that GS value prediction and, more generally, the assignment of reliable continuous scores remains a non-trivial task even for state-of-the-art systems.

6. Conclusions and Future Work

We presented GSI:detect, a new Italian resource for studying gender stereotypes in Italian short texts. The dataset introduces several innovations: (i) it includes cases of Gender Stereotype in naturally occurring contexts that goes beyond hate speech, (ii) it applies a perspectivist annotation that values disagreement, and (iii) it proposes for the first time a fine-grained taxonomy of gender stereotype categories. Experiments with LLMs show that the closed-source models considered in this study align more closely with human judgments, both in detecting the degree of stereotypicality and in categorical classification. However, given that only one relatively small open-source model (i.e, *Qwen3-14B*) was evaluated, and that it may have had more limited exposure to Italian data and its culturally grounded phenomena during training or instruction tuning, this result should be interpreted with caution and not generalized to all open-source systems, but rather viewed as an initial exploratory comparison in this direction. In addition, the relatively limited margin over the random baseline suggests that modeling the graded nature of stereotypicality remains challenging even for state-of-the-art (closed-source) systems.

Given the good response obtained by the participants in the shared task we organised based on GSI:detect, as future work we plan to consolidate our work by releasing a new version of our dataset applying the perspectivist approach to the classification of GSs. In addition to this, future work will focus on the sociolinguistic profiling of LLMs to better understand how their behaviour towards this topic aligns – or diverges – from human perspectives, with special attention to distinct demographic groups.

7. Acknowledgements

This paper is the result of a collaboration between all authors. Specifically, Davide Testa wrote Section 1, Section 5 and Section 6; Sofia Brenna wrote Section 4 together with Davide Testa; Manuela Speranza wrote Section 3 together with Gloria Comandini who also wrote Section 2.
This work has been carried out while Davide Testa was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with Fondazione Bruno Kessler (FBK). Bernardo Magnini was supported by the PNRR MUR project PE0000013-FAIR (Spoke 2).

8. Ethical considerations

The GSI:detect dataset includes naturally occurring texts that may contain sexist or offensive content, collected solely for research purposes to study gender stereotypes. All data come from public sources and were anonymized to protect privacy. The views and opinions expressed in the dataset do not necessarily reflect those of the authors, and in some instances neither those of the informants, as the authors may have selected only an extract from the original texts.

9. Bibliographical References

L. Arcuri and M.R. Cadinu. 1998. *Gli Stereotipi. Dinamiche psicologiche e contesto delle relazioni sociali*. Il Mulino, Bologna.

V. Basile, F. Cabitza, and A. Campagner. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Toward a Perspectivist Turn in Ground Truthing for Predictive Computing*, pages 6860–6868, Washington DC. Association for the Advancement of Artificial Intelligence.

Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of preaggregation on the evaluation of highly subjective tasks. In *Proceedings of the AIxIA 2020 Discussion Papers Workshop*, pages 31–40. CEUR Workshop Proceedings.

Irene Biemmi. 2020. *Educazione sessista. Stereotipi di genere nei libri delle elementari*. Rosenberg & Sellier, Torino.

Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theorygrounded 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*, page 1276–1295, Seattle. Association for Computational Linguistics.

S. Cavagnoli and F. Dragotto. 2021. *Sessismo*. Mondadori, Milano.

Amanda Cercas Curry, Gavin Abercrombie, and Zeerak Talat. 2024. Subjective isms? on the danger of conflating hate and offence in abusive language detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, page 275–282. Association for Computational Linguistics.

Gloria Comandini, Manuela Speranza, Sofia Brenna, Davide Testa, Stefania Cavagnoli, and Bernardo Magnini. 2026. Gsi:detect at evalita 2026: Overview of the task on detecting gender stereotypes in italian. In *Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026)*, Bari, Italy. CEUR.org.

E. Fersini, D. Nozza, and P. Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *EVALITA Evaluation of NLP and Speech Tools for Italian. Proceedings of the Final Workshop*, pages 59–66, Torino. Accademia University Press.

C. Harrison and K. D. Tanner. 2018. Language matters: Considering microaggressions in science. *CBE - Life Sciences Education*, 17:1–8.

Lynne M. Jackson. 2011. The psychology of prejudice: From attitudes to social action. *American Psychological Association*.

H. Kirk, W. Yin, B. Vidgen, and P. Röttger. 2023. Semeval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, page 2193–2210, Stroudsbourg. Association for Computational Linguistics.

Manfred Klenner, Anne G¨ohring, and Michael Amsler. 2020. Harmonization sometimes harms. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*. CEUR Workshop Proceedings.

Marco Madeddu, Simona Frenda, Mirko Lai, Viviana Patti, and Valerio Basile. 2023. Disaggreghate it corpus: A disaggregated italian dataset of hate speech. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 243–250. CEUR Workshop Proceedings.

M. Mitchell, G. Attanasio, I. Baldini, M. Clinciu, J. Clive, P. Delobelle, M. Dey, S. Hamilton, T. Dill, J. Doughman, R. Dutt, A. Ghosh, J. Zosa Forde, C. Holtermann, L.A. Kaffee, T. Laud, A. Lauscher, R.L. Lopez-Davila, M. Masoud, N. Nangia, A. Ovalle, G. Pistilli, D. Radev, B. Savoldi, V. Raheja, J. Qin, E. Ploeger, A. Subramonian, K. Dhole, K. Sun, A. Djanibekov, J. Mansurov, K. Yin, E. Villa Cueva, S. Mukherjee, J. Huang, X. Shen, J. Gala, H. Al-Ali, T. Djanibekov, N. Mukhituly, S. Nie, S. Sharma, K. Stanczak, E. Szczechla, T. Timponi Torrent, D. Tunuguntla, M. Viridiano, O. Van Der Wal, A. Yakefu, A. Névéol, M. Zhang, S. Zink, and Z. Talat. 2025. Shades: Towards a multilingual assessment of stereotypes in 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 11995–12041, Kerrville. Association for Computational Linguistics.

Benedetta Muscato, Chandana Sree Mala, Marta Marchiori Manerba, Gizem Gezici, and Fosca Giannotti. 2024. An overview of recent approaches to enable diversity in large language models through aligning with human perspectives. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, page 49–55. ELRA and ICCL.

OpenAI. 2024a. Gpt-4 technical report.

OpenAI. 2024b. Gpt-4o system card.

OpenAI. 2025. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf. Accessed: October 2025.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, page 1246–1266, New York. Association for Computing Machinery.

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.

G. Peruzzi, V. Bernardini, and R. Lombardi. 2019a. Le questioni di genere dei giovani. un'indagine sulle percezioni e le esperienze di studenti e studentesse universitari. In G. Peruzzi, V. Bernardini, R. Lombardi, C. Rinaldi, M. Bacio, L. Bainotti, and G. Viggiani, editors, *Il bias del gender*, pages 13–50. Durango, Andria.

G. Peruzzi, V. Bernardini, R. Lombardi, C. Rinaldi, M. Bacio, L. Bainotti, and G. Viggiani. 2019b. *Il bias del gender*. Durango, Andria.

L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, and P. Rosso. 2023. Overview of exist 2023: sexism identification in social networks. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*, pages 593–599, Cham. Springer.

Qwen-Team. 2025. Qwen3 technical report.

G. Rizos and B.J. Schuller. 2020. Average jane, where art thou? – recent avenues in efficient machine learning under subjectivity uncertainty. In M.J. Lesot, S. Vieira, M.Z. Reformat, J.P. Carvalho, A. Wilbik, B. Bouchon-Meunier, and R.R. Yager, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 42–55. Springer, Cham.

M. Sanguinetti, F. Poletto, C. Bosco, V. Patti, and Stranisci M. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2798–2805, Paris. European Language Resources Association (ELRA).

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

D.W. Sue. 2010. *Microaggressions in everyday life. Race, gender and sexual orientation*. John Wiley & Sons, Hoboken.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 769–779, Seattle. Association for Computational Linguistics.

M. Wojatzki, T. Horsmann, D. Gold, and T. Zesch. 2018. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgement. In *Proceedings of the 14th conference on Natural Language Processing (KONVENS 2018)*, pages 110–120.

Quantifying and Predicting Disagreement in Graded Human Ratings

Leixin Zhang Çağrı Çöltekin
Universität Tübingen, Germany
leixin.zhang@uni-tuebingen.de cagri.coeltekin@uni-tuebingen.de

Abstract

It is increasingly recognized that humans do not always agree, and disagreement is inherent in many annotation tasks. However, not all items in a given task elicit the same level of opinion divergence. In this paper, we study the extent to which item-level annotation variation and variation structure can be captured from text features, focusing on inappropriate language detection, including offensive language, hate speech, and toxic language detection. We model annotation variation to assess whether the degree of annotation divergence can be predicted from item-level textual features. We also propose the Opposition Index, a metric that quantifies the extent of opposing stances among annotators based on their Likert ratings.

Keywords: annotation variation, graded human ratings, opposing stances, human disagreement

1. Introduction

Recent work has demonstrated that many natural language processing (NLP) datasets and tasks exhibit inherent annotation variation (Plank, 2022; Sorensen et al., 2024). This variation occurs across multiple linguistic domains and task types. In *syntax*, it is observed in annotation tasks such as part-of-speech tagging (Plank et al., 2014; Zeman, 2010); in *semantics*, such as semantic similarity (Wang et al., 2023) and natural language inference (Huang and Yang, 2023; Jiang and de Marneffe, 2022; Pavlick and Kwiatkowski, 2019; Liu et al., 2023; Zhang and de Marneffe, 2021); in *pragmatics*, including irony detection (Casola et al., 2024) and dialogue act annotation (Verdonik, 2023); and in other *socially-relevant NLP tasks*, such as hate speech detection (Sang and Stanton, 2022; MacAvaney et al., 2019), offensive language (Kocoń et al., 2021; Davani et al., 2024), and toxic content detection (Kumar et al., 2021).

It is increasingly acknowledged that human annotators do not always make identical decisions or hold the same opinions on many NLP tasks (Plank, 2022; Uma et al., 2021; Basile et al., 2021). However, the annotation pattern is not consistent across all items in the same task: some cases are clear-cut and have near-perfect agreement from multiple annotators, while others can be more ambiguous, resulting in variance in annotation patterns. Such annotation variation can arise from item ambiguity or vagueness (e.g., insufficient context), language complexity (e.g., use of slang or jargon), or annotators' personal beliefs, values, expertise, or personality (Sap et al., 2020, 2022).

Reflected in Likert rating distributions, annotations for some items show sharper peaked distributions, indicating strong consensus among annotators, while others may display flatter or multi-modal distributions, reflecting interpretation variability or the presence of opposing opinions among annotators.

Estimating which items are likely to elicit disagreement has important practical and theoretical implications. In *annotation practice*, identifying disagreement-prone items allows researchers to optimize annotation workflows by prioritizing difficult or perspective-divergent cases. For instance, socially controversial items such as potentially offensive or politically charged content often require a larger number of annotators to capture the diversity of latent perspectives, whereas less controversial items may require fewer annotations.

From a *linguistic perspective*, analyzing uncertainty patterns allows researchers to uncover the latent factors underlying annotation uncertainty, such as detecting cases that lack sufficient context (Sandri et al., 2023), and provide insights into human perception of complex language phenomena, such as irony, sarcasm, or figurative language. Studying perspective disagreement can further reveal culturally dependent interpretations (e.g., Western vs. non-Western perspectives (Sap et al., 2022; Huang and Yang, 2023; Larimore et al., 2021), liberal vs. conservative viewpoints (Luo et al., 2020)) or conflicts in judgment. In *other decision-making domains*, including legal, political, and medical decision-making, annotation variation may reflect conflicting interests and opposing perspectives between different parties (e.g., employers vs. employ-

Proceedings of NLPerspectives @ LREC 2026, pages 33–43

ees, producers vs. consumers) (Angouri, 2012). Automatically identifying disagreement-prone items can help flag conflicting cases, prioritize expert review, and improve decision fairness (Patel et al., 2018).

In this work, we investigate whether the item-level annotation patterns can be inferred solely from item features, and we mainly focus on inappropriate language detection tasks in this work, including hate speech, offensive, and toxic language classification. These tasks have been extensively studied and are known to exhibit substantial annotator disagreement (Sang and Stanton, 2022), and there is a lack of universally accepted standards or definitions. For example, definitions of hate speech vary across research objectives (Talat and Hovy, 2016), legal frameworks (European Commission, 2016), and platform policies. It is often impractical to specify every possible case in annotation guidelines, particularly in crowdsourced settings where annotators are not formally trained. Thus, hate speech annotation often relies on annotators' perceptions, linguistic intuition, and individual understanding of what constitutes hate speech. In the era of large language models (LLMs), these challenges become even more critical (Weidinger et al., 2021). Given the rapidly growing user base of LLM-powered systems, detecting toxic and inappropriate language is essential for mitigating risks, preventing the amplification of harmful content, and ensuring safer user interactions.

To more faithfully capture the nuances of human judgment and annotation distributions, we use datasets annotated with Likert-scale ratings instead of discrete binary labels in this work. We also tailor the training objective by employing loss functions designed for ordinal data rather than simple discrete classes, including the Earth Mover's Distance (and its variant) and cumulative cross-entropy. This work focuses on estimating two aspects of human annotation variation:

- **Annotation variance:** whether the degree of dispersion in annotator responses can be inferred from item features.
- **Opposing stances:** whether the conflicting stances among annotators can be effectively modeled and predicted.

By modeling item-level annotation variance and stance opposition, our work takes an initial step toward characterizing patterns or structures of human annotation variation. We hope this study will inspire further research on annotation uncertainty and perspective-aware NLP systems.

2. Related Work and Positioning

Existing literature on annotation variation can be broadly divided into two main streams. The first stream focuses on analyzing human annotation and human interpretation variation (Hong et al., 2025; Jiang and de Marneffe, 2022), investigating types or causes of variation across annotators (Xu et al., 2023), disentangling noise from genuine disagreement (Weber-Genzel et al., 2024), and analyzing cultural background influence (Huang and Yang, 2023). The second stream focuses on modeling human annotation variation with machine learning approaches (Uma et al., 2021; Mostafazadeh Davani et al., 2022; Zhou et al., 2022). The methods include soft-label training (Uma et al., 2021; Fornaciari et al., 2021), incorporating socio-demographic features for group perspective simulation (Gordon et al., 2022), multi-task learning (Mostafazadeh Davani et al., 2022) with each task corresponding to a specific annotator, and using an annotator embedding layer (Mokhberian et al., 2024) to learn annotator-specific labels.

Work that explicitly predicts or infers disagreement from text features remains relatively scarce. In this direction, Zhang and de Marneffe (2021) aim to tease apart agreed and disagreed items in natural language inference (NLI) and propose an ensemble approach by integrating three specialized models trained to predict three labels: entailment, neutral, or contradiction. In subjective tasks such as hate speech detection, Wan et al. (2023) model disagreement by directly predicting whether or not annotation variation exists for an item as a binary classification problem, and a regression problem by predicting annotation variation with the value $1 - P_{majority}$, the proportion of annotations that do not fall into the majority label. Baumler et al. (2023) and van der Meer et al. (2024) estimate the uncertainty of human annotations and incorporate it into an active learning framework.

Despite these advances, prior work has largely treated opinion divergence as a categorical prediction problem (e.g., three-way NLI labels or binary hate speech decisions) and has not examined the degree of annotation uncertainty and whether the full structure of fine-grained annotation distributions can be recovered from item-level signals (Zhang, 2025). To address this gap, we perform a detailed analysis to model the full distribution of Likert-scale ratings and examine its effectiveness in inferring annotators' opinion divergence with item textual features.

3. Rating Variation Prediction

To model opinion divergence, including both variance and opposing views, we use datasets with Likert-scale ratings. Unlike discrete labels, these graded ratings capture fine-grained human perception differences, reflect gradations in perceptions of inappropriate language, enabling analysis of multimodal and polarized patterns in distribution. We aim to test whether the structure of annotation variation across items can be captured from textual features. Specifically, we examine the magnitude of opinion divergence in Section 3.1 and the presence of opposing stances among annotators in Section 3.2.

3.1. Inferring Annotation Variance

For discrete classes, the entropy of annotator labels is commonly used to quantify uncertainty. For Likert ratings, where the distance between ratings is meaningful, we treat them as equally spaced values. For each item i, the degree of annotation variation is computed as the unbiased variance of the ratings from N_i annotators:

$$\sigma_i^2 = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (r_{ij} - \bar{r}_i)^2, \quad (1)$$

where r_{ij} is the rating of annotator j on item i, and $\bar{r}_i$ is the mean rating for that item. Perfect agreement corresponds to $\sigma_i^2 = 0$.

To estimate annotation variation, we consider two approaches:

1. **Direct variance prediction**: we train a regression model to predict the unbiased variance σ_i^2 directly from item-level features.
2. **Distribution-based prediction**: we predict the full Likert rating distribution $\mathbf{p}_i = [p_{i1}, \ldots, p_{iK}]$ for item i, then compute the variance from the predicted distribution:

$$\hat{\sigma}_i^2 = \sum_{k=1}^{K} p_{ik}(k - \hat{\mu}_i)^2, \quad \hat{\mu}_i = \sum_{k=1}^{K} p_{ik} \cdot k. \quad (2)$$

To evaluate the performance of prediction, we mainly measure with the following metrics: (1) *Mean Squared Error (MSE)* between predicted and true variance, (2) *Spearman's rank correlation r* between predicted variance and true variance across items and (3) F_1 score of whether or not opinion divergence or rating difference is in the annotation of an item among annotators.

	Comments	Annotations
1	"Mr #Trump will be loving today. As it is the one day of the year when #FakeNews is acceptable. #aprilfools "	[0, 0, 0, 2, 2, 2]
2	"Nigga at da end of the day we all would be gone, or somewhere else. and speakin about it is not gonna fucking matter! "	[0, 0, 2, 2, 2]
3	"I hate when guys call their girls bitches and hoes. That's your girl. You respect her. "	[0, 0, 0, 2, 2]

Table 1: Examples of opposing stances in the offensive dataset (Sap et al., 2020). Three-ordinal ratings are used for labels, with scores from 0 to 2, with 0 as *not offensive* and 2 as *offensive*.

3.2. Identifying Opposing Opinions

Beyond variance, it is also crucial to assess distribution structures (Akhtar et al., 2021; Van der Eijk, 2001) and whether genuine opposing opinions exist for an item. Some cases with divergent judgments on offensive classification are shown in Table 1. The annotations suggest that disagreement can arise from different interpretations of what constitutes offensive content. In Example 1, disagreement emerges in a case that involves political satire. Some annotators interpret mocking a political figure as offensive, possibly due to perceived contempt or disrespect, whereas others regard it as legitimate political expression or humor rather than offensive content per se. In some cases, annotators label a comment as offensive due to the presence of offensive terms, even when the overall intent of the comment is not to insult. For example, in Example 3 of Table 1, a comment condemning derogatory terms toward women may still be marked offensive by some because it quotes them, while others focus on the critical intent and label it non-offensive. These examples show that annotation variation is often driven by differences in how annotators weigh lexical content, speaker intent, and contextual meaning about respect and harm.

Reflected in the Likert distribution, this manifests as bimodal patterns rather than a single Gaussian mode. We propose a metric to quantify opposing stances, referred to as **opposition index**.[1] Let the

[1]Traditional bimodality measures, such as the Bimodality Coefficient (BC), or mixture-model-based modality tests, are typically designed for continuous dis-

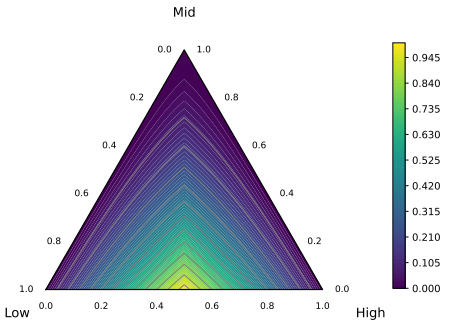


Figure 1: Opposition Index Illustration

predicted or observed distribution for an item be divided into three segments: the low, mid, and high ratings of the Likert scale, with probabilities denoted as $P_{\text{low}}, P_{\text{mid}},$ and P_{high}. P_{low} corresponds to the proportion of annotators giving the lowest ratings (e.g., not toxic). P_{high} corresponds to the proportion giving the highest ratings (e.g., extremely toxic); and P_{mid} corresponds to the proportion in the middle of the scale. We define the opposition stance index as:

$$\text{Index}_i = 2 \cdot \min(P_{\text{low}}, P_{\text{high}}) \cdot (1 - P_{\text{mid}}) \quad (3)$$

The final index value ranges from 0 to 1, with 1 indicating maximal polarization (half of the annotators select the low end and half the high end, with no intermediate ratings), with a value of 0 indicating consensus, reflected in a unimodal distribution centered in the middle or skewed toward either end of the scale, with no annotations spanning both extremes. Figure 1 illustrates how the index behaves: when $P_{\text{low}} = 0.5$ and $P_{\text{high}} = 0.5$ with negligible P_{mid}, the index reaches its maximum, reflecting clear opposing stances among annotators.

4. Objective Functions for Likert Distribution Prediction

To model both annotation variation and opposing opinions discussed in the previous section, we infer the full Likert rating distribution for each item, apart from predicting a single summary statistic of variance. In this section, we propose objectives specifically designed for Likert-scale ratings, leveraging their ordinal structure rather than treating them as categorical labels. For an item i with K Likert categories, the target distribution is represented as:

$$\mathbf{p}_i = [p_{i1}, p_{i2}, \ldots, p_{iK}], \text{ where } 1 < 2 < \cdots < K. \quad (4)$$

where p_{ik} denotes the proportion of annotators assigning rating k to item i, and $\sum_{k=1}^{K} p_{ik} = 1$. For example, in a 5-point Likert setting (ratings range from 0 to 4), if the annotations for an item are $\{1,1,0,1,2\}$ collected from 5 annotators, the empirical distribution is represented as a vector: $\mathbf{p}_i = [0.2, 0.6, 0.2, 0.0, 0.0]$.

To train the model to predict distributions, we experiment with loss functions listed below:

Earth Mover's Distance (EMD), also known as the Wasserstein distance (Rubner et al., 2000), explicitly accounts for the ordinal distance between rating categories. EMD penalizes prediction errors proportionally to the distance between categories.

EMD with Mean Regularization We further propose a multi-task objective that combines Earth Mover's Distance with an explicit constraint on the predicted mean rating with mean squared error.[2]

Ordinal Cumulative Cross Entropy To capture the ordinal structure of Likert-scale labels, we customize cross-entropy loss to measure the distributional difference of the Likert ratings. In our approach, a K-level Likert-scale problem is transformed into $K-1$ binary decisions with positive class as $y > k$. The total loss is then computed as the sum of the losses over all $K-1$ thresholds.

Kullback–Leibler Divergence We also consider the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) as as a comparison to cumulative probability approaches and use it to quantify the dispersion from the predicted distribution to the target distribution (soft-label representation, which we refer to as KL-soft in this work).[3]

tributions with sufficiently large sample sizes. When the number of annotators per item is small (e.g., around five), these statistics become unstable and lack statistical power, making them unsuitable for reliably detecting bimodality in item-level annotation distributions.

[2]While EMD captures overall distributional shifts and respects the ordinal structure of the Likert scale, it does not directly constrain the expected rating (i.e., the mean of the distribution). Two distributions with similar cumulative shapes may still differ in central tendency. To address this, we introduce an additional mean-squared error term on the expected rating.

[3]While KL divergence is widely used for multi-class classification, the standard categorical formulation treats all class mismatches equally, ignoring the ordinal relationships among Likert ratings.

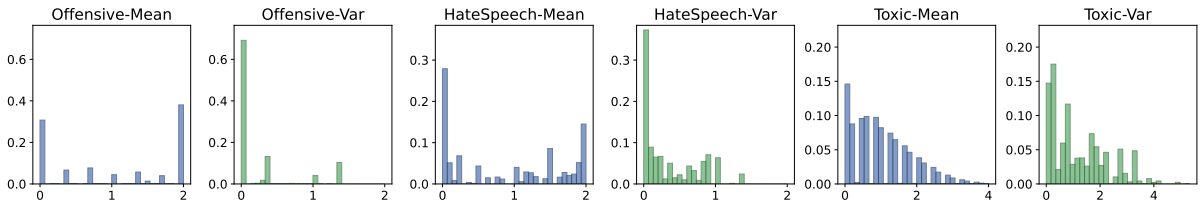


Figure 2: Summary of Dataset Statistics: Mean and Variance. The y-axis values indicate normalized density, which accounts for the relative proportion of data points in each bin.

5. Experiments and Implementation

This section introduces the datasets used for the experiments in this study and experiment implementation details.

5.1. Datasets

We conduct experiments on three subjective datasets annotated by multiple annotators. Figure 2 shows the summary of the statistics (annotation mean and variance) of three datasets.

Offensive Language. The Offensive Language dataset (Sap et al., 2020) is annotated with three categories of offensiveness: *no*, *maybe*, and *yes*. We map these categories to a three-point Likert scale (0 - 2). The dataset contains approximately 150k annotated items. To ensure reliable empirical annotation distributions, we filter out items annotated by fewer than three annotators. After filtering, the dataset comprises approximately 128.6k annotations over 35.8k unique items, with an average of 4 annotations per item.

Hate Speech. The hate speech dataset by Kennedy et al. (2020) provides graded annotations of hatefulness. Labels are provided on a three-level Likert scale, where 0 denotes non-hateful content and higher values indicate increasing hate speech severity. Items annotated by fewer than four annotators are removed, resulting in approximately 67k annotations over 5,990 unique items, and each item is roughly annotated by 11 annotators on average.

Toxicity. Annotations in the dataset (Kumar et al., 2021) follow a five-point Likert scale ranging from *not toxic* (0) to *extremely toxic* (4). The dataset contains approximately 107.6k text instances, most of which are annotated by 5 annotators. We retain items with at least five annotations and merge repetitive comments, resulting in approximately 106k items.

5.2. Implementation

Data Splits. Each dataset is randomly divided into training, validation, and test sets, following a 50%, 25%, 25% ratio. It is partitioned at the level of distinct text instances to prevent any items from appearing in multiple data splits.

Model Setting. Models are implemented in PyTorch, and text inputs are encoded using the pretrained Sentence-Transformer model `all-MiniLM-L6-v2`[4] (Reimers and Gurevych, 2019). To allow fair comparison across different prediction objectives, we keep the model architecture fixed, including input features, number of hidden layers, and layer dimensions, for all experiments. Models are trained with early stopping. Training is terminated if the validation performance does not improve for five consecutive epochs. The best-performing model on the validation set in the training history is selected for evaluation and reporting.

Baseline Setup. We use the aggregated binary distribution (where responses greater than $(K-1)/2$ for K-class Likert are treated as the positive class) as a baseline for each dataset, training with binary cross entropy loss, and compare its prediction with direct variance regression and full Likert distribution prediction.

Evaluation Protocol. For reliability, each experiment is repeated with five independent random splitting seeds, and the mean of the evaluation metrics is reported as the model performance.

Metric Computation. For the computation of the opposition index, we treat rating 0 as the low value and 2 as the high value for three-class Likert ratings, and treat ratings 0 and 1 as the low-value group and ratings 3 and 4 as the high-value group for five-class Likert ratings, representing two opposing stance camps.

[4] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

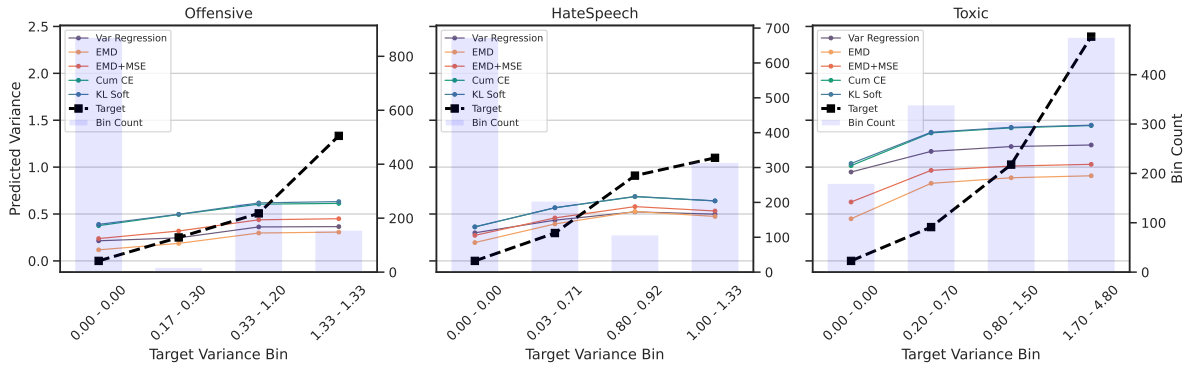


Figure 3: Item-level variance grouped by target variance bins: true versus predicted variance.

Model	Var_MSE ↓	Var_Corr ↑	Disagree_F1 ↑
Offensive			
Binary CE	0.218 ± 0.003	0.344	0.5917 ± 0.007
Var Reg	**0.195** ± 0.008	0.389	**0.6112** ± 0.018
EMD	0.227 ± 0.008	**0.393**	0.5957 ± 0.018
EMD+MSE	0.206 ± 0.003	0.386	0.6071 ± 0.018
Cum CE	0.248 ± 0.006	0.371	0.6076 ± 0.014
KL Soft	0.251 ± 0.014	0.372	0.6096 ± 0.022
Hate Speech			
Binary CE	0.275 ± 0.014	0.435	0.7275 ± 0.014
Var Reg	**0.197** ± 0.003	0.424	0.7233 ± 0.016
EMD	0.216 ± 0.019	**0.454**	0.7365 ± 0.013
EMD+MSE	**0.197** ± 0.011	0.445	0.7313 ± 0.017
Cum CE	0.204 ± 0.008	0.449	**0.7408** ± 0.016
KL Soft	0.206 ± 0.008	0.445	0.7371 ± 0.011
Toxic			
Binary CE	2.203 ± 0.021	0.290	0.9185 ± 0.001
Var Reg	**1.005** ± 0.027	**0.308**	0.9185 ± 0.007
EMD	1.179 ± 0.060	0.307	0.9186 ± 0.008
EMD+MSE	1.087 ± 0.049	0.303	0.9187 ± 0.007
Cum CE	1.056 ± 0.028	0.306	0.9191 ± 0.007
KL Soft	1.061 ± 0.012	0.298	0.9191 ± 0.007

Table 2: Comparison of models for estimating item-level annotation variance (mean ± std).

6. Results and Discussion

This section presents the experimental results for both annotation variance estimation (Section 6.1) and opposing stance prediction (Section 6.2).

6.1. Annotation Variance Estimation

Variance Prediction Models predict annotation variance values with reasonable accuracy. They achieve a variance MSE of around 0.2 on the 3-point Likert annotation tasks (Offensive and Hate Speech), and a variance MSE of approximately 1 for the 5-point Likert task (Toxic). Among all methods, directly predicting the unbiased variance using a regression model achieves the best performance. Among models that predict the Likert distribution and then compute variance, those trained with the EMD with mean regularization (EMD+MSE) achieve the second-best performance. They consistently outperform the EMD and KL-soft models.

Prediction across Annotation Variance Bins Apart from overall performance, we also analyze variance prediction across bins divided based on human-annotated variance levels (see Figure 3).

The bin-grouped analysis reveals a similar pattern across models. All variance predictions exhibit a monotonic trend: items with near-perfect agreement are assigned the lowest predicted variance scores, and items with higher empirical variance tend to receive higher predicted variance. However, the predicted variance values tend to concentrate around a middle range. The difference between low, medium, and high variance bins is attenuated. For instance, the magnitude differences are underestimated for the highest variance category. Models do not distinguish well between items with moderate and high variance. It may be due to the relatively smaller number of examples in these bins for the offensive and hate speech datasets. For the lowest variance bins, predicted variance values rarely reach exactly 0, particularly for distribution-based models. As a result, the lowest bin is not as low as the target variance.

Spearman Correlation Models show a moderate positive correlation with human annotation variance, ranging from approximately 0.3 to 0.45 across three datasets. Some models (e.g., EMD) trained to predict Likert distributions often achieve higher Spearman correlations with human annota-

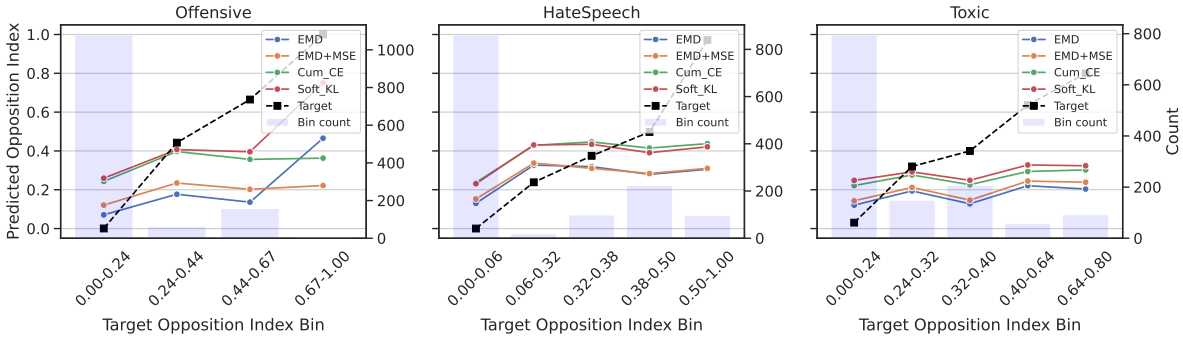


Figure 4: Opposition index values binned according to target scores.

tion variance compared to regression models that directly predict the variance. This suggests that even when variance is not explicitly supervised, distribution-based training objectives can recover the annotation variation structure across items. By contrast, directly regressing the variance can force the model to overfit the noise in the dataset. In contrast, predicting the full rating distribution allows the model to capture more structured patterns in annotator judgments. Rather than fitting a single summary statistic, the model learns the overall shape of the response distribution, providing a richer and more robust representation of opinion divergence. However, as expected, the loss function (KL-soft) that does not account for rating distances performs the worst, particularly for tasks with more Likert classes, such as the 5-rating Toxic dataset.

F1 Score F_1 score in this study measures whether an item's annotations exhibit rating differences (that is, whether the variance is greater than 0). Across all three datasets, the F1 scores from different models are generally very similar. On the Offensive dataset, the Variance Regression model achieves the highest F1 score (0.611 ± 0.018), slightly outperforming the other models. For hate speech, distribution prediction with cumulative cross-entropy loss achieves the best performance. On the Toxic dataset, all models achieve high (~0.92) and nearly identical F1 scores, likely because the dataset is annotated using a 5-point Likert scale, and most items have variance greater than 0. The imbalance tendency toward the class of presence of annotation variation (around 85% items) may lead to most results at class 1 (the presence of annotation divergence). Overall, the models achieve strong performance in detecting variation, with F_1 scores exceeding 0.6 for nearly all models across the three datasets.

6.2. Opposing Stance Prediction

For opposing stances measured by the polarization index, we group items into intervals based on their target opposition index values and analyze the results within each interval.

Ideally, items with higher true opposition values should also receive higher predicted scores from models. However, this pattern does not fully hold. While models successfully distinguish between items with no opposition (index close to 0) and items with moderate opposition, they struggle to capture extreme polarization. For items whose true opposition index approaches 1, predicted values tend to remain in a mid-range (approximately 0.4), indicating underestimation of highly polarized cases.

Several factors may explain this phenomenon. First, as shown in Figure 4, the number of items in the highest opposition index bin is relatively small, which limits the model's ability to learn stable patterns for extreme polarization when such cases are rare in the training set.

Secondly, items with a high polarization index are more prone to noise when a few annotators deviate significantly from the majority. When annotation noise inflates the opposition index, the resulting patterns may not reflect stable item features, causing the model to regress toward the mean.

Finally, extreme opposition may partly result from other factors beyond the text itself, such as annotators' ideological differences, personal experience, or differing interpretations of the guidelines, which cannot be inferred from textual features alone. Across the three datasets, items are labeled by annotators with diverse socio-demographic backgrounds, which are not evenly distributed across items. As a result, certain influences cannot be fully inferred from item-level features alone.

7. Limitations of Current Experiments

Firstly, although this paper examines the predictability of annotation variation from textual features, it cannot be assumed that the state-of-the-art encoder model, which converts texts into embeddings, perfectly captures all textual information (Lucy and Gauthier, 2017; Zhang et al., 2024; Zhang and Çöltekin, 2024). There can be information loss during the embedding process, and some linguistic cues may not be fully represented. Secondly, the number of annotations per item is limited, which limits the reliability of opinion divergence and distributional estimates. Additionally, the observed rating distributions may be sensitive to sampling noise. The datasets used in this work are crowd-sourced. Although crowd-sourced data increases annotator diversity, it also introduces additional noise, making human opinion modeling more challenging. Finally, Likert scales are restricted to three or five categories in the datasets we experiment with. With few annotators and coarse-grained scales, the space of possible variance or distributional values becomes highly discrete. For example, when only three annotators are available, certain distribution proportions (e.g., 0.33 or 0.66) occur frequently due to combinatorial constraints rather than meaningful underlying differences. This discreteness reduces the granularity of human opinion divergence and can affect the interpretability of predicted distributions.

8. Conclusion

This study investigates the extent to which annotation variation can be inferred from item-level features alone. We explore two aspects of annotation variation: human annotation variance estimation and opposing stances prediction. Our results show that variance derived from predicted distributions achieves performance comparable to direct variance regression when appropriate loss functions, such as Earth Mover's Distance and its variant with mean regularization, are used. Beyond predicting a single summary statistic like variance, distribution-based approaches can better capture disagreement structure, such as annotators' opposing stances. To quantify this, we propose the opposition index and demonstrate its use across three datasets. These findings have practical implications for future annotation design: resources can be allocated more efficiently by assigning more annotators to items likely to exhibit opinion divergence, while reducing effort on items with clear consensus.

9. Bibliographical 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*.

Jo Angouri. 2012. Managing disagreement in problem solving meeting talk. *Journal of Pragmatics*, 44(12):1565–1579.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 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.

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.

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.

Aida Mostafazadeh Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. *arXiv preprint arXiv:2404.10857*.

European Commission. 2016. Code of conduct on countering illegal hate speech online. Accessed: 2026-03-25.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, et al. 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. Association for Computational Linguistics.

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 *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Pingjun Hong, Beiduo Chen, Siyao Peng, Marie-Catherine de Marneffe, Benjamin Roth, and Barbara Plank. 2025. Agree, disagree, explain: Decomposing human label variation in NLI through the lens of explanations. *arXiv preprint arXiv:2510.16458*.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609.

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.

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*.

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.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

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.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90. Association for Computational Linguistics.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Li Lucy and Jon Gauthier. 2017. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 76–85, Vancouver, Canada. Association for Computational Linguistics.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315. Association for Computational Linguistics.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

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.

Pinal Patel, Disha Davey, Vishal Panchal, and Parth Pathak. 2018. Annotation of a large clinical entity corpus. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2033–2042, Brussels, Belgium. 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.

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. 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121.

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.

Yisi Sang and Jeffrey Stanton. 2022. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In *International Conference on Information*, pages 425–444. Springer.

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.

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*.

Zeerak Talat and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

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.

Cees Van der Eijk. 2001. Measuring agreement in ordered rating scales. *Quality and Quantity*, 35(3):325–341.

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.

Darinka Verdonik. 2023. Annotating dialogue acts in speech data: Problematic issues and basic dialogue act categories. *International Journal of Corpus Linguistics*, 28(2):144–171.

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.

Yuxia Wang, Shimin Tao, Ning Xie, Hao Yang, Timothy Baldwin, and Karin Verspoor. 2023. Collective human opinions in semantic textual similarity. *Transactions of the Association for Computational Linguistics*, 11:997–1013.

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.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Shanshan Xu, Santosh T.y.s.s, Oana Ichim, Isabella Risini, Barbara Plank, and Matthias Grabmair. 2023. From dissonance to insights: Dissecting disagreements in rationale construction for case outcome classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9558–9576, Singapore. Association for Computational Linguistics.

Daniel Zeman. 2010. Hard problems of tagset conversion. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, pages 181–185.

Leixin Zhang. 2025. Proposal: From one-fit-all to perspective aware modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1016–1025, Vienna, Austria. Association for Computational Linguistics.

Leixin Zhang, David Burian, Vojtěch John, and Ondřej Bojar. 2024. Unveiling semantic information in sentence embeddings. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024*, pages 39–47, Torino, Italia. ELRA and ICCL.

Leixin Zhang and Çağrı Çöltekin. 2024. Tübingen-CL at SemEval-2024 task 1: Ensemble learning for semantic relatedness estimation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1019–1025, Mexico City, Mexico. 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.

HurtLens: A Perspectivist Corpus Analysis of Hurtful Language

Samuele D'Avenia, Eliana Di Palma, Marta Marchiori Manerba, Valerio Basile
Computer Science Department, University of Turin, Turin, Italy
{samuele.davenia, eliana.dipalma, marta.marchiorimanerba, valerio.basile}@unito.it

Abstract

Offensive language detection systems often rely on majority-aggregated annotations, overlooking the diversity of perspectives that shape how different communities perceive harm. In this contribution, we introduce HurtLens, a perspectivist corpus of hurtful language leveraging four disaggregated datasets which are automatically enriched through HurtLex lemmas, a multilingual resource of offensive and derogatory terms. Using mixed-effects modeling, we investigate how annotators' sociodemographic backgrounds, the presence of specific types of offensive language (through Hurtlex categories) and their interaction influence offensiveness ratings. Our analysis reveals that offensiveness ratings are influenced both by annotators' sociodemographic characteristics (particularly when considering them in intersection) and by the presence of specific types of offensive language. Additionally, we identify significant interaction effects showing that different demographic groups vary in their sensitivity to texts containing particular types of offensive language.

Keywords: offensive language, perspectivism, language resources, sociodemographic analysis

WARNING: *This paper contains examples of offensive or upsetting content.*

1. Introduction

A single word can hurt, but not everyone is offended by the same type of words in a specific context. Offensiveness is a complex phenomenon, where context and individual-subjectivity play a significant role (Kiritchenko et al., 2021).

The problem of offensive speech detection is not new in literature, neither in NLP nor in other related fields such as linguistics, social sciences, and communication sciences (Fortuna and Nunes, 2018). However, when it comes to offensive speech detection, what is mainly done is to annotate datasets based on the majority opinion of annotators, thus losing essential information such as minority views (Mostafazadeh Davani et al., 2022). A turning point in this regard has been provided by the perspectivist approach, which calls for the release of disaggregated resources and systems that learn from disagreement (Cabitza et al., 2023a).

In this work, we introduce HurtLens, a perspectivist corpus of hurtful language[1] that preserves disaggregated annotations across sociodemographic groups. This corpus enriches four existing resources under the lens of HurtLex lemmas, enabling the analysis of how different sociodemographic groups respond to different types of hurtful expressions. HurtLex is a multilingual lexicon containing offensive and hateful words, which are mapped to 17 categories indicating the semantic area of the word used to offend. For example "*wh*re*" falls in the category of words related to prostitution, while "*pig*" in that of animals (Bassignana et al., 2018).

Importantly, HurtLens includes both offensive and non-offensive uses of potentially hurtful words, reflecting the inherently contextual nature of offensiveness. Using our resource, we analyze how sociodemographic characteristics (*who*), the presence of specific types of offensiveness (*what*) and their interaction (*who reacts to what*) jointly shape perceived offensiveness. We articulate our work into three research questions:

RQ1 How do annotators' sociodemographics influence their offensiveness ratings of texts?

RQ2 How does the presence of lemmas belonging to specific HurtLex categories influence the offensiveness ratings?

RQ3 Do certain demographic groups exhibit different sensitivity to texts including certain HurtLex categories of offensive language?

Following previous works by Homan et al. (2024), we leverage *multilevel modelling* (Gelman and Hill, 2006) (or mixed-effects modelling), to analyze these three levels. This approach enables the examination of how sociodemographic factors, categories of hurtful language, and group-specific sensitivities interact, while accounting for inherent variability in both the text itself and the annotators.

Our analysis reveals that **intersectionality of sociodemographic traits** provides a more comprehensive explanation of rating behaviour compared to independent sociodemographics. Furthermore, we observe that the presence of specific types of hurtful words also informs the ratings pre-

[1] Throughout this work, we use the terms *hurtful* and *offensive* interchangeably to refer to language that may cause harm or be perceived as disrespectful, since both terms encompass a spectrum of harmful language phenomena (Poletto et al., 2020).

Proceedings of NLPerspectives @ LREC 2026, pages 44–55
12 May 2026.

diction. Finally, analyzing the interaction between sociodemographics and types of hurtful lemmas, **we uncover different sensitivities across *age* and *race* groups to different types of offensive lemmas.**

The full code is publicly available[2].

2. Related Works

In recent years, the growing presence of offensive and discriminatory language in public debate and on online platforms has attracted increasing attention, becoming a major concern for society and the scientific community in various fields. An approach to addressing this issue has been to develop models for the recognition of hate speech (Basile et al., 2019; Zampieri et al., 2020), requiring linguistic resources for model training and benchmarking mostly based on annotated texts taken from social media platforms (Poletto et al., 2020; Alkomah and Ma, 2022; Yu et al., 2024; Fortuna et al., 2020; Ollagnier, 2024).

Although previous research has focused mainly on textual resources, there are also studies that treat words as clues for analysing hate speech, using lexical knowledge to identify offensive language. Lexica based on this assumption are presented, for example, in Wiegand et al. (2018) and Bassignana et al. (2018). HateWiC (Wiegand et al., 2018) provides a basic and automatically expanded lexicon of words in context, based on the assumption that offensive terms constitute a subset of negatively polarized expressions. HurtLex (Bassignana et al., 2018), on the other hand, is a multilingual lexicon of hate originally developed for Italian and organised into 17 semantic categories. The lexicon was then expanded through links to synset-based lexical resources such as MultiWordNet and BabelNet, and extended to multiple languages through semi-automatic translation and expert annotations.

In the same years, a new paradigm in linguistic annotation has emerged, highlighting the inherently subjective nature of human annotation (Aroyo and Welty, 2015). In this context, disagreement is no longer seen as background noise, but as a meaningful signal (Uma et al., 2021; Plank, 2022). Building on this view, the perspectivist approach seeks to model and preserve annotators' viewpoints (Basile, 2021; Cabitza et al., 2023b). This has led to the spread of disaggregated corpora and datasets (Sap et al., 2022; Sachdeva et al., 2022; Frenda et al., 2023), and a shift in research towards the analysis of the perspectives that emerge from the annotation process itself (Mostafazadeh Davani et al., 2022; Homan et al., 2024; Sap et al., 2022).

Previous works have analyzed the effect of sociodemographic variables in subjective NLP tasks. Hu and Collier (2024) analyse the role of non-intersectional persona variables, finding that they explain up to 10% of the variability in those datasets. Homan et al. (2024) show that safety judgments in conversational AI are highly subjective and shaped by intersectional demographic factors, with Bayesian multilevel models revealing how gender, race/ethnicity, age, and education jointly influence annotators' perceptions of conversational harm and surface underrepresented perspectives.

3. HurtLens: A Corpus of Perspectivist Hurtfulness

In this section, we describe the construction of HurtLens, a corpus of social media posts featuring lemmas from the HurtLex lexicon of hurtful words (Bassignana et al., 2018), encompassing both offensive and non-offensive usages. We first present the lexicon and the source datasets upon which HurtLens relies, and then describe the construction process.

3.1. HurtLex

As previously mentioned, HurtLex (Bassignana et al., 2018) is a multilingual computational lexicon of offensive and hateful expressions, originally derived from the Italian lexicon *Le Parole per Ferire* developed by De Mauro (2016). We chose this lexical resource for its detailed categorical structure, which enables fine-grained analyses across different semantic types of offensive language. In addition, it is a widely adopted resource explicitly designed with a multilingual perspective (Stanković et al., 2020; Stamou et al., 2022; Tontodimamma et al., 2023; Osenova, 2024) and its effectiveness for offensive language detection has been demonstrated in numerous studies (Koufakou et al., 2020; Giordano and Di Buono, 2023; rcos and Pérez, 2023).

In this work, we leverage the English HurtLex-core lexicon, containing 501 lemmas. Each lemma is associated with a semantic category tag that reflects the taxonomy defined in the original Italian lexicon and preserved across its multilingual extension. These categories capture the kind of harmful or sensitive concept the word expresses, with some referring to social groups (e.g. ethnic slurs), or others to personal characteristics (e.g. physical or cognitive disability).

One of the authors of this paper, with a Linguistics training, manually reviewed each entry and removed entries with no known offensive usage. Entries were retained if they exhibited at least one of the following: (i) explicit offensiveness or abusive

[2] https://github.com/SDavenia/hurt_persp

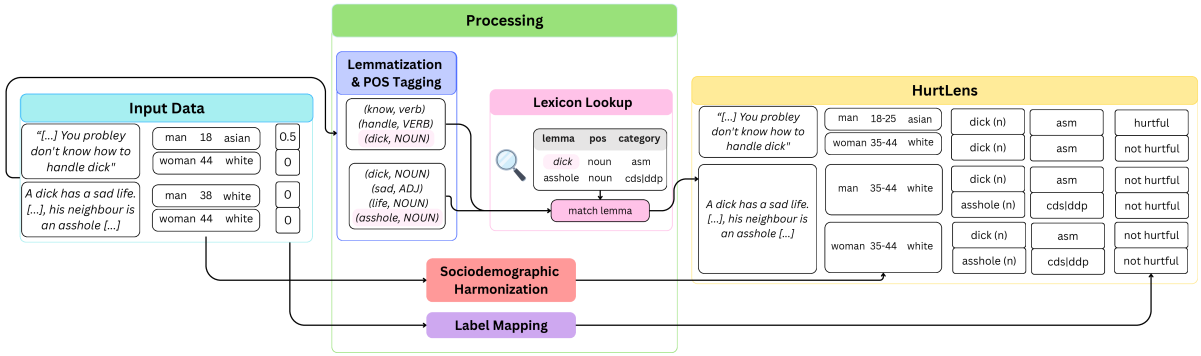


Figure 1: HurtLens construction steps. Example with entries from the SBIC dataset.

meaning (e.g., "*wh*re*"), (ii) non-literal derogatory usage arising from semantic shift or metaphor (e.g., "*pig*"), (iii) negative evaluative connotation without direct insult and (iv) vulgarity, such as terms related to sexuality that may occur in derogatory contexts.

Additionally, the same author reviewed the semantic categories to handle incorrect assignments due to translation (such as "*idiot*" associated with the category *plant*). We also include an additional category indicating lemmas that fall within the sphere of sexuality (*sexual words*) which are distinguished from words strictly related to *male genitalia* or *female genitalia*. The resulting lexicon includes 341 of the 501 entries available in the original published resource and a total of 18 lexical categories. The full set of categories, along with their definitions and example lemmas are included in Appendix A

3.2. Datasets and Harmonization

The selection of source datasets is guided by the following criteria: (i) availability of English-language data, (ii) focus on abusive or related phenomena, and (iii) provision of disaggregated annotations, including annotator-level sociodemographic information.

Table 1 provides an overview of the four datasets included in HurtLens. Below we detail what annotators were asked to label in each dataset and how we mapped the original labels onto a binary offensive/not offensive distinction.

- Attitudes (Sap et al., 2022): Annotators were asked *"How toxic/hateful/disrespectful or offensive does this post seem to you?"*, *"How much could this post be seen as toxic/hateful/disrespectful or offensive to anyone?"* and *"In your opinion, how racist is this post?"* on a 1-5 scale.
- Voiced (Weerasooriya et al., 2023): Annotators were generally asked if the text was offensive to them, using a four-point scale (*Not at all Offensive, Slightly Offensive, Moderately Offensive, Very Offensive*).
- SBIC (Sap et al., 2020): Annotators were asked *whether a post could be considered "offensive to anyone"* with three possible categorical answers (*yes, maybe, no*).
- POPQUORN (Pei and Jurgens, 2023): Annotators were asked *"Consider you read the above comment on Reddit, how offensive do you think it is?"*. on a 1-5 Likert scale, where 1 means *"Not offensive at all"* and 5 means *"Very offensive"*.

Given that each dataset adopts its own conceptualization of offensive speech, we perform a harmonization process over both the offensive speech labels and the annotators' demographic variables.

For the offensive speech labels, we reduce the label space to a binary setting, mapping all instances to either the offensive or non-offensive class. Specifically, we adopt a conservative harmonization in which all labels indicating any degree of offensiveness are grouped into the offensive class [3]. This strategy was chosen to retain all possible ways of using a lemma offensively, regardless of the degree, to capture a broader range of potential harmfulness. We further discuss the implications of this choice in Section 7.

For the sociodemographic variables, we normalize heterogeneous annotations into a unified schema. Gender is mapped to four categories (*man, woman, nonBinary, transman*), while for race we consider seven groups (*asian, black, hispanic, white, native, arab, other*). Age is discretized into standard intervals (*18–24, 25–34, 35–44, 45–54, 55–64, 65 or older*). Political ideology is reduced to three categories (*left, right, other*), and education levels are grouped into five (*less_than_high_school, high_school_diploma, bachelors_degree, graduate_degree, other*).

[3] https://github.com/SDavenia/hurt_persp/blob/main/utils/dataset_lexicon_processing.py

Dataset	Platform	#Annotations	#Annotators	#Texts	#Off. Texts	Avg/Annotator	Avg/Text
Attitudes	Twitter	3454	184	627	586(93.0%)	18.77	5.51
Voiced	Reddit	44676	726	2338	2327(∼ 100%)	61.54	19.11
SBIC	Tw./Red./Gab/Storm.	144649	304	45223	30863(68.0%)	377.71	2.54
POPQUORN	Reddit	13036	262	1500	1338(89.0%)	49.76	8.69

Table 1: Summary of datasets with annotation statistics containing unique number of annotators, unique total number of annotations, number of annotators, number of texts and number of offensive texts. We report the number of offensive texts by counting instances where at least one annotator flagged it as offensive.

This harmonization enables consistent cross-dataset comparisons while preserving the core demographic distinctions captured in the original annotations.[4]

3.3. HurtLens

For each HurtLex lemma, we construct HurtLens by retrieving from the selected datasets instances in which the lemma appears, encompassing both hurtful and non-hurtful usages. We use `spaCy` to perform lemmatization and part-of-speech tagging on the texts, and then we match the extracted lemma–POS pairs with the corresponding HurtLex entries.

A visual workflow of HurtLens is reported in Figure 1, while Table 2 reports its main statistics, where triplets correspond to the extracted text-lemma-annotation units.

Dataset	#Triplets	#Texts	#Offensive	#Annotators
Attitudes	3052	364(58.0%)	356(97.8%)	148(80%)
Voiced	41468	1332(57%)	1331 (∼ 100%)	726(100%)
SBIC	59429	18066(40%)	14331(79.3%)	235(77%)
POPQUORN	6665	569(38%)	546(96.0%)	262(100%)
HurtLens	110614	20331(40.9%)	16564(81.5%)	1371(92.9%)

Table 2: Summary of HurtLens statistics, triplets are the extracted text-lemma-annotations. We report also the number of texts extracted, the number and percentage of texts originally annotated as offensive, the number of unique annotators, and the percentage of annotators retained from the original datasets after retrieval.

4. Methodology

We model the offensiveness label, which is our dependent variable y as a binary variable using a generalized linear mixed model, using logistic regression with a logit link function.

[4] https://github.com/SDavenia/hurt_persp/blob/main/data/sociodemographic_mappings.json

4.1. Preprocessing for Modelling

For this analysis, we only consider the sociodemographic variables which are available in all datasets considered, namely *race*, *age* and *gender*. To ensure reliable estimates of the model parameters, we excluded all sociodemographic levels for which any combination with other variables (corresponding to the interaction effects) has less than 30 observations. A similar filtering is conducted on the HurtLex categories. We filter these cases for the purposes of the present analysis; however, the released resource retains them for other downstream uses.

Moreover, we remove all instances where some sociodemographic is set to *other*, as it is deemed not informative. After this filtering step we are left with the following levels for each variable: age (*18-24, 25-34, 35-44, 45-54, 55-64*), therefore excluding *65 or older*; race (*white, black, asian*), excluding *hispanic, native, arab*; gender (*man, woman*), excluding *non-binary, transman*, while for the HurtLex categories we exclude *is_or* (plants), with a total of 17 types of offensive language.

After this filtering step, the dataset upon which we build our models consists of 70062 text-annotator instances, from 1247 unique annotators on 19945 unique texts.

For modeling purposes, both sociodemographic variables and lexical category tags correspond to the fixed effects under investigation for this analysis, while a by-annotator and by-text intercept for *annotator_id* and *text_id* are included as random effects to account for text and annotator variability. Additionally, for the sociodemographic variables we set the reference level to the most common traits, namely *white, man, 25-34*.

4.2. Models Definition

The **null model** (**N**) does not include any fixed effexts and only considers a by-annotator and by-text random intercept. In R notation:

$$y \sim 1 + (1|\text{rater\_id}) + (1|\text{text\_id})$$

Sociodemographic-only Models These models only consider the sociodemographic variables as fixed effects.

For the first model we consider these variables as independent, non-intersecting predictors. We denote this model as the **sociodemographic model** (**S**), in R notation:

$$y \sim \text{race} + \text{age} + \text{gender} + (1|\text{rater\_id}) + (1|\text{text\_id})$$

For the second model, we focus on the interaction between *race* and the other two sociodemographics, grounded in previous literature on intersectionality which showed that it is a common predictor to interact with other variables (Homan et al., 2024). We denote this model as the **sociodemographic race-intersectional model** (**SRi**), in R notation:

$$y \sim \text{race} * (\text{age} + \text{gender}) + (1|\text{rater\_id}) + (1|\text{text\_id})$$

Tags Model We consider a model using each binary variable denoting the presence of lemmas from certain HurtLex categories as independent fixed-effects. We denote this model as the **tag model** (**T**), in R notation:

$$y \sim \text{is\_ps} + ... + \text{is\_re} + (1|\text{rater\_id}) + (1|\text{text\_id})$$

Sociodemographics-Tags Interaction Models For this set of models, we consider both sociodemographic variables and whether the text contains lemmas from certain HurtLex categories.

We first include a model where the race-intersectional model is enriched with the various tags. This model is denoted as **race-intersectional + tags model** (**SRi-T**), in R notation:

$$\begin{aligned} y \sim\ & \text{race} * (\text{age} + \text{gender}) + \\ & + \text{is\_ps} + \cdots + \text{is\_re} \\ & + (1 \mid \text{rater\_id}) + (1 \mid \text{text\_id}) \end{aligned}$$

Finally, we define a model that allows us to investigate how different sociodemographic traits interact with texts containing different types of offensive language, identified via HurtLex categories. Fitting an interaction term between each category tag and sociodemographic variable would lead to an overly-complex model. As such, we conduct an exploratory analysis to identify which interaction terms are of interest.

The methodology for this exploration is described in Section 4.3 with results in Section 5.1, leading to the inclusion of 4 interaction terms between *age* and category tags and 3 for *age*. The final model is denoted as **race-intersectional + tag-sociodemographic model** (**SRi-TS**), in R notation:

$$\begin{aligned} y \sim\ & \text{race} * (\text{age} + \text{gender}) + \\ & + \text{is\_ps} + \cdots + \text{is\_re} \\ & + (\text{is\_asm} + \text{is\_ddp} + \text{is\_ps} + \text{is\_asf}) * \text{race} \\ & + (\text{is\_ddf} + \text{is\_pa} + \text{is\_is}) * \text{age} \\ & + (1 \mid \text{rater\_id}) + (1 \mid \text{text\_id}) \end{aligned}$$

4.3. Exploratory Analysis Methodology

To select meaningful interaction terms for the SRi-TS model, we conduct an exploratory analysis to identify which texts, grouped by the presence of lemmas from specific HurtLex categories, cause different demographic levels to diverge most in their offensiveness ratings.

For every pair of levels of a sociodemographic variable (A, B) and category c, we identify the textual instances containing at least one lemma belonging to category c that were rated by at least one annotator from each group, and obtain a within-group majority label (excluding ties). We compute the divergence between groups A, B on category c as:

$$w^c_{AB} = (n^c_{A>B} - n^c_{B>A})/n_{AB}$$

Where $n^c_{A>B}$ indicates the number of instances annotated as offensive by sociodemographic level A but not B and $n^c_{B>A}$ the opposite, over a total of n^c_{AB} instances with lemmas from category c that were rated by both. This coefficient $w_{AB} \in [-1, 1]$ serves as a sensitivity index: values near 0 indicate consensus, while values toward the extremes indicate that one group consistently perceives those texts as more offensive than the other. To avoid drawing conclusions from a few observations, we exclude pairs with fewer than 20 comparisons, and include interaction terms between the sociodemographic variable and c if there is at least a coefficient $w^c_{AB} > 0.25$. This procedure is exploratory and heuristic rather than a formal model-selection method, and its limitations are discussed in Section 7.

5. Results

5.1. Exploratory Analysis

For our exploratory analysis, we compute w^c_{AB} for every pair of values A, B for a specific sociodemographic variable and category c. We visualize only comparisons for pairs and tags where there is at least a coefficient larger in absolute value than 0.20. Cells highlighted in green indicate that annotators from the first group on the x-axis annotated more instances as offensive, while those in pink indicate the opposite.

Figure 2: w_{AB} values for levels of *race*, only showing pairs and tags where at least one entry is greater or equal than 0.20. A positive value (in green) indicates that the first entry in the comparison annotated more as offensive, while a negative one indicates the opposite.

Figure 2 shows the results for *race*. We observe that across the observed categories *white* annotators tend to label more instances as offensive than both *black* and *native* annotators, and the same behaviour is observed for *asian* annotators compared to *black* ones.

From this exploration we decided to include an interaction term between *race* and *asf, ddp, asm, ps* (corresponding to female genitalia, cognitive disability, male genitalia and negative stereotypes/ethnic slurs). An interaction term with *pr* (words related to prostitution) is not included as it appears above our threshold only in a comparison with *race=hispanic*, which is not included in the model as stated before.

Figure 3 shows the results of the exploratory analysis for *age*. We observe that in general across the observed categories, young annotators tend to label more instances as offensive compared to older ones, with only some exceptions.

Similarly to what we did for *race*, we decided to include interaction terms between *age* and *ddf, is, pa* (corresponding to physical disability, social/economic disadvantage and professions/occupations). An interaction term with *ddp* (corresponding to cognitive disability) is not included since level *65 or older* is excluded from the model.

Concerning *gender*, across all combinations, no w_{AB} is above our pre-specified threshold and no interaction term is included in the model.

5.2. Model Comparison

We report number of degrees of freedom, marginal and conditional R^2, Akaike (AIC) and Bayesian (BIC) Information Criteria across all models in Table 3. Degrees of freedom reflect model complexity in terms of the number of estimated parameters. Marginal R^2 represents the proportion of variance explained by fixed effects alone, while conditional R^2 captures the variance explained by both fixed and random effects. AIC and BIC are information criteria used for model comparison, balancing goodness of fit with model complexity; lower values indicate a better trade-off between fit and parsimony.

Model	df	M-R^2(↑)	C-R^2(↑)	AIC(↓)	BIC(↓)
N	3	—	81.1	62661	62689
S	10	3.2	81.6	62592	62683
SRi	20	5.3	81.7	62573	62756
T	19	4.2	81.6	61189	**61363**
SRi-T	36	9.3	**82.2**	61101	61431
SRi-TS	56	**9.4**	**82.2**	**61087**	61600

Table 3: Model comparison statistics, reporting number of degrees of freedom (df), marginal and conditional R^2 (M-R^2, C-R^2) as percentage of total variability, AIC and BIC.

The results from the null model reveal that 81.1% of variability can be accounted for by the random effects, indicating that a large part of variance is explained by text-specific and annotator-specific characteristics.

The results from the sociodemographic models indicate that sociodemographic variables treated independently (S) account for 3.2% of the total variability. However, when modeled with an intersectional approach (SRi) they account for 5.3% of total variability, **emphasizing the importance of considering interaction effects with an intersectional approach**. These findings agree with Hu and Collier (2024), who observed that sociodemographic variables accounted for 4.5% and 2.9% on AnnwithAttitudes and POPQUORN respectively. Similarly, in line with previous work on intersectionality by (Homan et al., 2024), when considering the intersection of *race* with the other sociodemographic variables the explained variability increases.

The results from the tag model (T) indicate that the presence of specific HurtLex lemma category tags accounts for 4.2% of total variability when treated independently. This effect is similar in magnitude to that of the sociodemographic variables.

When considering the models leveraging both tags and sociodemographics information (SRi-T,

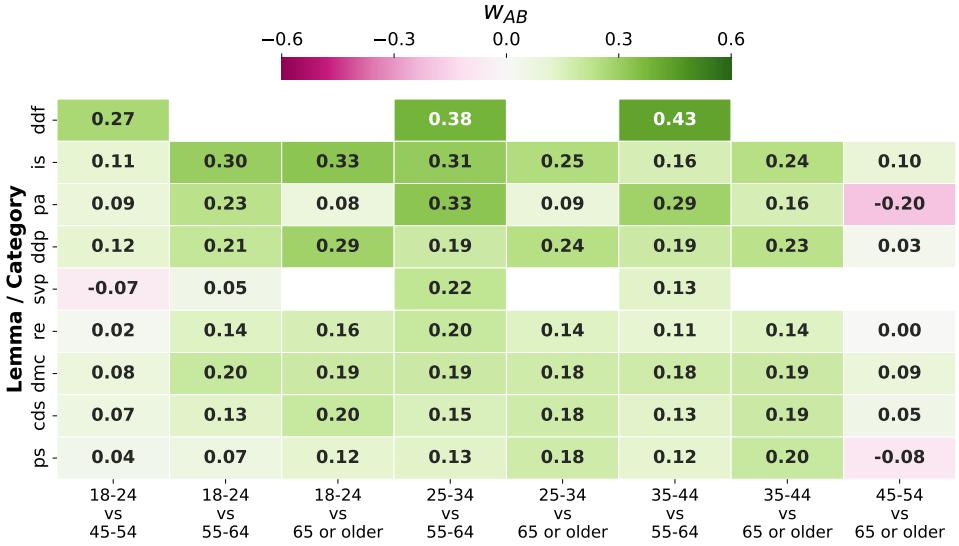


Figure 3: w_{AB} values for levels of *age*, only showing pairs and tags where at least one entry is greater or equal than 0.20. A positive value (in green) indicates that the first entry in the comparison annotated more as offensive, while a negative one indicates the opposite.

SRi-TS), the variability explained by those factors jumps to 9.3% for the model not considering sociodemographic-tags interactions and 9.4% for the other. This increase compared to both tags-only and sociodemographic-only models indicates that the sociodemographics and presence of specific lemma category tags model different aspects of the variability.

While the interaction model (SRi-TS) only results in a small increase in marginal R^2, we choose to utilize this model as it allows us to investigate the interaction between sociodemographic groups and the presence of lemmas from HurtLex belonging to the identified categories. Moreover, while BIC favours simpler models due to its penalty for larger parameters, the decrease in AIC compared to all other models justifies the inclusion of the interaction terms of the SRi-TS model.

5.3. Effects and Interaction of Sociodemographics and Tags

For the rest of the analysis, the predicted probabilities of offensive ratings are obtained with the SRi-TS model using the Average Marginal Effect (AME) method via the `ggeffects` package `predict_response` function with `margin="average"`, ensuring the predicted probabilities are averaged over the distribution of all observations (L decke, 2018). Additionally, we include the 95% confidence interval error bars for these predicted probabilities.

Effect of Sociodemographics To answer **RQ1**, we investigate how the effect of *age* and *gender* varies across different levels of *race*. This allows us to visualize the intersectional component of the model, allowing for the influence of *age* and *gender* to vary depending on the annotator's *race*.

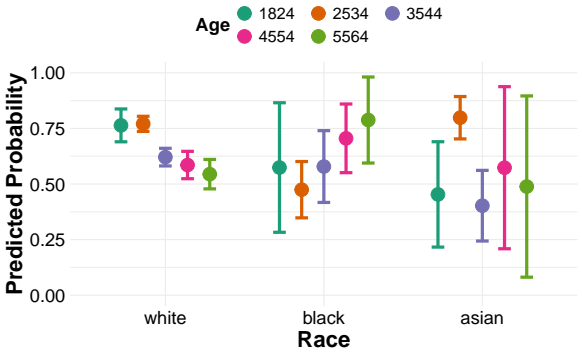


Figure 4: Predicted probability of offensiveness rating (AME) for *race* and *age* interaction.

Figure 4 shows the effect of *age* across different levels of *race*. We note that while for *white* annotators older age groups have lower predicted probability of rating a text as offensive than younger groups, an opposite trend is observed on average when focusing on *black* annotators, with only little overlap between the error bars for $25-34$ and $55-64$ age groups. Within *asian* annotators, we observe that the $25-34$ group is identified as being the one most likely to rate content as offensive. However, for the other groups, particularly the older ones, the limited number of observations leading to large confidence intervals does not allow us to derive strong conclusions.

Similarly, Figure 5 shows the effect of *gender*

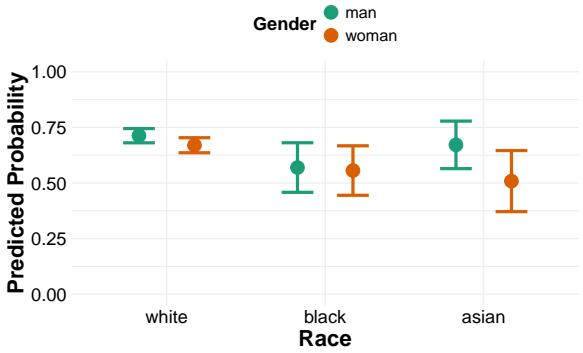


Figure 5: Predicted probability of offensiveness rating (AME) for *race* and *gender* interaction.

across different levels of *race*. We observe that across both *white* and *asian* annotators, women appear to be less likely to rate texts as offensive, while this behaviour is not observed in *black* annotators.

Effect of Tags To answer **RQ2**, we investigate how the presence of lemmas belonging to certain HurtLex Categories impacts the predicted probability of offensiveness of those texts. Figure 6

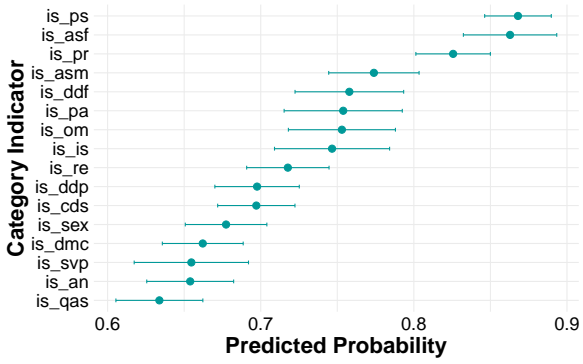


Figure 6: Predicted probability of offensiveness rating (AME) for texts containing lemmas from different HurtLex category tags.

shows how the predicted probability of a text being offensive changes based on which HurtLex tag the text contains. We observe that texts containing words related to stereotypes, slurs (*is_ps*), female genitalia (*is_asf*) and prostitution (*is_pr*) are much more likely than the others to be annotated as offensive, with the CI lower bound above 0.80. On the contrary, texts containing potentially negative words (*is_qas*), animals (*is_an*), seven deadly sins (*is_svp*) and moral/behavioral defects (*is_dmc*) are less likely than the others to be annotated as offensive, with the upper bound of the CI falling below 0.7.

Effect of Sociodemographics and Tags Interactions Finally, to answer **RQ3** we investigate how different sociodemographic groups vary in their offensiveness ratings on texts containing words belonging to different HurtLex categories. For our SRi-TS model, described above, we included 3 interaction effects with *age* and 4 with *race*. Here we report only those where we observed at least a significant interaction effect compared to the base group (i.e. $25-34$ for *age* and *white* for *race*).

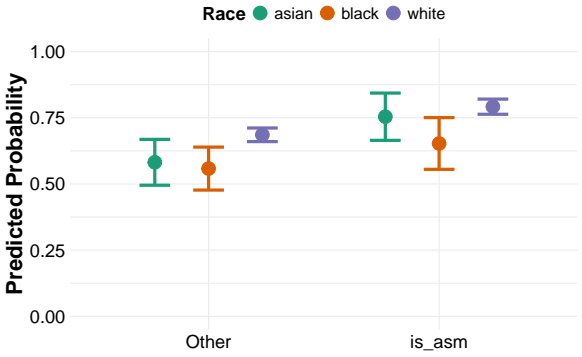


Figure 7: Predicted probability of offensiveness rating (AME) for texts containing words related to male genitalia (*is_asm*) for varying levels of *race*.

Figure 7 shows how annotators with different *race* change their ratings comparing texts without male-genitalia related words to those containing them. We observe that while across all groups the presence of *is_asm* (male genitalia) words increases the predicted probability of offensiveness ratings, this increase is more pronounced for *asian* annotators. In particular, *asian* annotators are more similar to *black* annotators when those terms do not appear, and less likely than *white* annotators to identify texts as offensive, but their offensiveness ratings are more similar to *white* annotators for texts containing *asm* related lemmas. This indicates that *asian* annotators appear to be particularly sensitive to lemmas belonging to this category. This agrees with our exploratory findings where we did not identify any difference between *white* and *asian* annotators on these terms.

Similarly, Figure 8 shows how annotators with different *race* change their ratings comparing texts with cognitive-disability related words to those containing them. We observe different effects across the different *race* groups, where *asian* and *white* annotators become slightly more likely to identify texts as offensive when they contain *ddp* (cognitive disability) lemmas, while for *black* annotators the opposite behaviour is observed. This complements our exploratory findings, where we observed that both *asian* and *white* annotators were more likely than *black* annotators to annotate as offensive texts containing cognitive-disability related

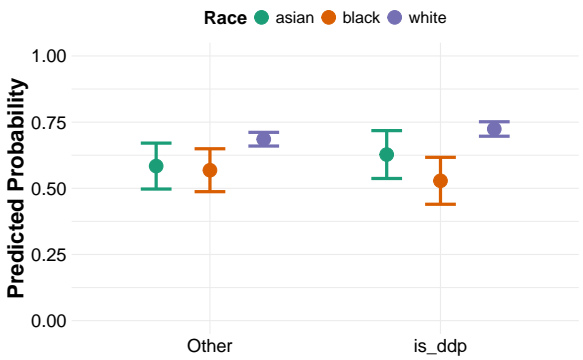


Figure 8: Predicted probability of offensiveness rating (AME) for texts containing words related to cognitive disability (*is_ddp*) for varying levels of *race*.

words.

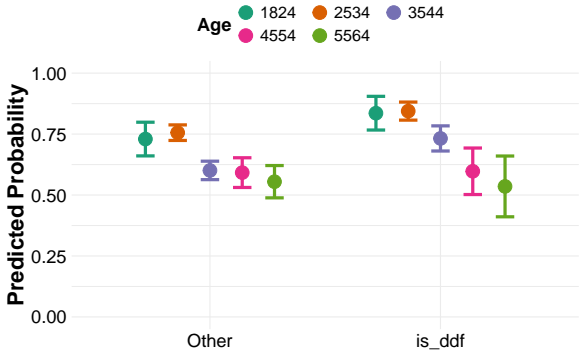


Figure 9: Predicted probability of offensiveness rating (AME) for texts containing words related to physical disability (*is_ddf*) for varying levels of *age*.

Finally, Figure 9 shows how annotators with different *age* change their ratings comparing texts with physical-disability related words to those containing them. We observe different effects across various *age* groups. In particular, while for younger age groups (up to $35-44$ group), the annotators are more likely to rate texts containing *ddf* (physical disability) lemmas as offensive, this change is not observed in the older demographics. This points to younger generations being more sensitive to the topic of physical disability.

6. Conclusions

In this work, we introduce HurtLens by enriching existing disaggregated datasets with hurtful lexicon-based information. We demonstrate that offensiveness perception is shaped by both the sociodemographic characteristics of annotators, the lexical categories of hurtful words present in texts and the interaction between them. Through the proposed mixed-effects modelling approach (SRi-TS) we found that intersectional modelling increased explained variance (marginal R^2) compared to models including independent sociodemographic predictors. In particular, *race* and *age* groups exhibited different rating shifts depending on the presence of specific lexical categories, uncovering **novel insights into group-specific sensitivities to distinct types of hurtful language.**

These findings underscore the importance of adopting perspectivist approaches to offensive language detection.

Moreover, our methodology is generalizable to other pragmatic phenomena, provided a prior structured knowledge, e.g., a lexicon capturing the target phenomenon, and suitable disaggregated source datasets.

As part of future work, we intend to further investigate not only the categories but also the individual lemmas in HurtLex, examining how the perception of specific lemmas varies according to sociodemographic variables and context of use. By shifting the analysis to the lemma level, it will also be possible to incorporate additional lexical resources in order to expand the lexical coverage.

Furthermore, we aim to leverage perspective-specific examples from HurtLens to tackle over-moderation issues in Large Language Models on offensive speech detection. Current systems often struggle to identify hurtless usage of certain lemmas, leading to many False Positives (Draetta et al.). By utilizing a perspectivist corpus which also includes non-offensive usage of certain words and demographic-specific interpretations could help models disambiguate between different uses of these words.

Finally, we intend to adopt community-led and participatory approaches to validate and refine the resource with input from the demographic groups represented in our analyses.

7. Ethics Statement and Limitations

Ethics Statement The primary goal of HurtLens is to provide a resource for combating online hate by enabling more nuanced models that can better account for the multifaceted nature of offensiveness perception across different demographic groups, ultimately supporting more equitable content moderation systems. We acknowledge that resources documenting hurtful language carry inherent risks of misuse. HurtLens is intentionally designed to expose the diversity of perspectives on offensiveness rather than to provide a definitive catalog of harmful terms. The resource should not be used to target or harass specific demographic groups, nor to train models that disproportionately silence marginalized voices. Our perspectivist approach is motivated by the goal of reducing bias

in content moderation by making systems aware of how different communities perceive harm differently, thereby avoiding both over-moderation of minority perspectives and under-moderation of actual harmful content. We strongly discourage any application of this resource that would amplify harm or reinforce existing power imbalances in online spaces.

Limitations We identify limitations of our work. First, it is restricted to the English language, which may limit the generalization of the findings to other linguistic and cultural contexts. Similarly, our methodology builds upon the HurtLex lexicon, and therefore inherits its coverage limitations despite our revision efforts. Additionally, the resource focuses on explicit lexical realizations of hurtful language and does not account for implicit expressions of hate or stereotypes, which often require deeper contextual and pragmatic interpretation. Furthermore, our use of spaCy for lemmatization may fail to resolve non-standard slang or intentional misspellings.

From a modeling perspective, our selection of interaction terms between sociodemographic levels and HurtLex categories was guided by an exploratory analysis rather than an exhaustive search using more appropriate model selection criteria. While this approach allowed us to identify salient interaction terms, it is not exhaustive. Moreover, we did not consider intersectional groups interactions with lexical categories, which could capture additional patterns. As a methodological note, employing Bayesian approaches could provide more robust and interpretable estimates, but given the dataset size we were limited by computational requirements for this exploratory analysis.

Another limitation arises as we reduced heterogeneous annotation schemes to a binary offensive vs. non-offensive label which may hide important differences in the original scales. In particular, our conservative choice to treat any degree of offensiveness as offensive merges mild, ambiguous, and severe cases into a single case. While this supports a broad view of harmful usage, different thresholding choices (e.g., excluding mid-scale values) could lead to different distributions and effects.

Finally, combining multiple source datasets introduces potential confounds, as they differ in annotation guidelines, platform, and annotator composition. As a result, some observed demographic or lexical effects may partly reflect dataset-specific artifacts rather than general patterns. While our unified analysis aims at capturing broader trends, controlling for dataset effects (e.g., via per-dataset models or dataset indicators) could help disentangle these factors, and we leave this for future work.

Acknowledgements

This work was funded by the partnership with Amazon Science "Multilingual personalization through perspective-aware Language Modeling" and the project NEIKEA (Bando CSP TRAPEZIO - Linea 1 - Paving the way to research excellence and talent attraction).

8. Bibliographical References

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Inf.*, 13(6):273.

Iv n rcos and Jaime Pérez. 2023. Detecting hurtful humour on twitter using fine-tuned transformers and 1d convolutional neural networks. In *IberLEF@SEPLN*.

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. 2021. *It's the End of the Gold Standard as We Know It: Leveraging Non-aggregated Data for Better Evaluation and Explanation of Subjective Tasks*, page 441–453. Springer International Publishing.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, pages 52–57, Turin, Italy. CEUR Workshop Proceedings.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023a. 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.

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

Tullio De Mauro. 2016. Le parole per ferire. *Internazionale*. Compiled for the "Joe Cox" Committee on intolerance, xenophobia, racism and hate phenomena of the Italian Chamber of Deputies, which issued a Final Report in 2017.

Lia Draetta, Soda Marem Lo, Samuele D'Avenia, Valerio Basile, and Rossana Damiano. Testing llms' sensitivity to sociodemographics in offensive speech detection.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30.

Paula Fortuna, Juan Soler Company, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. pages 6786–6794.

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.

Andrew Gelman and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Luca Giordano and Maria Pia Di Buono. 2023. Assessing Italian news reliability in the health domain through text analysis of headlines. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 538–548, Vienna, Austria. NOVA CLUNL, Portugal.

Christopher Homan, Gregory Serapio-Garcia, Lora Aroyo, Mark Diaz, Alicia Parrish, Vinodkumar Prabhakaran, Alex Taylor, and Ding Wang. 2024. Intersectionality in AI Safety: Using Multilevel Models to Understand Diverse Perceptions of Safety in Conversational AI. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 131–141, Torino, Italia. ELRA and ICCL.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the Persona Effect in LLM Simulations.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *J. Artif. Intell. Res.*, 71:431–478.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

Daniel L decke. 2018. ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, 3(26):772.

Aida Mostafazadeh Davani, Mark D az, 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.

Anais Ollagnier. 2024. CyberAgressionAdo-v2: Leveraging pragmatic-level information to decipher online hate in French multiparty chats. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4287–4298, Torino, Italia. ELRA and ICCL.

Petya Osenova. 2024. On a hurtlex resource for Bulgarian. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, pages 214–219, Sofia, Bulgaria. Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences.

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.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

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.

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.

Vivian Stamou, Iakovi Alexiou, Antigone Klimi, Eleftheria Molou, Alexandra Saivanidou, and Stella Markantonatou. 2022. Cleansing & expanding the HURTLEX(el) with a multidimensional categorization of offensive words. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 102–108, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Ranka Stanković, Jelena Mitrović, Danka Jokić, and Cvetana Krstev. 2020. Multi-word expressions for abusive speech detection in Serbian. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 74–84, online. Association for Computational Linguistics.

Alice Tontodimamma, Lara Fontanella, Stefano Anzani, and Valerio Basile. 2023. An italian lexical resource for incivility detection in online discourses. *Quality & Quantity: International Journal of Methodology*, 57(4):3019–3037.

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.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. 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 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

Zehui Yu, Indira Sen, Dennis Assenmacher, Mattia Samory, Leon Fröhling, Christina Dahn, Debora Nozza, and Claudia Wagner. 2024. The unseen targets of hate - A systematic review of hateful communication datasets. *CoRR*, abs/2405.08562.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and ağrı öltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

9. Language Resource References

Bassignana, Elisa and Basile, Valerio and Patti, Viviana. 2018. *HurtLex*. PID https://github.com/valeriobasile/hurtlex.

Pei, Jiaxin and Jurgens, David. 2023. *POPQUORN*. PID https://github.com/Jiaxin-Pei/potato-prolific-dataset.

Sap, Maarten and Gabriel, Saadia and Qin, Lianhui and Jurafsky, Dan and Smith, Noah A. and Choi, Yejin. 2020. *SBIC*. PID https://huggingface.co/datasets/allenai/social_bias_frames.

Sap, Maarten and Swayamdipta, Swabha and Vianna, Laura and Zhou, Xuhui and Choi, Yejin and Smith, Noah A. 2022. *Attitudes*.

Weerasooriya, Tharindu and Dutta, Sujan and Ranasinghe, Tharindu and Zampieri, Marcos and Homan, Christopher and KhudaBukhsh, Ashiqur. 2023. *Voiced*. PID https://huggingface.co/datasets/Lab-PL/voiced.

A. Additional Details

A.1. HurtLex categories

Column Tag	Definition	Example
ps	negative stereotypes dearethnic slurs	jewish
pa	professions and occupations	cop
ddf	physical disabilities and diversity	disabled
ddp	cognitive disabilities and diversity	dumbass
dmc	moral and behavioral defects	liar
is	words related to social and economic disadvantage	poor
or	plants	melon
an	animals	snake
asm	male genitalia	dick
asf	female genitalia	pussy
pr	words related to prostitution	slut
om	words related to homosexuality	twink
qas	with potential negative connotations	camp
cds	derogatory words	baby
re	felonies and words related to crime and immoral behavior	abuse
svp	words related to the seven deadly sins of the Christian tradition	rage
sex	words related to sexual acts	fuck

Table 4: Hurtlex categories with their definitions and example lemmas.

A Measure of Systematic Disagreement

Valerio Basile
University of Turin

Abstract

This paper introduces a new metric, called σ, that quantifies the degree of systematicity in inter-annotator disagreement. The metric is inspired by Structural Balance Theory and is designed to approximate the clusterability of annotators in a dataset. When paired with a standard inter-annotator agreement measure such as Krippendorff's α, σ provides a complementary signal designed to capture the extent to which disagreement stems from genuine subjective factors rather than from ambiguity or annotation noise. The metric is applied to over twenty datasets encoding a broad variety of annotations, showing a tendency to produce higher values for tasks conventionally considered subjective.

Keywords: Inter-annotator agreement, Subjective tasks, Perspectivist NLP

1. Introduction

Inter-annotator agreement has been the main lens to quantify, or at least approximate, the quality of a human-annotated dataset (Artstein and Poesio, 2008). While this connection is unchallenged, the recent attention on the phenomenon of Human label variation (Plank, 2022) in the NLP research community has spurred proposals to investigate the aspects that impact observed disagreement and its multiple causes. Basile et al. (2021) argue that annotator disagreement can be traced to a variety of causes, which they cluster in two main sources:

- **Ambiguity**, encompassing all the exogenous factors such as lack of clear annotation guidelines, less-than-ideal annotation interfaces, human distraction, or genuine errors;
- **Subjectivity**, the characteristic of a language annotation task to depend strongly on the individual perception, as well as the personal or cultural background of the annotator.

The study of subjectivity-bound disagreement, already an important aspect mentioned in the seminal work on disagreement by Aroyo and Welty (2015), has led to interesting developments in the NLP research community, including learning with disagreements (Leonardelli et al., 2025; Uma et al., 2022) and the perspectivist turn in NLP (Cabitza et al., 2023).

While disagreement is a measurable signal (Section 2), at the moment we lack a straightforward procedure to determine the contribution of individual factors. The objective of this paper is to introduce a computational tool to measure the degree of systematicity of the disagreement of annotators who expressed judgments on the same data.

2. Related Work

Cohen's κ (Cohen, 1960) and Scott's π (Scott, 1955) are quantitative indexes of the amount of agreement between two annotators. With respect to simpler measures (e.g., percent agreement), κ and π account for the probability of the annotators to agree by chance, just by virtue of imbalanced nature of the label distribution, while differing slightly in the definition of chance agreement. Fleiss' κ extends both Cohen's κ and Scott's π to an arbitrary number of annotators. Krippendorff's α further extends π to cases where the annotation matrix is sparse, i.e., not all annotators annotated every instance, which is a common scenario, e.g., in a crowdsourcing context.

While the aforementioned measures are widespread, we note that they quantify disagreement independently from its origin, or, in other words, disregarding any knowledge about the identity of the annotators. Checco et al. (2017a) identify a set of pitfalls in the use of κ-like metrics, especially in crowdsourcing contexts. Dumitrache et al. (2018) introduce a set of metrics that consider the distribution of the annotated instances and the distribution of the annotators jointly, to account for different annotator behaviors.

Akhtar et al. (2019) introduce the polarization index, a measure of systematic disagreement at the instance level. While the authors note that the average polarization over an annotated dataset can approximate an overall measure of systematic disagreement, the polarization index needs a predetermined partition of the annotator cohort into groups. Recently, Tsirmpas and Pavlopoulos (2026) build on the polarization idea and introduce statistical tools to quantify the level of polarization in relation with determined annotator groups (e.g., by sociodemographic traits). Alacam et al. (2025) propose a method to discriminate the effect of subjectivity vs. uncertainty in hate speech annotation by leveraging the confidence measured through gaze data. To

Proceedings of NLPerspectives @ LREC 2026, pages 56–65
12 May 2026.

cope with the need to know, or somehow box in the identity of the annotators, several works propose methods to learn annotator representations, mainly as a step towards modeling human perspectives in supervised classification contexts. Lo and Basile (2023) apply clustering algorithms to vectors representing the entirety of each annotators activity. Conversely, the approaches of Mostafazadeh Davani et al. (2022) and Mokhberian et al. (2024) learn annotator embedding from the annotated data. These works are highly related to the present paper, where a metric is defined that approximates the clusterability of annotators. More precisely, this paper proposes a metric that validates the assumptions made by Lo and Basile (2023) and others, i.e., that subjective tasks tend to produce annotations with more separable clusters of annotators.

3. Systematicity of Inter-annotator Disagreement

In Social Psychology, like/dislike relationships between humans are modeled through signed undirected graphs, where the nodes represent the individuals and and edge between A and B represent their relationship (if present) as positive (+) or negative (-). The theory of **Structural Balance** (Cartwright and Harary, 1956) models triadic relationships and their possible states. As a classic example, if a person A has a positive relation (e.g. affection) for a person B, and if B is responsible for an entity X (e.g. an event or an artifact, then there will be a tendency for A to like or approve of X. However, if the direct attitude of A towards X (without considering B) is negative, the triangle A-B-X is "imbalanced". At a more abstract level, the theory posits that any three entities in relationship with each other (a *triangle*) in such a graph tend towards a balance achieved by either all edges being + or a situation where one edge is + and the other two are -. The other two possible configurations are instead regarded as imbalanced, as summarized in Figure 1.

Davis (1967) applies the notion of structural balance to graphs, calling a balanced graph an undirected signed graph where all triangles (i.e., cycles of length 3) are balanced, and proving that a balanced graph has a unique clustering. I extend this definition to a degree of balancedness, that is, the rate of triangles in an undirected signed graph that are balanced:

$$\sigma = \frac{(\#\text{balanced triangles})}{(\#\text{triangles})}$$

Next, the outcome of an annotation task is represented as a signed undirected graph, where each node represents an annotator, and the $+-$ sign indicates whether the pair agrees $(+)$ or disagrees $(-)$.

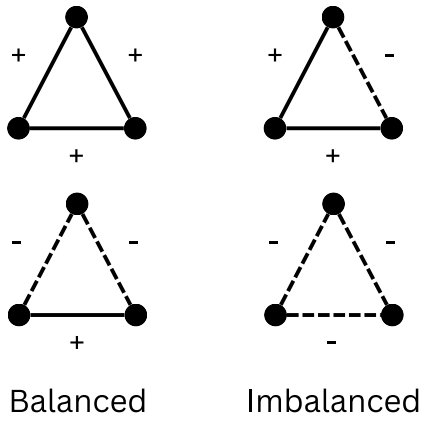


Figure 1: Possible configurations of graph triangles according to the sign of their edges in Structural Balance theory.

The pairwise Krippendorff's α is computed ($\alpha_{\{i,j\}}$ where i and j are two annotators) and compared to overall agreement (α). Formally, the sign s of an edge connecting two annotator-representing nodes i and j is computed as follow:

$$s_{\{i,j\}} = \begin{cases} +, & \text{if } \alpha_{\{i,j\}} \geq \alpha \\ -, & \text{if } \alpha_{\{i,j\}} < \alpha \end{cases}$$

In this paper, I argue that σ, by virtue of approximating the clusterability degree of an annotator graph, will tends to higher values when the annotation is related to more subjective tasks.

4. Experimental validation

σ is tested as a reliable measure of systematic disagreement on a collection of manually annotated datasets. The experiment reported in this section has the goal of verifying the hypothesis that datasets annotated according to subjective phenomena should exhibit higher values of σ by virtue of their disagreement being more systematic. On the contrary, datasets where the agreement is not systematic should exhibit lower σ.

For this experiment, it is crucial to have access to disaggregated datasets, where individual annotations are distributed rather than a single aggregate annotation for each instance. The datasets collected, listed in Section 4.2, have the general form of an $m \times n$ matrix A where m is the number of annotators, n is the number of instances, and $A_{i,j}$ is the label given by annotator i on instance j. A can be fully populated or sparse (typical outcome of crowdsourcing annotation).

4.1. Experimental Setting

All the mainstream IAA measures, and in particular those listed in Section 2 are computed starting from an instance vs. label contingency table. Note that in this process, the identity of the annotators is lost, because the contingency table does not model any relationship between the annotations provided by the same annotator. Therefore, any operation $A_{i,j} := A_{k,j}, A_{k,j} := A_{i,j}$ (swapping two annotations on instance j provided by annotators i and k) results in exactly the same IAA.

This characteristic of IAA measures is exploited to test the hypothesis by generating randomized variations of a dataset that preserve its overall IAA (here computed as α). Firstly, the columns of A are randomly divided into two groups. The rows of the sub-matrix with columns from the first group are shuffled, and so are the rows of the sub-matrix with columns from the second group. The two shuffles are independent from each other. Finally, the entire procedure is repeated ten times, producing the derived annotation matrix A_{rnd}. While it is ensured that $\alpha(A) = \alpha(A_{\text{rnd}})$, the shuffling procedure destroys the systematicity of the annotator agreement. Therefore, if there is a certain degree of systematicity in the original annotation, we should observe $\sigma(A) > \sigma(A_{\text{rnd}})$.

4.2. Data

I collect a number of datasets of varying size and shape, annotated according to different language phenomena, with the only common characteristic of being distributed with disaggregated labels.

I start by retrieving the datasets harmonized in structure and made available by two popular benchmarks for perspectivist classification. From PersEval (Lo et al., 2025), I obtained the following datasets: BREXIT (Akhtar et al., 2020), made of English tweets about Brexit annotated with hate speech (hs), aggressiveness (ag), offensiveness (of), and stereotype (st); MD-Agreement (Leonardelli et al., 2021), English tweets annotated for offensive language; Measuring Hate Speech (Sachdeva et al., 2022), with crowdsourced annotations of hate speech across may targets by a diverse set of annotators; DICES (Aroyo et al., 2024), a collection of human-chatbot conversations annotated for AI safety. The final dataset from PersEval is EPIC (Frenda et al., 2023), containing post-reply pairs annotated for irony, which is replaced by its newer multilingual version MultiPICo (Lo et al., 2024).

I further collected three datasets distributed in the context of the 2025 edition of the Learning with Disagreements challenge (Leonardelli et al., 2025): VariErrNLI (Weber-Genzel et al., 2024) on Natural Language Inference; The Paraphrase Detection dataset released specifically for the challenge, annotated with paraphrastic relation between pairs of questions; Conversational Sarcasm Corpus (CSC) by Jang and Frassinelli (2024), containing short dialogues annotated with perceived sarcasm.

Other three datasets were added to the set selecting from publicly available lists of disaggregated datasets, in particular the Perspectivist Data Manifesto[1] and the Awesome Human Label Variation repository[2]: ConvAbuse (Cercas Curry et al., 2021), a corpus of conversations with AI assistants annotated with abusive language; Tweet Annotation Sensitivity (Kern et al., 2023), made of tweets annotated with hate speech (hs) and offensive language (of); jobQ3MT+ (Liu et al., 2019), a collection of tweets annotated according to the three questions on the interpretation of the job market aspects of the messages (Q1: point of view of job /employment-related information in the target tweet; Q2: employment status of the subject in the tweet; Q3: mention of job/employment transition event in the tweet).

The datasets listed so far have been published with disaggregated labels mostly because they relate to study on annotator disagreement and perspectivist approaches to NLP (Frenda et al., 2025). As a consequence, they mostly cover NLP tasks typically considered subjective, i.e., where the individual perception of the annotator strongly influences their annotation, with the exception of VariErrNLI (natural language inference). Unsurprisingly, analogous corpora annotated for less subjective phenomena are harder to come by. However, I collected three more datasets in this broad category: Frame Disambiguation (Dumitrache et al., 2019) contains crowdsourced annotations for frame disambiguation of sentence-word pairs; Phrase Detectives (Poesio et al., 2019) is a corpus of documents annotated for anaphora with four labels (NR: non-referring; PR: predicative NPs; DN: discourse-new mention; DO: discourse-old mentions). Visual Features (Cheplygina and Pluim, 2018) consists of 100 images from dermoscopic a medical AI challenge, annotated according to four features: asymmetry, border, color, dermoscopic structures.

Table 1 summarizes the datasets along with their size and annotation statistics.

4.3. Results

The result of the metrics computed as defined in Section 4.1 on the dataset listed in Section 4.2 are shown in Table 2. Under my hypothesis, the datasets encoding phenomena whose annotation is more dependent on subjective perception will

[1] https://pdai.info

[2] https://github.com/mainlp/awesome-human-label-variation

Dataset	Instances	Annotators	Avg. annotations per instance (st. dev.)
BREXIT (Akhtar et al., 2020)	1120	6	6.0(0)
BREXIT-hs	"	"	"
BREXIT-ag	"	"	"
BREXIT-of	"	"	"
BREXIT-st	"	"	"
MD-Agreement (Leonardelli et al., 2021)	10753	819	5.0(0)
MHS (Sachdeva et al., 2022)	39565	7912	3.4(26.96)
DICES (Aroyo et al., 2024)	350	123	123.0(0)
MultiPICo (Lo et al., 2024)	18778	506	5.0(1.53)
VariErrNLI (Weber-Genzel et al., 2024)	480	4	3.9(0.89)
Paraphrase (Leonardelli et al., 2025)	500	4	4.0(0)
CSC (Jang and Frassinelli, 2024)	7036	872	4.5(0.89)
ConvAbuse (Cercas Curry et al., 2021)	2894	8	4.4(7.67)
TAS (Kern et al., 2023)	3013	263	4.1(3.09)
TAS-hs	"	"	"
TAS-of	"	"	"
jobQ3MT+ (Liu et al., 2019)	2000	1185	
jobQ3MT+-Q1	"	"	10.06(0.29)
jobQ3MT+-Q2	"	"	10.06(0.29)
jobQ3MT+-Q3	"	"	10.55(0.96)
Frame (Dumitrache et al., 2019)	433	51	21.03(4.08)
Phrase Detectives (Poesio et al., 2019)			
PD-DN	5997	290	10.90(4.38)
PD-DO	3029	326	14.48(9.42)
PD-NR	155	103	7.41(3.22)
PD-PR	1826	282	7.77(3.98)
Visual Features(Cheplygina and Pluim, 2018)	100	6	
VF-asymmetry	"	"	5.93(0.38)
VF-border	"	"	17.86(0.77)
VF-color	"	"	11.90(0.54)
VF-dermo	"	"	23.83(0.97)

Table 1: Statistics of the datasets used in the experimental validation.

show a more systematic structure (σ). The step of randomizing the annotation matrix has the effect of lowering the systematicity (σ_{rnd}) to an extent proportional to the subjectivity of the task. This is confirmed by ordering the result table by $\sigma_{\text{rnd}} - \sigma$ and grouping the tasks into subjective and "objective"[3]. Most of the former datasets at the top of the table (high subjectivity), while the latter ones cluster at the bottom (low subjectivity). Note that at the bottom of the table the statement $\sigma > \sigma_{\text{rnd}}$ does not hold, signaling that the agreement in the annotation of those dataset is not systematic.

The two outliers are arguably justified. In VF-asymmetry, the task requires annotators to judge the symmetry of certain skin formations in medical imagery. The subjectivity of this task founds confirmation in the scientific literature both in the clinical domain (Kunz et al., 2021) and from a computational modeling perspective (Amirshahi et al., 2017), including highlighting correspondences between the task of judging symmetry and the individual perception of human emotions (Evans et al., 2012). The BREXIT-ag dataset has labels of aggressiveness annotated according to the intention to be aggressive, harmful, or even to incite, in various forms, to violent acts against a given target. Examining the examples in the annotation guidelines (Sanguinetti et al., 2018), this task seems to be correlated, at least partially, with overt lexical features (an example contains the work extermination, for instance), which is arguably a more formal rather than subjective task.

[3]"Objective" is written in quotes in this context following the observation of Cabitza et al. (2023) who prefer the term low intersubjective to highlight the difficulty of considering any annotation task completely objective.

4.4. Visual Analysis

In order to gain better insights into the explanatory potential of the new measure, I visualize

Dataset	α	σ	σ_{rnd}	$\sigma_{\text{rnd}} - \sigma$
Paraphrase	.155	1.000	.350	−.650
BREXIT-hs	.347	1.000	.500	−.500
VF-asymmetry	.344	1.000	.550	−.450
CSC	.121	.494	.120	−.374
BREXIT-of	.364	.800	.470	−.330
ConvAbuse	.578	.714	.386	−.328
jobQ3MT+-Q3	.276	.392	.159	−.233
jobQ3MT+-Q2	.353	.427	.201	−.226
MD-Agreement	.359	.494	.283	−.211
jobQ3MT+-Q1	.247	.339	.154	−.185
PD-NR	.085	.416	.264	−.152
VF-dermo	.072	.600	.450	−.150
BREXIT-st	.294	.600	.480	−.120
TAS-hs	.397	.558	.452	−.106
DICES	.210	.601	.518	−.083
MultiPICo	.264	.496	.439	−.057
Frame	.250	.558	.505	−.053
MHS	.516	.677	.637	−.040
TAS-of	.469	.451	.421	−.030
PD-DO	.040	.384	.377	−.007
VF-border	.112	.500	.510	.010
PD-DN	.076	.376	.423	.047
BREXIT-ag	.299	.500	.580	.080
VF-color	.146	.400	.500	.100
PD-PR	−.048	.532	.645	.113
VariErrNLI	.344	.000	.550	.550

Table 2: Results of the experiment described in Section 4.1, in ascending order of difference between original σ and the same metric computed on randomly shuffled datasets (σ_{rnd}). Datasets of tasks traditionally considered "objective" are highlighted with a darker background.

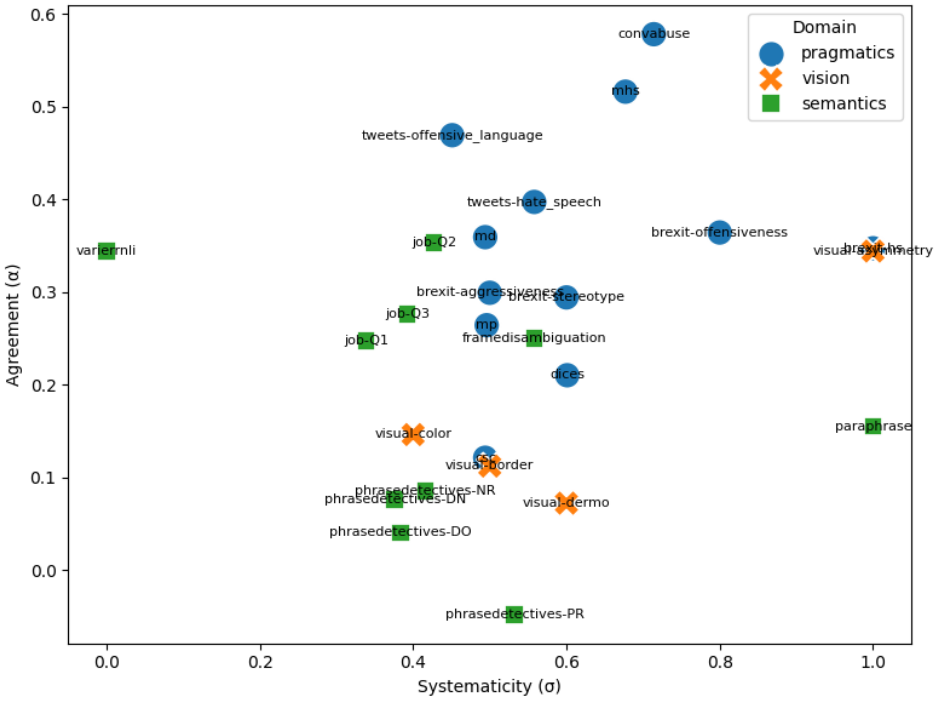


Figure 2: Agreement (α) and systematicity (σ) of the analysed datasets by task domain.

the datasets used in the experiment in a scatterplot (Figure 2). The datasets are coarsely grouped into three domains related to the task they model, namely semantics (VariErrNLI, Paraphrase, jobQ3MT+, Frame, and Phrase Detectives), pragmatics (BREXIT, MD-Agreement, MHS, DICES, ConvAbuse, MultiPICo, CSC, and TAS), and visual (the four VF-* datasets).

Despite a certain amount of variability and the few outliers, the visual analysis shows a clear pattern: the pragmatics dataset generally have a higher agreement (α) than semantics and vision, and the agreement on pragmatics is more systematic (σ).

Besides the visual analysis at the dataset level, Structural Balance Theory provides useful analytical tools for inspecting the inner structure of a dataset annotation. The signed graphs computed as an intermediate steps in the calculation of σ (Section 3) are visualized for some of the datasets involved in the experiment. In the figures, nodes represent annotators, solid lines represent a pairwise agreement above average (i.e., a + edge), and dashed lines represent a pairwise agreement below average (i.e., a - edge).

Figure 3 shows the graph of BREXIT-hs, one of the most subjective datasets according to the analysis. In this graph there are two clear 3-size clusters of annotators, internally connected by + edges and connected with members of the other cluster by - edges. Consequently, the disagreement of the annotators of BREXIT-hs is highly systematic.

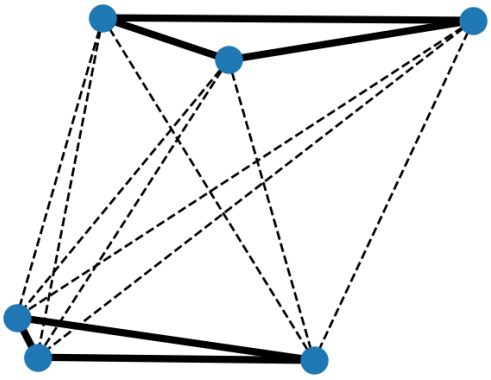


Figure 3: Signed graph of BREXIT-hs.

Figures 4 and 5 show graphs from the annotation of less subjective tasks (VF-color and BREXIT-ag, respectively), which do not exhibit a clustered structure. Interestingly, the distribution of pairwise agreement can vary: while in VF-color 60% of the pairs (9 out of 15) are linked by a +, only 40% of the pairs agree more than the average amount in BREXIT-ag.

Finally, the visualization of the graph of VF-asymmetry reveals a specific structure with a single annotator disagreeing from all the others, who agree among them, possibly contributing to explain the outlier result on this dataset highlighted in Section 4.3.

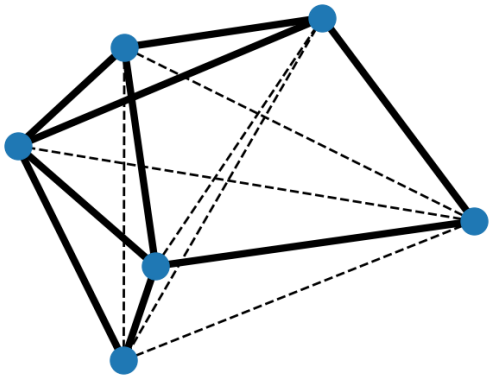


Figure 4: Signed graph of VF-color.

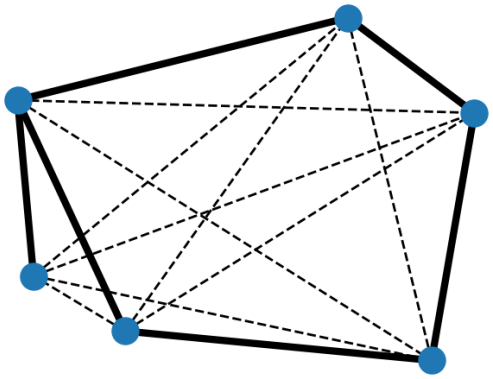


Figure 5: Signed graph of BREXIT-ag.

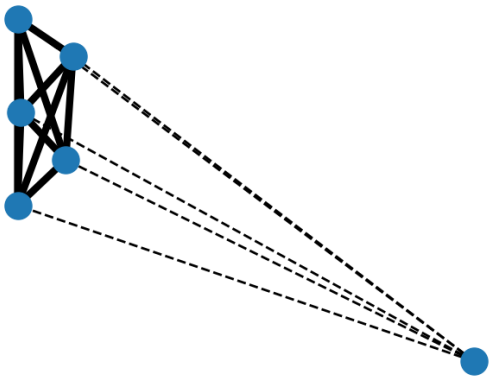


Figure 6: Signed graph of VF-asymmetry.

5. Conclusions

I introduced σ, a quantitative measure of the systematicity of inter-annotator agreement and validated it on a large number of diverse annotated datasets. In particular, the experiment shows that σ captures the systematic patterns of individual annotators and groups, that traditional IAA metrics like Krippenrodffs α are not designed to model.

While in this work σ is validated on a variety of real datasets, its application can be further extended. Intra-annotator agreement could be analyzed under the lens of σ, if data is available with multiple annotations from the same people. Correlations between σ and groups, e.g., by sociodemographics or moral values, are also worth exploring, as well as the impact of persona and perspective-taking prompts in LLM-based annotation.

Besides further tests on more annotated datasets, and extensions to other annotation styles such as rating and ranking, the planned future work also includes the integration of σ into predictive models in order to produce better, more separable representation of the annotators.

6. Limitations

While the experimental section of this paper aims at exploring a wide array of datasets and language phenomena, the resulting figures cannot definitely be grounded in anything other than intuition and the collective consensual experience of a research community. Ironically, there is no objective notion of the subjectivity of a task.

On the practical side, as the number of edges of the graph scales quadratically with the number of nodes, the computational efficiency of σ on a very large dataset with many annotators may dramatically decrease.

Finally, while σ does not need any additional information on the annotators, it relies on Krippendorff's α, which may exhibit an anomalous behavior under particular circumstances (Checco et al., 2017b), a potential limitation already noted by Tsirmpas and Pavlopoulos (2026).

Acknowledgments

I would like to thank the anonymous reviewers who provided valuable insights, some of which made it into the final version. This work also has a debt of gratitude towards Alessandro Mazzei, Daniele Radicioni, and Samuele D'Avenia, for the discussions over the theoretical and experimental aspects which strongly contributed to shape the current version of this paper.

7. Bibliographical 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*, volume 11946 of *Lecture Notes in Computer Science*, pages 588–603. Springer.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 151–154.

Özge Alacam, Sanne Hoeken, Andreas Säuberli, Hannes Gröner, Diego Frassinelli, Sina Zarrieß, and Barbara Plank. 2025. Disentangling subjectivity and uncertainty for hate speech annotation and modeling using gaze. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28707–28724, Suzhou, China. Association for Computational Linguistics.

Seyed Ali Amirshahi, Asha Anoosheh, Stella X. Yu, Jakob Suchan, Carl P. L. Schultz, and Mehul Bhatt. 2017. Symmetry in the eye of the beholder. *Journal of Vision*, 17:300–300.

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2024. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36.

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 and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

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. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, Washington DC, USA.

Dorwin Cartwright and Frank Harary. 1956. Structural balance: a generalization of heider's theory. *Psychological review*, 63 5:277–93.

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.

Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2017a. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2017, 23-26 October 2017, Québec City, Québec, Canada*, pages 11–20. AAAI Press.

Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2017b. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 5, pages 11–20.

Veronika Cheplygina and Josien P. W. Pluim. 2018. Crowd disagreement about medical images is informative. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 105–111, Cham. Springer International Publishing.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

James Allan Davis. 1967. Clustering and structural balance in graphs. *Human Relations*, 20:181 – 187.

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.

Anca Dumitrache, Oana Inel, Lora Aroyo, Benjamin Timmermans, and Chris Welty. 2018. Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement (short paper). In *Proceedings of the 1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management (SAD 2018 and CrowdBias 2018) co-located the 6th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2018), Zürich, Switzerland, July 5, 2018*, volume 2276 of *CEUR Workshop Proceedings*, pages 11–18. CEUR-WS.org.

David W. Evans, Patrick T. Orr, Steven M. Lazar, Daniel Breton, Jennifer Gerard, David H. Ledbetter, Kathleen Janosco, Jessica Dotts, and Holly Batchelder. 2012. Human preferences for symmetry: Subjective experience, cognitive conflict and cortical brain activity. *PLOS ONE*, 7(6):1–9.

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

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: Multiperspective 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.

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.

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.

Felix Kunz, Matthias Hirth, Tilmann Schweitzer, Christian Linz, Bernhard Goetz, Angelika Stellzig-Eisenhauer, Kathrin Borchert, and Hartmut

Böhm. 2021. Subjective perception of craniofacial growth asymmetries in patients with deformational plagiocephaly. *Clinical oral investigations*, 25(2):525–537.

Elisa Leonardelli, Silvia Casola, Siyao Peng, Giulia Rizzi, Valerio Basile, Elisabetta Fersini, Diego Frassinelli, Hyewon Jang, Maja Pavlovic, Barbara Plank, and Massimo Poesio. 2025. LeWiDi-2025 at NLPerspectives: Third edition of the learning with disagreements shared task. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 182–195, Suzhou, China. Association for Computational Linguistics.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher M. Homan. 2019. Learning to Predict Population-Level Label Distributions. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 68–76.

Soda Marem Lo and Valerio Basile. 2023. Hierarchical clustering of label-based annotator representations for mining perspectives. In *Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP co-located with 26th European Conference on Artificial Intelligence (ECAI 2023), Kraków, Poland, September 30th, 2023*, volume 3494 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Soda Marem Lo, Silvia Casola, Simona Frenda, 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.

Soda Marem Lo, Silvia Casola, Erhan Sezerer, Valerio Basile, Franco Sansonetti, Antonio Uva, and Davide Bernardi. 2025. PERSEVAL: A framework for perspectivist classification evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22345–22370, Suzhou, China. Association for Computational Linguistics.

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.

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.

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.

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 at LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the 11th Conference on Language Resources and Evaluation (LREC2018), May 2018, Miyazaki, Japan*, pages 2798–2895.

William A. Scott. 1955. Reliability of content analysis:the case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.

Dimitris Tsirmpas and John Pavlopoulos. 2026. Quantifying and attributing polarization to annotator groups.

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.

Fine-Grained Perspectives: Modeling Explanations with Annotator-Specific Rationales

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

Abstract

Beyond exploring disaggregated labels for modeling perspectives, annotator rationales provide fine-grained signals of individual perspectives. In this work, we propose a framework for jointly modeling annotator-specific label prediction and corresponding explanations, fine-tuned on the annotators' provided rationales. Using a dataset with disaggregated natural language inference (NLI) annotations and annotator-provided explanations, we condition predictions on both annotator identity and demographic metadata through a representation-level User Passport mechanism. We further introduce two explainer architectures: a post-hoc prompt-based explainer and a prefixed bridge explainer that transfers annotator-conditioned classifier representations directly into a generative model. This design enables explanation generation aligned with individual annotator perspectives. Our results show that incorporating explanation modeling substantially improves predictive performance over a baseline annotator-aware classifier, with the prefixed bridge approach achieving more stable label alignment and higher semantic consistency, while the post-hoc approach yields stronger lexical similarity. These findings indicate that modeling explanations as expressions of fine-grained perspective provides a richer and more faithful representation of disagreement. The proposed approaches advance perspectivist modeling by integrating annotator-specific rationales into both predictive and generative components.

Keywords: Explanation, Perspectives, Annotator

1. Introduction

Perspectivist NLP argues that annotations should reflect the specific judgments of individual annotators rather than converge on a single consensus label (Pavlick and Kwiatkowski, 2019). In tasks such as natural language inference (NLI), stance detection, and hate speech classification (Xu et al., 2024), it is legitimate for annotators to disagree due to differences in background, interpretation, or socio-demographic perspective. Modeling such disagreement has become an important focus of recent shared tasks (Leonardelli et al., 2023; Uma et al., 2021) and research initiatives, shifting the emphasis away from majority voting toward preserving variation.

Most perspectivist approaches implicitly model perspectives using linguistic and contextual signals such as sociodemographic information, user IDs, and group affiliations (Davani et al., 2022; Plepi et al., 2022). These signals are used to infer the potential sources of variation and diversity in annotations. However, beyond disaggregated labels, annotators' perspectives often remain abstracted and only indirectly represented in the model. Although some datasets require annotators to provide rationales for selecting a particular label (Weber-Genzel et al., 2024), these explanations are rarely integrated explicitly into perspectivist modeling. Incorporating annotator rationales enables more nuanced and fine-grained representations of perspective.

Explainability in perspectivist approaches has gradually emerged as a critical component of trustworthy NLP systems, as it supports model-level interpretability through the analysis of attention patterns or internal structures used to justify predictions (Mastromattei et al., 2022a). In recommendation systems, for example, natural language generation (NLG) methods have been proposed to generate flexible, free-text explanations based on user-generated content (Li et al., 2021b). While such approaches demonstrate the potential of generative models to produce fluent and varied explanations, they also expose limitations: generated content may be off-topic, insufficiently grounded in the input, repetitive Li et al. (2021a), or insufficiently personalized. These challenges highlight the need for controllable and faithful explanation generation, particularly when explanations are expected to reflect specific user or annotator viewpoints.

Within perspectivist NLP, explainability has been approached in different ways. Some studies treat it as post-hoc model interpretation Mastromattei et al. (2022a), identifying linguistic features or structural patterns that influence perspective-aware predictions (Muscato et al., 2025). Others rely on prompting strategies to simulate user perspectives in large language models (Hayati et al., 2024). However, relatively little work has explicitly modeled annotator-

Proceedings of NLPerspectives @ LREC 2026, pages 66–75
12 May 2026.

specific explanations alongside disaggregated labels, partly because few datasets contain both disagreement and distinct rationales.

In this study, we integrate perspectivist modeling with perspectivist explanation by explicitly conditioning explanation generation on annotator-specific representations. Using a dataset with disaggregated NLI labels and annotator-provided explanations, we model perspective in both label prediction and rationale generation. We explore a prompt-based post-hoc explainer and a representation-prefix bridge that transfers classifier representations enriched with annotator information into a generative model. In doing so, we treat explanations as expressions of perspective rather than merely post-hoc justifications of a model's decisions.

2. Related works

Perspectivist NLP aims to preserve the nuanced information hidden within disagreement by modeling annotator-specific labels rather than aggregating them into a single label (Cabitza et al., 2023). However, explainability within this paradigm remains relatively understudied and fragmented (Frenda et al., 2025). Existing research primarily approaches explainability either through model interpretability as in Mastromattei et al. (2022a) or by explicitly prompting Large Language Models (LLMs) for explanations (Orlikowski et al., 2025). However, most work has yet to explore annotator-specific rationales grounded in internal representations as a primary approach for perspectivist explainability.

2.1. Current Approaches to Perspective-Aware Explanations

One line of research addresses explainability in perspectivist models by identifying the linguistic components in Hate speech tasks with the use of recognizers that incorporates syntactic dependency trees to provide post-hoc justifications for classifications (Mastromattei et al., 2022a). In these instances, explainability focuses on revealing the mechanics of the model's prediction rather than capturing the annotator's subjective reasoning. Similarly, Mastromattei et al. (2022b) explored explainable syntax-based models within hate speech detection to identify trigger words that influence target classification. In a different vein, Nirmal et al. (2024) implicitly extracted user rationales from input text using LLMs to guide classifier outcomes, aiming for a more interpretable architectural framework.

2.2. Personalized Generation and Recommendation

A shift toward personalized explanation is evident in the work of Li et al. (2021b), who designed a specialized Transformer for explainable recommendation. This model utilizes user IDs and items alongside linguistic cues to generate recommendations and justifications that reflect individual user interests. Similarly, Li et al. (2020) utilized a neural template approach to address user ratings within recommender systems. More recently, Plepi et al. (2024) introduced twin-encoder architectures that separately encode auxiliary user information to facilitate perspective-taking in conflict situations. This allows the model to conceptualize user viewpoints through self-disclosure statements. While this approach structurally integrates user context, it does not explicitly disentangle annotator-specific explanatory reasoning in disaggregated datasets, where annotators might agree on a label but diverge significantly in their underlying logic. In this study, we address explainability through the lens of annotator rationales, seeking to understand the *why* behind a label from the human's perspective. Our approach models annotator perspectives at both the classification and explanation levels. Furthermore, we introduce a representation-level bridge that conditions explanation generation directly on annotator-specific internal representations. By doing so, we treat explanation not merely as a post-hoc interpretability tool, but as an explicit expression of annotators perspectives tied directly to disaggregated labels they represent.

3. Methods and data

We study perspectivism in generative explainability using the VariErrNLI dataset (Weber-Genzel et al., 2024), which contains disaggregated annotator labels and annotator-specific rationales. Unlike most existing disaggregated datasets, VariErrNLI preserves both label disagreement and explanation diversity, making it suitable for modeling fine-grained perspectives.

Our framework consists of two components: (i) an annotator-aware classifier that predicts label sets for each annotator, and (ii) an annotator-conditioned explainer that generates corresponding rationales. We explicitly model annotator identity using learned embeddings and metadata features, which are fused with the contextual representation of the input (context and statement) to produce annotator-specific predictions.

We compare two explanation approaches. The first is a post-hoc, prompt-based explainer that generates explanations from textual inputs. The second is a prefixed bridge explainer that conditions

generation on the classifier's internal annotator-specific representations. This allows the model to incorporate both predicted labels and underlying annotator-specific reasoning signals.

3.1. VariErrNLI Dataset

We use VariErrNLI (Variation vs. Error), a perspectivist NLI dataset designed to disentangle human label variation from annotation error. VariErrNLI contains approximately 500 NLI items sampled from ChaosNLI (MNLI subset) and annotated in two rounds by four independent annotators.

In Round 1, annotators assigned one or more NLI labels, Entailment (E), Neutral (N), or Contradiction (C) to each item and provided a one-sentence explanation for each label assigned, preserving fine-grained reasoning diversity. This round of annotation produced 1,933 label-explanation pairs.

In Round 2, annotators independently evaluated the validity of each label–explanation pair (including their own) by judging whether the explanation plausibly supports the assigned label. This second stage enables distinguishing plausible human label variation from annotation errors. The dataset, therefore, provides not only disaggregated labels and rationales but also meta-judgments about their validity.

Although VariErrNLI was originally designed to study annotation error versus variation, we use it for a different purpose. Specifically, we leverage its disaggregated labels and annotator-specific explanations to model and generate annotator-conditioned reasoning. For this study, we use the version released for the Learning with Disagreement (LeWiDi) 2025 Shared Task (Leonardelli et al., 2026), which provides predefined training, development, and test splits. The dataset statistics are presented in Table 1

3.2. Problem Formulation

We formalize annotator-specific prediction and explanation as a joint task. Each instance in the VariErrNLI dataset consists of a context c, a statement s, and annotations from annotators $a \in \mathcal{A}$. Each annotator provides a judgment (in some instances, multi-label) over the label set

$$\mathcal{L} = \{C, E, N\}, \tag{1}$$

corresponding to contradiction (C), entailment (E), and neutral (N); and an explanation that justifies their labeling decision. For each annotator a, there is an annotator-specific label

$$y_a \subseteq \mathcal{L}, \tag{2}$$

and an associated explanation r_a, where r_a is a short sentence describing the reasoning for y_a. Because two annotators can assign the same label for different reasons, we treat explanation generation as an explicitly perspectivist problem. Our model therefore has two goals: (i) predict the annotator-specific label set for each annotator a, and (ii) generate the corresponding annotator explanation r_a, which we define as the annotator's expressed perspective. For each instance (c, s) and annotator a, we learn an annotator-aware classifier and an annotator-conditioned explainer trained on the provided human rationales.

3.3. Annotator-Aware Classification

We implement the *User Passport* method to explicitly model annotator-specific perspectives within our classification framework (Sarumi et al., 2025), using DeBERTa-v3-base as the backbone encoder. This approach incorporates annotator identity and metadata directly at the representation level rather than through input text modification or token-based methods (Welch et al., 2022). The resulting classifier serves as the underlying prediction component for both the post-hoc and prefixed bridge explanation models.

Formally, we consider an annotated dataset defined by $\mathcal{D} = (X, A, Y)$, where X is the set of text instances $\{x_1, x_2, \ldots, x_n\}$. Each instance $x_i \in \mathcal{X}$ is a pair (c_i, s_i) representing the context and statement. The set $A = \{a_1, a_2, \ldots, a_k\}$ represents unique annotators, and the annotation matrix is defined as:

$$Y : X \times A \to \{0, 1\}^3 \tag{3}$$

To handle varying annotator coverage, a masking mechanism is applied during training and evaluation. The annotator-level loss is computed only for instances where a label exists, using a binary mask to ensure missing annotations do not contribute to the training objective.

The encoder extracts a pooled representation $h \in \mathbb{R}^H$ capturing the relationship between c_i and s_i. To incorporate individual variation, we define a learnable embedding space where each annotator a_j is mapped to a unique, d-dimensional vector $u_j \in \mathbb{R}^E$:

$$u_j = \text{Embedding}(a_j) \tag{4}$$

Simultaneously, each annotator's structured demographic metadata is transformed into a fixed-size vector m_j and projected into the latent space of the text encoder. We then perform a *representation-level fusion* by concatenating the instance representation, the annotator embedding, and the metadata projection. The resulting fused representation z_{ij} is passed to the classification head:

$$z_{ij} = [\, h \,;\, u_j \,;\, m_j] \tag{5}$$

This allows the model to explicitly account for both the annotator's identity and their demographic context by learning systematic patterns between these

Statistic	Train	Dev	Test	Total
Split-level statistics				
Instances	388	50	50	488
Annotators	4	4	4	4
Annotations	1,505	187	199	1,891
Avg. annotations / instance	3.88	3.74	3.98	3.88
Explanations	1,505	187	199	1,891
Avg. explanation length (words)	13.90	13.12	14.28	13.86
Label distribution (count, %)				
Entailment	446 (29.6%)	34 (18.2%)	61 (30.7%)	541 (28.6%)
Neutral	767 (51.0%)	96 (51.3%)	93 (46.7%)	956 (50.6%)
Contradiction	292 (19.4%)	57 (30.5%)	45 (22.6%)	394 (20.8%)
Annotations per annotator (count) and demographics				
Ann1 (F,22,CN,MSc)	367	45	47	459
Ann2 (M,33,DE,Postdoc)	376	45	47	468
Ann3 (F,25,CN,MSc)	379	46	54	479
Ann4 (M,25,CN,MSc)	383	51	51	485

Table 1: VariErrNLI dataset statistics by split. Demographics are abbreviated as Gender, Age, Nationality, Education (CN=Chinese, DE=German; MSc=Master student).

features and labeling behavior through latent feature fusion.

3.4. Annotator Explanation Modeling

To generate annotator-specific rationales, we implement two explanation approaches that produce an explanation r_a but differ in how they incorporate classifier information.

3.4.1. Post-hoc Explainer

Our first approach trains a standard encoder-decoder model $Flan-T5$ (Chung et al., 2024) to generate an annotator explanation using a text-only prompt. For each training record, we construct an input prompt that contains: the context and statement, the annotator's gold labels, annotators persona: derived from the annotator metadata information, and an annotator control token. The annotator control token is linked to the annotator ID in the dataset and prepended to the prompt by extending the tokenizer vocabulary with a unique, learnable special token (Sarumi et al., 2024; Plepi et al., 2022). At inference time, we insert the classifier's predicted probabilities (p_C, p_E, p_N) into the prompt. The explainer then generates a short explanation. In this setup, there is no differentiable connection between the classifier and the explainer.

3.4.2. Prefixed Bridge Explainer

Our second approach introduces a stronger coupling between classification and explanation using the classifier's continuous internal representation, rather than text-only features. We first run the annotator-aware classifier on (c, s) to obtain the fused representation z_{ij}. We then learn a small neural Prefixed Bridge (a 2-layer MLP) that projects this vector into a sequence of prefix embeddings with the same dimensionality as the T5 encoder embedding space. These prefix embeddings are prepended to the T5 encoder input embeddings before encoding. We then train by freezing the classifier parameters and optimizing the bridge and the T5 parameters to minimize explanation generation loss. At inference time, explanation generation is performed using the prefix produced by the bridge, which is concatenated with the prompt token embeddings before encoding and generation (see Figure 1).

4. Experiments

In our experiments, we used two base models that follow the encoder-decoder architecture. We also implemented the User Passport method for incorporating annotator-meta information.

4.1. Experimental set-up

We train the annotator-aware classifier for 50 epochs using the AdamW optimizer with a learning rate of 2×10^{-5} and weight decay 0.01. A linear scheduler with warmup (ratio 0.06), gradient clipping (max norm 1.0), and early stopping on development macro-F1 (patience 3) is applied. The backbone model is DeBERTa-v3-base (He et al., 2023), with a maximum input length of 256 and batch size 32. To model annotator-specific predictions, we incorporate annotator information through a learnable annotator embedding (dimension 64) and a projected metadata representation, fused

Explainer	F1 (Macro)	Exact Match	ROUGE-L	Semantic Similarity
User Passport (Sarumi et al., 2025)	70.5	—	—	—
Post-hoc Explainer	92.3	92.2	**24.5**	51.0
Prefixed Bridge Explainer	**93.9**	**92.4**	24.0	**53.4**

Table 2: Aggregated evaluation scores across all annotators. We report the results of the User Passport model from previous work, without explanation, as the baseline. Bold values indicate the best scores. All scores are reported as the mean of three runs.

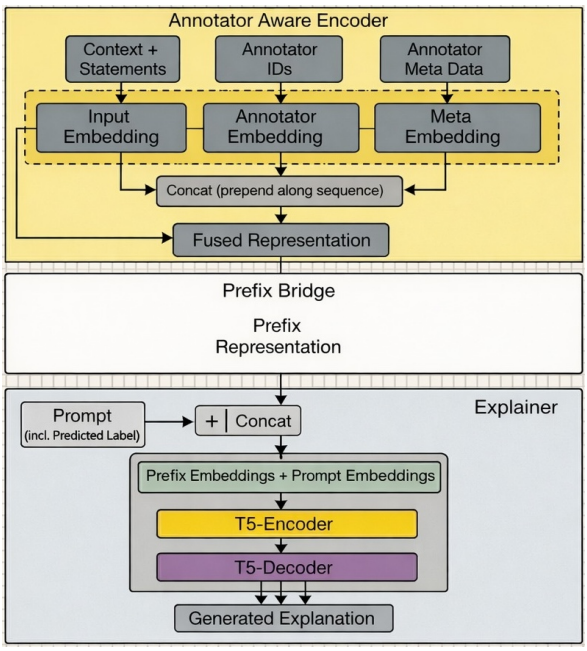


Figure 1: The Prefixed Bridged Explainer

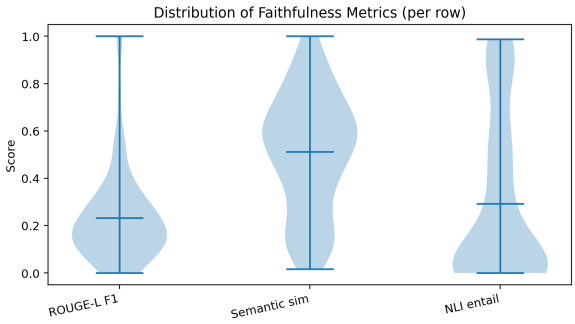


Figure 2: Prefixed Bridged Faithfulness Evaluation

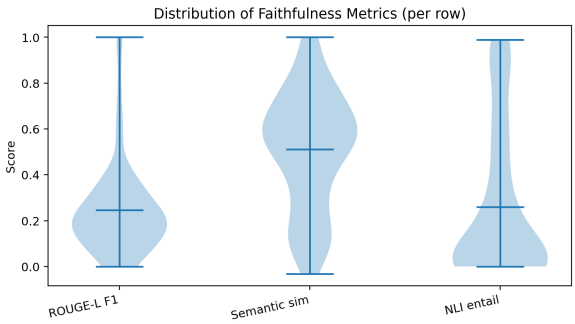


Figure 3: Post-hoc Faithfulness Evaluation

with the instance representation at the feature level. Training uses masked binary cross-entropy with an auxiliary soft-label alignment objective ($\lambda_{\text{soft}} = 1.0$). Class imbalance is handled using masked focal BCE with class-specific positive weighting.

For explanation generation, both explainer variants are trained using Flan-T5-base with a maximum input length 512 and target length 128. Models are trained for up to 50 epochs with early stopping on validation loss, using AdamW with learning rate 8×10^{-5} and weight decay 0.01. Label thresholds are tuned on the development set with grid search over $[0.1, 0.9]$, selecting the configuration that maximizes mean Jaccard similarity with gold annotator label-sets.

All experiments are conducted on a single NVIDIA A100 80GB PCIe GPU (CUDA 13.1). Average end-to-end runtime (training and evaluation) is approximately 15-20 minutes per model. All reported results are averaged over three runs.

5. Result and Discussion

Aggregated Evaluation of Explainers

Table 2 presents the aggregated performance comparison between the baseline annotator-aware classifier (User Passport) from previous work, the Post-hoc Explainer, and the Prefixed Bridge Explainer. The baseline achieves a Macro-F1 score of 70.5, indicating that incorporating explanation modeling substantially improves classification performance.

Both explanation-based approaches outperform the baseline, with the Prefixed Bridge Explainer achieving the highest Macro-F1 (93.9) and Exact Match (92.4), indicating stronger agreement with the gold labels. The Post-hoc Explainer also performs well with Macro-F1(92.3), but remains slightly below the bridge model.

In terms of explanation quality, ROUGE-L is marginally higher for the Post-hoc Explainer, suggesting better lexical overlap with reference explanations, which is consistent with its text-based ap-

Prefixed Bridge Explainer								
Annotator	Gender	Age	Nationality	Education	Macro F1	Exact Match	ROUGE-L	Semantic Sim
Ann1	Female	22	Chinese	MSc.	**94.3**	**94.2**	23.8	55.2
Ann2	Male	33	German	Postdoc	92.5	92.0	**34.1**	**59.6**
Ann3	Female	25	Chinese	MSc	92.0	88.7	21.2	53.5
Ann4	Male	25	Chinese	MSc	**95.5**	**94.7**	17.9	45.8
Post-hoc Explainer								
Annotator	Gender	Age	Nationality	Education	Macro F1	Exact Match	ROUGE-L	Semantic Sim
Ann1	Female	22	Chinese	MSc.	87.9	90.6	24.5	49.9
Ann2	Male	33	German	Postdoc	**96.7**	**97.1**	**31.3**	**55.9**
Ann3	Female	25	Chinese	MSc	90.4	88.7	22.9	50.0
Ann4	Male	25	Chinese	MSc	92.4	92.7	19.6	48.7

Table 3: Comparison of Explainers per annotator: A descriptive Analysis. Bold values highlight key patterns discussed in section 5: improved predictive performance with the bridge model (Ann1, Ann4), stronger lexical overlap and semantic strength with both prefixed and post-hoc model (Ann2), and overall Macro-F1 score, Exact Match (notably Ann2).

proach. In contrast, the Prefixed Bridge Explainer achieves higher semantic similarity, indicating that its generated explanations are better aligned in meaning. This improvement can be attributed to its use of the classifier's internal representations, which provide richer contextual features for generation.

These results show that explanation modeling significantly improves performance over the baseline, while tighter integration between prediction and generation further enhances classification consistency and semantic alignment.

Faithfulness Distribution and Qualitative Analysis

The faithfulness distributions in Figures 2 and 3 show that, for both models, semantic similarity scores cluster around moderate values (median $\sim$ 0.5), while ROUGE-L remains relatively low (median $\sim$ 0.24), indicating lexical divergence despite semantic alignment. However, the Prefixed Bridge Explainer exhibits a more balanced NLI entailment distribution, with a larger proportion of high-entailment cases compared to the Post-hoc model, suggesting stronger inferential alignment between predictions and explanations. To further examine these differences, we present qualitative examples in Figures 4 and 5, focusing on cases where the predicted label is consistent but the generated explanations differ in structure and depth. In both examples, the two models correctly identify that the context supports investment in information technology rather than the financial sector. However, the nature of the generated explanations differs. The Prefixed Bridge Explainer produces explanations that are more concise and directly grounded in the key contrast between the context and the statement, closely mirroring the underlying reasoning required for the prediction. In contrast, the Post-hoc Explainer tends to generate more verbose explanations, introducing additional statements that are not explicitly stated in the context. While both explanations are semantically aligned with the gold rationale, the prefixed bridged explanation more precisely captures the core inference without introducing random words. This qualitative difference supports the distributional findings. The Prefixed Bridge Explainer demonstrates stronger alignment between prediction and explanation, not only quantitatively through higher scores, but also qualitatively in the clarity and focus of the generated explanation.

Comparison of Explainers per annotator: A descriptive Analysis.

Table 3 presents a detailed examination across the four annotators. Differences are observed in Macro-F1, Exact Match, ROUGE-L, and Semantic similarity, suggesting that both models interact differently with individual annotator patterns.

The Prefixed Bridge Explainer generally produces more stable performance across annotators in terms of Macro-F1 and Exact Match. In particular, Ann1 and Ann4 show improvements in both metrics compared to the Post-hoc Explainer, indicating that incorporating classifier-level representations contributes to more reliable alignment between predictions and annotator-specific labels. This suggests that the shared representation between the two encoders better captures variability in annotator decision patterns, especially when explanations differ in structure or clarity.

Ann2 achieves the highest overall performance, particularly under the Post-hoc Explainer (Macro-F1: 96.7, Exact Match: 97.1), outperforming the Prefixed Bridge model. Ann3 and Ann4 exhibit comparatively lower or more variable performance across certain metrics, particularly in ROUGE-L. For Ann4, while Macro-F1 and Exact Match improve under the Prefixed Bridge Explainer, ROUGE-

```
"statement": "Investment in the financial sector explains most of the acceleration
in labor productivity.",
    "context": "Indeed, recent economic research suggests that investment in
information technology explains most of the acceleration in labor productivity
growth-a major component of overall economic growth-since 1995.",
    "gen_expl": "The context suggests that investment in information technology
explains most of the acceleration in labor productivity growth, which suggests that
the financial sector is not involved in this explanation. We don't know whether the
investment in the financial sector is involved in this explanation.",
    "gold_expl": "Investment in information technology explains most of the
acceleration in labor productivity, not investment in financial sector.\nWe don't
know whether investment in information technology is a subsector of investment in
financial sector, or reverse."
  },
```


Figure 4: The Prefixed Bridged Explanation Example


```
"statement": "Investment in the financial sector explains most of the
acceleration in labor productivity.",
    "context": "Indeed, recent economic research suggests that investment
in information technology explains most of the acceleration in labor
productivity growth-a major component of overall economic growth-since
1995.",
    "gen_expl": "The context suggests that investment in information
technology explains most of the acceleration in labor productivity growth,
not the financial sector. The model probabilities are (0.25, 0.22, 0.55)",
    "gold_expl": "The reason of the acceleration in labor productivity is
the investment in information technology, not in the financial sector."
  },
```


Figure 5: The Post-hoc Explanation Example

L and Semantic similarity remain relatively low across both models.

A closer examination of the VariErrNLI dataset (Weber-Genzel et al., 2024) provides important context for interpreting these results. The dataset explicitly distinguishes between variation and annotation error through a second round of self- and peer-validation, where explanations are assessed for whether they plausibly support the assigned labels. As shown in the original study, agreement increases substantially after validation, indicating that a portion of annotator disagreement is attributable not to genuine perspectives and differences, but to inconsistencies and errors.

This distinction is reflected in our findings. Annotators whose explanations are more consistently grounded in the input text and validated by peers are more reliably modeled by both approaches. In

particular, Ann2 achieves the highest predictive performance across metrics, especially under the Post-hoc Explainer, aligning with the dataset's validation framework where more coherent and text-aligned reasoning leads to more stable label–explanation pairs. Notably, Ann2 is also the annotator with the highest age (33) and level of education (Postdoc) in the dataset. While this may be associated with clearer or more structured explanations, stronger task understanding or domain expertise, we do not draw definitive conclusions from this observation due to the limited number of annotators. Instead, this serves as an indicative pattern that can be further investigated in larger and more controlled settings.

In contrast, annotators exhibiting more variability in explanation quality are more challenging to model. For example, Ann4 shows comparatively lower or less consistent performance across certain metrics, particularly in lexical overlap (ROUGE-L), despite improvements in predictive performance under the Prefixed Bridge Explainer. This pattern is consistent with the dataset observations, where some explanations may be less well-aligned with the assigned labels or expressed in ways that deviate from reference formulations. As a result, the model relies more heavily on underlying representations rather than surface-level cues.

A consistent pattern across annotators is the divergence between lexical and semantic metrics. The Post-hoc Explainer tends to produce higher ROUGE-L scores, indicating closer surface-level similarity to reference explanations. In contrast, the Prefixed Bridge Explainer achieves higher or comparable semantic similarity across most annotators, suggesting better alignment in meaning. This reflects the underlying modeling difference: the Post-hoc approach relies primarily on textual prompts, whereas the bridge model leverages classifier-derived internal representations, enabling richer contextual grounding of explanations.

These per-annotator differences highlight that identical labels do not imply identical reasoning processes. The variation observed across metrics suggests that annotators may express similar decisions through different explanatory structures, levels of detail, or linguistic forms. By incorporating explanation generation, both models move beyond label prediction and provide additional insight into how annotator perspectives are represented. The Prefixed Bridge Explainer, in particular, better preserves the relationship between predictions and underlying reasoning, especially when explanations are less consistent in form.

It is important to note that these observations are based on a small number of annotators, with limited demographic diversity and a relatively small test set. As such, we do not perform statistical significance testing and instead rely on descriptive analysis. The patterns observed should therefore be interpreted as indicative trends rather than generalizable findings.

Overall, the per-annotator analysis suggests that incorporating explanation modeling improves the representation of annotator perspectives, and that tighter integration between prediction and explanation, as in the Prefixed Bridge Explainer, provides more consistent and semantically aligned outputs across diverse annotator behaviors.

6. Conclusion

This work demonstrates the importance of Modeling fine-grained annotator perspectives jointly with explanation generation in natural language inference. Rather than treating explanations as post-hoc rationalizations, we show that integrating annotator-expressed rationales into the predictive architecture enables more robust modeling of human diversity. By leveraging explanation-level supervision tied to individual annotations, the model captures not only label outcomes but also the reasoning patterns underlying them, allowing for more faithful representation of disagreement and interpretative nuance.

Methodologically, we implement an encoder-to-encoder bridge architecture that explicitly connects prediction and explanation modules. This structural coupling enables the model to condition its explanatory representations on the same signals that drive classification decisions, thereby improving macro-level stability and inferential alignment across annotators. Our results show that Modeling perspectives through annotator rationales strengthens semantic consistency and predictive robustness, particularly in semantically complex categories. Overall, this work highlights the value of integrating explanation modeling into annotator-aware architectures for developing more transparent and perspective-sensitive NLP systems.

7. Limitation

A primary limitation of this work is the dataset used, which was originally constructed to investigate annotation errors in human label variation. Although the inclusion of annotator-specific rationales represents a substantial step toward preserving individual reasoning patterns, the explanations were not designed to systematically capture controlled variations in demographic, linguistic, or cultural background, but were instead targeted toward annotation error detection. As a result, the scope remains limited, which may constrain the generalizability of the proposed encoder-to-encoder bridge framework.

We initially proposed extending the ChaOSNLI instances used in VariErrNLI with explanations written by native English speakers to systematically examine how bilingual versus native annotator rationales affect modeling outcomes. This extension would allow a more controlled investigation of linguistic background effects on explanation faithfulness and predictive performance. Future work will focus on expanding the dataset in this direction to strengthen the empirical foundation of perspective modeling.

Additionally, an ensemble of the Post-hoc and Prefixed Bridge approaches presents an interesting direction for future work, as it could leverage the strengths of the individual models to produce a more well-rounded output.

All code and resources developed for this study are publicly available[1] to facilitate reproducibility and further research.

Bibliographical References

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.

Aida Mostafazadeh Davani, Markos Markatou, Tommaso Fornaciari, Silviu Paun, Dirk Hovy, Joel Tetreault, and Cecilia Ovesdotter Alm. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

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

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2024. How far can we extract diverse perspectives from large language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5336–5366, Miami, Florida, USA. Association for Computational Linguistics.

[1] https://github.com/Responsible-NLP/LRECNLPerspectives2026-Fine-Grained-Perspective

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.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021a. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics.

Lei Li, Yongfeng Zhang, and Li Chen. 2020. Generate neural template explanations for recommendation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*, pages 755–764, New York, NY, USA. ACM.

Lei Li, Yongfeng Zhang, and Li Chen. 2021b. Personalized transformer for explainable recommendation. 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 4947–4957, Online. Association for Computational Linguistics.

Matteo Mastromattei, Leonardo Ranaldi, Francesca Fallucchi, and Fabio Massimo Zanzotto. 2022a. Syntax and prejudice: ethically-charged biases of a syntax-based hate speech recognizer unveiled. *PeerJ Computer Science*, 8:e859.

Michele Mastromattei, Valerio Basile, and Fabio Massimo Zanzotto. 2022b. Change my mind: How syntax-based hate speech recognizer can uncover hidden motivations based on different viewpoints. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 117–125, Marseille, France. European Language Resources Association.

Benedetta Muscato, Lucia Passaro, Gizem Gezici, and Fosca Giannotti. 2025. Perspectives in play: A multi-perspective approach for more inclusive nlp systems. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI-2025, page 9827–9835. International Joint Conferences on Artificial Intelligence Organization.

Ayushi Nirmal, Amrita Bhattacharjee, Paras Sheth, and Huan Liu. 2024. Towards interpretable hate speech detection using large language model-extracted rationales. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 223–233, Mexico City, Mexico. Association for Computational Linguistics.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions.

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.

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, and Daniel Braun. 2025. NLP-ResTeam at LeWiDi-2025:performance shifts in perspective aware models based on evaluation metrics. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 219–227, Suzhou, China. 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.

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.

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.

Language Resource References

Chung, Hyung Won and Hou, Le and Longpre, Shayne and Zoph, Barret and Tai, Yi and Fedus, William and Li, Yunxuan and Wang, Xuezhi and Dehghani, Mostafa and Brahma, Siddhartha and Webson, Albert and Gu, Shixiang Shane and Dai, Zhuyun and Suzgun, Mirac and Chen, Xinyun and Chowdhery, Aakanksha and Castro-Ros, Alex and Pellat, Marie and Robinson, Kevin and Valter, Dasha and Narang, Sharan and Mishra, Gaurav and Yu, Adams and Zhao, Vincent and Huang, Yanping and Dai, Andrew and Yu, Hongkun and Petrov, Slav and Chi, Ed H. and Dean, Jeff and Devlin, Jacob and Roberts, Adam and Zhou, Denny and Le, Quoc V. and Wei, Jason. 2024. *Scaling instruction-finetuned language models*. JMLR.org.

Pengcheng He and Jianfeng Gao and Weizhu Chen. 2023. *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing*.

Elisa Leonardelli and Silvia Casola and Siyao Peng and Giulia Rizzi and Valerio Basile and Elisabetta Fersini and Diego Frassinelli and Hyewon Jang and Maja Pavlovic and Barbara Plank and Massimo Poesio. 2026. *LeWiDi-2025 at NLPerspectives: Third Edition of the Learning with Disagreements Shared Task*.

Weber-Genzel, Leon and Peng, Siyao and De Marneffe, Marie-Catherine and Plank, Barbara. 2024. *VariErr NLI: Separating Annotation Error from Human Label Variation*. Association for Computational Linguistics.

Structured Disagreement in Health-Literacy Annotation: Epistemic Stability, Conceptual Difficulty, and Agreement-Stratified Inference

Olga Kellert*, **Sriya Kondury, Candice Koo, Nemika Tyagi, Steffen Eikenberry**
Arizona State University
{Olga.Kellert, skondury, ckoo4, ntyagi8, seikenbe}@asu.edu

Abstract

Annotation pipelines in Natural Language Processing (NLP) commonly assume a single latent ground truth per instance and resolve disagreement through label aggregation. Perspectivist approaches challenge this view by treating disagreement as potentially informative rather than erroneous. We present a large-scale analysis of graded health-literacy annotations from 6,323 open-ended COVID-19 responses collected in Ecuador and Peru. Each response was independently labeled by multiple annotators using proportional correctness scores, reflecting the degree to which responses align with normative public-health guidelines, allowing us to analyze the full distribution of judgments rather than aggregated labels. Variance decomposition shows that question-level conceptual difficulty accounts for substantially more variance than annotator identity, indicating that disagreement is structured by the task itself rather than driven by individual raters. Agreement-stratified analyses further reveal that key social-scientific effects, including country, education, and urban–rural differences, vary in magnitude and in some cases reverse direction across levels of inter-annotator agreement. These findings suggest that graded health-literacy evaluation contains both epistemically stable and unstable components, and that aggregating across them can obscure important inferential differences. We therefore argue that strong perspectivist modeling is not only conceptually justified but statistically necessary for valid inference in graded interpretive tasks.

Keywords: Perspectivist NLP, Health literacy, Low-resource languages, Annotation disagreement

1. Introduction

Manual annotation underlies most NLP systems. The dominant paradigm assumes that each instance has a single latent ground truth and resolves disagreement through aggregation methods such as majority voting or probabilistic label modeling. Within this framework, variability is typically interpreted as annotation noise or insufficient guideline clarity. Perspectivist approaches challenge this assumption by treating disagreement as potentially informative rather than erroneous (Basile et al., 2021a). While perspectivism has been widely discussed in overtly subjective tasks such as toxicity or stance detection (Davani et al., 2022; Kanclerz et al., 2021), less attention has been paid to more objective graded interpretive tasks. This raises a broader question: when does disagreement reflect instability in the task itself rather than annotator unreliability? Health-literacy evaluation provides a compelling case. Assessing whether an open-ended response to a public-health question is correct often involves interpreting partial knowledge, implicit reasoning, and degrees of completeness. Such judgments lie between fact verification and subjective stance evaluation, particularly in multilingual and low-resource settings where linguistic variation and unequal access to formal health information widen interpretive space.

In this paper, we analyze structured disagreement in graded COVID-19 health-literacy annotation across Ecuador and Peru. The dataset consists of 6,323 open-ended response-question items annotated independently by four raters using proportional correctness scores on a five-point scale, reflecting the degree to which each response aligns with normative answers derived from WHO and national public health guidelines, yielding 17,305 annotation-level observations. The corpus includes Spanish and Quechua-Kichwa responses along with sociodemographic metadata and represents one of the largest publicly releasable datasets linking graded health-literacy judgments with Indigenous Andean communities. We examine disagreement using variance decomposition and agreement-stratified inference. Question-level conceptual difficulty explains substantially more variance than annotator identity, indicating that disagreement is primarily task-structured rather than rater-driven.

Although lexical modeling captures substantial signal in responses, it does not eliminate epistemic variability. Crucially, social-scientific effects such as education and urban-rural differences vary in magnitude and, in some cases, direction depending on levels of inter-annotator agreement. Aggregation, therefore, obscures important inferential differences. In this work, we argue that graded health-literacy evaluation contains both epistemically stable and unstable components and that strong perspectivist modeling is statistically neces-

*Main and Corresponding Author.

Proceedings of NLPerspectives @ LREC 2026, pages 76–83

sary for valid inference in graded interpretive tasks.

2. Background and Related Work

Perspectivism distinguishes three central concepts (Frenda et al., 2025):

- **Disagreement**: observable variability in labels.
- **Subjectivity**: interpretive dependence on individual perspective.
- **Reliability**: annotator consistency independent of stance.

Disagreement may arise from subjectivity, ambiguity, or conceptual difficulty rather than annotator incompetence (Plank et al., 2014; Uma et al., 2022). Weak perspectivism preserves disaggregated labels, whereas strong perspectivism incorporates disagreement into model training, evaluation, and explanation (Basile et al., 2021a). The latter treats variability as a signal rather than noise. Most perspectivist research has focused on clearly subjective domains such as hate speech, stance classification, and aggressiveness detection in user-specific settings (Davani et al., 2022; Kanclerz et al., 2021). However, disagreement also appears in tasks typically framed as objective, including semantic similarity and inference (Biester et al., 2022). This suggests that epistemic instability may arise not only from annotator differences but also from properties of the task itself.

Public-health evaluation has traditionally relied on aggregated correctness judgments and closed-ended assessment tools, leaving graded interpretive variability underexplored. Prior work on health literacy assessment (Altin et al., 2014) and COVID-19 knowledge (Meneses-Navarro et al., 2020; Mejia et al., 2022) further suggests that educational background and community context can shape how public health information is interpreted. In multilingual and Indigenous settings, these challenges are especially pronounced. By examining health-literacy annotations through a perspectivist lens, this study extends disagreement-aware analysis to a graded public-health evaluation domain that combines factual knowledge, conceptual difficulty, and socially structured variation.

3. Data

3.1. Survey Design

Open-ended COVID-19 knowledge responses were collected in Ecuador and Peru as part of a broader cross-national health communication study (Kellert et al., in press). The instrument included questions covering transmission, symptoms, vaccination, risk groups, mask use, and protective measures. The survey was designed to elicit open-ended responses rather than forced-choice answers in order to capture graded health literacy and interpretive variability across respondents.

Size

- 17,305 annotation-level observations
- 6,323 response-question items
- 25 question identifiers
- 6,280 non-empty responses

3.2. Dataset & Participants

This corpus links graded health-literacy judgments with respondent-level sociodemographic metadata and self-reported information sources in historically underserved Quechua/Kichwa-speaking communities in Peru and Ecuador. It combines (i) open-ended responses, (ii) graded expert annotations, and (iii) metadata on language use, education, and location. The dataset is therefore relevant to public-health research, low-resource NLP, and perspectivist annotation studies that require naturally occurring interpretive variability rather than artificially constructed ambiguity.

Participants. Participants (N = 299) were recruited through targeted snowball sampling, a non-probability recruitment method in which referrals are guided toward specific subgroups of interest (e.g., rural, Indigenous, and linguistically diverse communities) in selected urban and rural sites (Lima and Apurímac in Peru; Cañar and El Tambo in Ecuador). The strategy aimed to capture variation in language use, education, and information access, particularly among Indigenous communities. Although non-probabilistic, the sampling was designed to increase representation from rural and linguistically diverse populations central to the study aims.

Survey items and analytic subset. The questionnaire included 30 items (including predominantly open-ended questions, along with a small number of structured items and metadata questions). For inferential analyses, we selected seven open-ended questions that were semantically specific enough to allow comparison with official WHO and national public-health guidance and that were frequently answered across participants. These items were selected prior to analysis and were used to construct composite knowledge scores and question-level models. In total, 18 questions were

annotated for correctness, of which seven open-ended questions were used for the primary inferential analyses reported in this paper.

Normative references and fieldwork. Normative answers were derived from official national health ministry materials and WHO recommendations available at the time of fieldwork and were used to develop the annotation rubric. Data were collected from November to December 2022 by trained local fieldworkers in participants' preferred language (Spanish, Quechua, or Kichwa). All responses were anonymized prior to analysis.

3.3. Annotation Protocol

Annotators. Four annotators participated in the labeling process. All were undergraduate students (aged 19-22) at Arizona State University, including two studying data science and two with backgrounds in biology and healthcare. Two annotators were fluent in Spanish and additionally assisted with dataset translation. This combination of technical and health-related expertise was intended to support both analytical consistency and domain-informed evaluation. All annotators were drawn from the same institutional context, which may introduce shared interpretive biases. We mitigate this by focusing on agreement structure and question-level variability rather than treating annotations as independent ground-truth judgments.

Scoring procedure. Responses were evaluated using a proportional correctness scoring scheme anchored to a five-point scale:

$$\{0, 0.25, 0.5, 0.75, 1.0\}$$

Proportional correctness reflects the degree to which a response aligns with normative answers derived from official national health ministry materials and WHO guidelines. Scoring was adapted to question type. Binary questions were scored as 0 (incorrect) or 1 (correct). For structured selection questions, scores were assigned proportionally based on the number of correct options selected. For open-ended responses, annotators assigned scores based on the extent to which key concepts from the normative answer were present, with partial matches receiving intermediate values.

While the five-point scale served as a common reference, some items admitted finer-grained proportional scores depending on the number of possible correct elements. Label distributions were skewed toward fully correct responses (score = 1), with a substantial proportion of partially correct responses (e.g., score = 0.5), indicating that intermediate values capture meaningful partial knowledge rather than annotator uncertainty.

Annotators worked independently and did not discuss individual cases during scoring. They were blinded to respondent sociodemographic metadata to minimize bias. Prior to annotation, raters received training using a pilot subset of responses to ensure consistent interpretation of the scoring rubric. For baseline lexical modeling, proportional scores were binarized into incorrect (< 0.5) and correct (≥ 0.5). This threshold reflects the distinction between predominantly incorrect and predominantly correct responses while preserving graded variation for subsequent disagreement analyses.

4. Agreement and Baseline Modeling

Metrics. Throughout the analysis, *accuracy* refers to the degree to which a response matches the normative answer derived from official public-health guidance. We report the following three measures throughout the paper:

- **Weighted Fleiss' κ.** We use the weighted version of Fleiss' κ, which accounts for ordinal distance between categories and therefore captures partial agreement more appropriately than the unweighted variant. Following conventional interpretation guidelines, values below 0.40 indicate low agreement, values between 0.40 and 0.60 moderate agreement, and values above 0.60 substantial agreement (Fleiss, J. L., 1971).
- **TF-IDF (Term Frequency-Inverse Document Frequency).** TF-IDF weights tokens by their frequency within a document relative to their frequency across the corpus. This representation captures discriminative lexical patterns and is used here as a baseline feature space for predicting binary correctness labels.
- **Intraclass Correlation (ICC).** To quantify the structure of disagreement, we report intraclass correlation coefficients (ICC), which estimate the proportion of total variance attributable to grouping factors (e.g., question or annotator). A high ICC for questions indicates that a substantial portion of the variance is structured by question-level properties rather than annotator identity.

Inter-annotator agreement, measured using weighted κ across the seven analytic questions, ranged from 0.42 to 0.70, indicating moderate to substantial agreement. As a lexical baseline, we trained a logistic regression classifier over TF-IDF features to predict binary correctness (incorrect < 0.5, correct ≥ 0.5). Model performance was evaluated using stratified 5-fold cross-validation to preserve class balance across splits. The classifier achieved:

$$\textbf{Accuracy} = 0.8398 \quad (SD = 0.0062)$$

The reported accuracy reflects the mean performance across held-out folds, with standard deviation indicating fold-level variability. This result demonstrates substantial lexical predictability: responses judged as correct and incorrect exhibit systematic differences in word usage. However, lexical separability does not imply epistemic stability. The classifier captures surface-level signal in token distributions but does not eliminate structured disagreement. Despite this predictive performance, residual variability remains strongly tied to question-level conceptual difficulty and agreement regimes. High lexical predictability therefore does not equate to inferential stability.

Variance Structure. To quantify the structure of disagreement, we estimated a mixed-effects model with random intercepts for question and annotator using proportional correctness scores as the dependent variable. Total variance was:

$$\sigma^2_{total} = 0.13196$$

Component	Variance	Percent
Question-level	0.04451	33.73%
Annotator-level	0.00151	1.14%
Residual	0.08594	65.13%

Table 1: Variance decomposition.

The intraclass correlation for questions ($ICC_{question} \approx 0.337$) substantially exceeds that for annotators ($ICC_{annotator} \approx 0.011$), indicating that disagreement is primarily structured by question-level conceptual difficulty rather than annotator inconsistency. In other words, variability reflects properties of the task more than properties of the raters.

5. Question-Level Variability and Education Gradients

The variance decomposition results indicated that disagreement clusters at the question level rather than at the annotator level. To further examine the source of this variability, we conducted a per-question variability analysis across all accuracy items.

5.1. Ranking Questions by Variability

For each question, we computed the variance of accuracy scores across respondents. The highest-variance questions were:

- 2.11. Is it possible to reuse masks? (Var = 0.162)
- 2.13. What is the safe distance between people? (Var = 0.149)
- 2.5. Asymptomatic infection (Var = 0.104)
- 2.10. How do masks prevent illness? (Var = 0.103)
- 2.1. How does COVID-19 spread? (Var = 0.101)
- 2.14. Who are the risk groups for COVID-19?

These items assess transmission mechanisms, mask efficacy, asymptomatic infection, and protective distancing, which are concepts that require both biomedical knowledge and interpretive reasoning.

5.2. Education Gradients

Education range refers to the difference in mean accuracy between the lowest and highest education groups on a 0-1 scale. To distinguish conceptual ambiguity from differences in knowledge across education levels, we computed the range in mean accuracy between the lowest and highest education groups for each question. The highest-variance questions exhibited large differences between education groups:

- 2.1. Spread of COVID-19: education range = 0.515
- 2.5. Asymptomatic infection: education range = 0.473
- 2.10. Mask mechanism: education range = 0.455
- 2.11. Mask reuse: education range = 0.400
- 2.13. Safe distancing: education range = 0.395
- 2.14. Risk groups: education range = 0.129

5.3. Interpretation

High variability in accuracy does not appear to arise solely from annotation ambiguity. Instead, it systematically co-occurs with large education-based differences in performance. Disagreement therefore clusters around conceptually demanding questions where knowledge is unevenly distributed across respondents. This pattern suggests that variability reflects the interaction between task-level conceptual difficulty and socially structured knowledge differences, rather than pure subjectivity or annotator inconsistency. Questions with higher variance are precisely those for which biomedical mechanisms

and implicit reasoning are required, making partial understanding more likely and graded judgments more sensitive to interpretive nuance. In particular, the presence of intermediate scores reflects graded partial correctness rather than binary uncertainty, reinforcing that disagreement arises from differing degrees of knowledge rather than purely subjective interpretation.

Moreover, high-variance items tend to exhibit less stable patterns across agreement levels, indicating that education-related effects depend on how clearly and consistently responses can be evaluated. Disagreement in this setting thus reflects structured epistemic difficulty rather than arbitrary labeling variation.

6. Agreement-Stratified Inference

To examine whether substantive conclusions vary across levels of epistemic stability, we stratified questions by inter-annotator agreement using Fleiss' κ. Items were classified as High ($\kappa > 0.60$), Medium ($0.40 \leq \kappa \leq 0.60$), and Low ($\kappa < 0.40$). For each stratum, we recomputed sociodemographic effects to assess whether relationships remain stable or shift across agreement levels. Sociodemographic differences were evaluated using standard statistical tests appropriate to each comparison (e.g., t-tests for binary group comparisons and one-way ANOVA for multi-level factors).

6.1. Country Effects

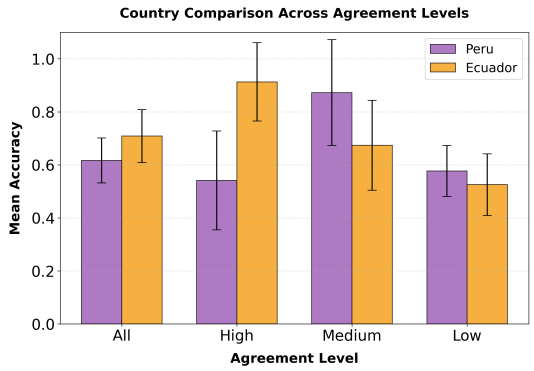


Figure 1: Country Comparison Across Agreement Levels

As shown in Figure 1, agreement levels vary across countries. Across all questions, Ecuador exhibits higher mean accuracy than Peru. Under high agreement, the country difference strengthens substantially. Under medium agreement, however, the direction reverses, with Peru outperforming Ecuador. Under low agreement, the difference attenuates but remains moderate in magnitude. Thus, both the magnitude and direction of country differences vary across agreement levels.

6.2. Gender Effects

Gender shows no significant differences across any agreement level, as shown in Figure 2. The null result remains stable under high, medium, and low agreement, indicating that the absence of gender effects is not driven by aggregation.

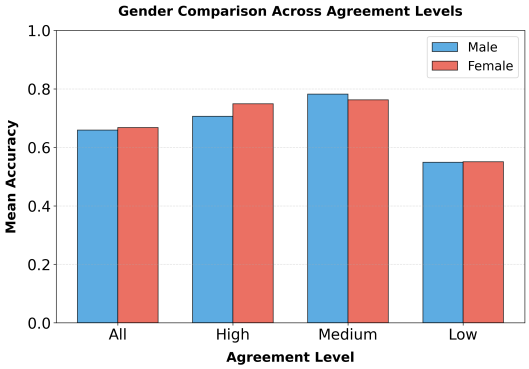


Figure 2: Gender Comparison Across Agreement Levels

6.3. Education Effects

Education is a significant predictor in the aggregated model as shown in Figure 3. Under high agreement, the education gradient remains strong ($p < .001$). Under medium agreement, however, the effect disappears entirely ($p = .837$), indicating no systematic education-based difference. Under low agreement, the effect re-emerges, though with reduced magnitude.

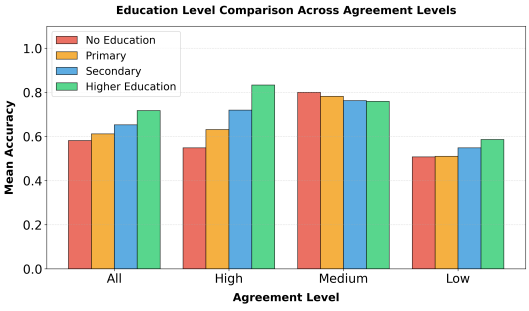


Figure 3: Education Comparison Across Agreement Levels

Thus, education-based inference collapses under moderate disagreement and is highly sensitive to the agreement level.

6.4. Urban-Rural Effects

As shown in Figure 4, urban respondents outperform rural respondents in the aggregated data (Ur-

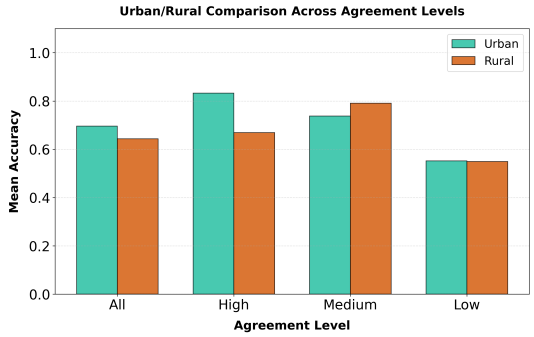


Figure 4: Urban and Rural Comparisons Across Agreement Levels

ban: $n = 117$, $M = 0.6961$; Rural: $n = 181$, $M = 0.6438$). Under high agreement, the urban advantage strengthens ($t = 6.2274$, $p < .001$). Under medium agreement, the effect reverses ($t = -2.12$, $p = .0351$), indicating higher accuracy in rural respondents. Under low agreement, the difference disappears entirely ($t = 0.24$, $p = .8124$).

Again, we see that inference is agreement-dependent. Figure 5 synthesizes the agreement-stratified results across predictors, demonstrating that effect magnitudes, and in some cases even their direction, shift across high, medium, and low agreement regimes, thereby exposing the instability introduced by epistemic variation. These reversals under medium agreement suggest that responses in this regime are more ambiguous or inconsistently interpretable, leading to unstable group comparisons. In such cases, annotators may agree on partial correctness but differ in how strongly responses align with normative criteria, which can attenuate or invert observed effects. This highlights that medium-agreement items occupy a transitional zone between clearly interpretable and highly ambiguous responses, where inference is particularly sensitive to how correctness is operationalized.

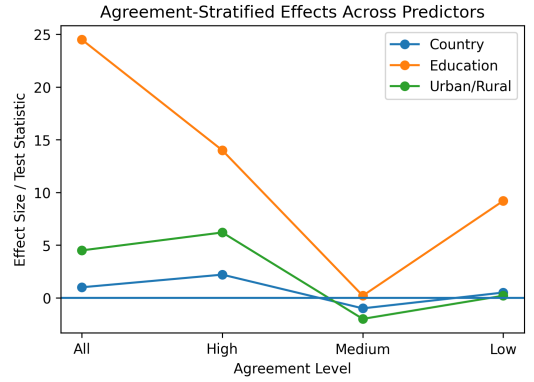


Figure 5: Agreement-Stratified Effects Across Predictors

6.5. Multivariate Interactions

We further examined whether education effects differ by location and country.

- **Location × Education (All Questions).** Urban areas showed a significant education effect ($F = 9.7841$, $p < .001$), while rural areas showed an even stronger effect ($F = 12.2732$, $p < .001$). This indicates that education has a stronger association with accuracy in rural areas than in urban areas.
- **Country × Education (All Questions).** Peru also showed a significant education effect ($F = 3.2749$, $p = .0230$), but the effect was substantially stronger in Ecuador ($F = 14.7899$, $p < .001$). This suggests that the education gradient is more pronounced in Ecuador than in Peru.
- **High Agreement Subset.** In the high-agreement subset, the same general pattern remains visible. The location-by-education effect is still stronger in rural areas (Urban: $F = 2.7109$, $p = .0485$; Rural: $F = 8.8561$, $p < .001$), and the country-by-education effect remains stronger in Ecuador.

The multivariate analyses further show that these patterns are context-sensitive: the strength of education effects differs across urban-rural and country comparisons, especially in the high-agreement subset. These multivariate results suggest that education-related differences are not uniform across social contexts, but vary by both geography and country.

7. Discussion

This study provides an empirical test of core perspectivist claims within a graded public health evaluation setting. Across analyses, disagreement is shown to be structured rather than incidental.

Structured Disagreement. Variance decomposition indicates that disagreement is primarily driven by question-level conceptual difficulty rather than annotator identity. The very low annotator-level ICC suggests that variability does not arise from rater inconsistency, but from properties of the task itself. Disagreement in this setting, therefore, reflects structured epistemic complexity rather than annotation noise.

Disagreement and Inference. Agreement-stratified analyses demonstrate that substantive conclusions vary across levels of inter-annotator agreement. Effects that appear robust in aggregated models may strengthen, attenuate,

disappear, or reverse direction when evaluated within specific agreement strata. Aggregation thus collapses heterogeneous evaluation contexts into a single estimate, masking instability in underlying relationships.

Beyond Overt Subjectivity. These findings extend perspectivist theory beyond clearly subjective domains such as toxicity or stance detection. In a graded health-literacy assessment, disagreement emerges at the intersection of conceptual difficulty and socially distributed knowledge. Variability is not merely interpretive difference, but a signal of uneven epistemic access across respondents.

Methodological Implications. The results suggest that strong perspectivist modeling is not simply a normative preference but a requirement for valid inference in graded interpretive tasks. Agreement-sensitive and hierarchical approaches allow researchers to distinguish stable knowledge patterns from contexts of interpretive uncertainty rather than collapsing them into a single summary label.

Practical Implications. For NLP, these findings support disagreement-aware evaluation and modeling strategies in tasks that combine factual reasoning with graded interpretation. For public health research, inter-annotator agreement functions as a diagnostic indicator of epistemic stability and should inform how knowledge assessments are interpreted.

Taken together, the results demonstrate that epistemic stability must be modeled explicitly rather than assumed, and that aggregation without regard to disagreement structure risks obscuring substantively important variation.

8. Conclusion

Health-literacy annotation exhibits structured disagreement driven primarily by conceptual difficulty rather than annotator inconsistency. Substantive social inference varies across levels of inter-annotator agreement: effects that appear robust under aggregation may strengthen, attenuate, or reverse across agreement strata. Agreement-stratified and multivariate analyses further show that these patterns are context-sensitive, with education-related differences varying by geography and country. Together, these results demonstrate that aggregation can obscure meaningful epistemic and social variation, and that disagreement encodes differences in partial knowledge rather than noise. These findings highlight the importance of disagreement-aware approaches for modeling graded, real-world knowledge in NLP and public health settings.

9. Limitations.

Several limitations should be noted. First, annotators were drawn from a single institutional context, which may introduce shared interpretive biases despite efforts to standardize training and blind annotations. Second, proportional correctness relies on normative public-health guidelines, which may not fully capture local knowledge practices or alternative valid interpretations. Finally, while agreement-stratified analyses reveal important patterns, they do not establish causal relationships between disagreement and social factors. Future work should develop multilevel and agreement-aware modeling approaches that explicitly incorporate agreement as a moderating factor and extend this framework to other graded evaluation domains.

10. Ethical Consideratin

Data were collected as part of a broader health communication study conducted by trained local fieldworkers. Participants provided informed consent for anonymous data collection and analysis. All responses were anonymized prior to annotation and analysis to protect participant privacy.

Data Availability

The deidentified dataset and annotation guidelines are released on GitHub at: https://github.com/olga-kel/Health-Communication. The resource supports research in low-resource and Indigenous language NLP, multilingual information access, and disagreement-aware annotation modeling.

Acknowledgments

The authors thank the local fieldworkers and community collaborators who supported participant recruitment, transcription, and data collection in Ecuador and Peru. In particular, we thank Fernando Ortega, Claudia Crespo, and Marleen Haboud for their contributions to the design and execution of the data collection, as well as their fieldwork assistants for their support in gathering the data. We are especially grateful to Maria Rosa Guamán (Cañar, Ecuador) and Mery Salas Santa Cruz (Apurímac, Peru) for conducting interviews and supporting field data collection. We also thank Talan Herrera and Noah Arellano for translating the datasets from Spanish into English and assisting with annotation. This research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), Project Number 468416293. Olga Kellert was the Principal Investigator (PI) and Stavros Skopeteas was the Co-PI

of the funded project that made the data collection possible. Finally, we thank all participants for their time and willingness to take part in the study.

References

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. arXiv preprint arXiv:2109.04270v3.

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. Association for Computational Linguistics.

Laura Biester, Vinodkumar Prabhakaran, Anjalie Field, Naomi Saphra, and Rada Mihalcea. 2022. Analyzing the effects of annotator gender across NLP tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @ LREC2022*, pages 10–19. European Language Resources Association.

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.

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

Kamil Kanclerz, Alicja Figas, Marcin Gruza, Tomasz Kajdanowicz, Jan Kocon, Daria Puchalska, Przemyslaw Kazienko. 2021. Controversy and Conformity: from Generalized to Personalized Aggressiveness Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021)*, pages 5915–5926.

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

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Valerio Basile. 2021. 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*, Lecture Notes in Computer Science, pages 441–453. Springer.

Sergio Meneses-Navarro, María Graciela Freyermuth-Enciso, Blanca Estela Pelcastre-Villafuerte, Roberto Campos-Navarro, David Mariano Meléndez-Navarro, and Liliana Gómez-Flores-Ramos. 2020. The challenges facing indigenous communities in Latin America as they confront the COVID-19 pandemic. *International Journal for Equity in Health*, 19(1):63.

Christian R. Mejia, Telmo Raul Aveiro-Robalo, Luciana Daniela Garlisi Torales, Maria Fernanda Fernández, Francisco E. Bonilla-Rodríguez, Enrique Estigarribia, Johanna Magali Coronel-Ocampos, Cecilia J. Caballero-Arzamendia, Renato R. Torres, Aram Conde-Escobar, Yuliana Canaviri-Murillo, Diana Castro-Pacoricona, Victor Serna-Alarcón and Dennis Arias-Chávez. 2022. Basic COVID-19 knowledge according to education level and country of residence: Analysis of twelve countries in Latin America. *Frontiers in Medicine*, 9:978795.

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

Sibel Vildan Altin, Isabelle Finke, Sibylle Kautz-Freimuth and Stephanie Stock. 2014. The evolution of health literacy assessment tools: a systematic review. *BMC Public Health*, 14(1):1207.

Olga Kellert, Fernando Ortega, Claudia Crespo, Marleen Haboud, Salma Atfah, Hannah Sommer, and Stavros Skopeteas. In press. ¿Cómo impactaron las fuentes de información y los factores sociodemográficos en la transmisión del conocimiento relacionado con la COVID-19 a nivel de minorías étnicas y lingüísticas? In Luis Moreno Hernández, Alma Delia Zárate Flores, Blanca Cortez Rodríguez, Mauro García Solano, Karla Marisol Teutli Mellado, Hazel Cordero Perea (Coords.), *Epidemiología y prevención en salud bucal: Estudios y retos*, pages 77–100. Benemérita Universidad Autónoma de Puebla.

SubData: Bridging Heterogeneous Datasets to Enable Theory-Driven Evaluation of Political and Demographic Perspectives in LLMs

Pietro Bernardelle[1] **Leon Fröhling**[2] **Stefano Civelli**[1] **Gianluca Demartini**[1]
[1]The University of Queensland, Australia
[2]GESIS - Leibniz Institute for the Social Sciences, Germany
{p.bernardelle,s.civelli,g.demartini}@uq.edu.au leon.froehling@gesis.org

Abstract

As increasingly capable large language models (LLMs) emerge, researchers have begun exploring their potential for subjective tasks. While recent work demonstrates that LLMs can be aligned with diverse human perspectives, evaluating this alignment on downstream tasks (e.g., hate speech detection) remains challenging due to the use of inconsistent datasets across studies. To address this issue, in this resource paper we propose a two-step framework: we **(1)** introduce SubData, an open-source Python library designed for standardizing heterogeneous datasets to evaluate LLMs perspective alignment; and **(2)** present a theory-driven approach leveraging this library to test how differently-aligned LLMs (e.g., aligned with different political viewpoints) classify content targeting specific demographics. SubData's flexible mapping and taxonomy enable customization for diverse research needs, distinguishing it from existing resources. We illustrate its usage with an example application and invite contributions to extend our initial release into a multi-construct benchmark suite for evaluating LLMs perspective alignment on natural language processing tasks.

Keywords: Dataset Standardization, Perspective Alignment in LLMs, Hate Speech Detection Resources

1. Introduction

The ever-increasing capabilities of large language models (LLMs) have enabled them to capture increasingly nuanced human perspectives (Bommasani et al., 2021; Brown et al., 2020). Researchers have begun exploring their potential for subjective tasks, with particular focus on "perspective alignment"—the ability of models to reflect diverse human viewpoints across different contexts (Durmus et al., 2023; Kirk et al., 2024). Ensuring robust evaluation of this alignment is crucial as LLMs increasingly mediate information access and influence decisions in socially sensitive domains where human perspectives naturally differ (Blodgett et al., 2020; Khamassi et al., 2024; Weidinger et al., 2021).

Recent work has explored how well LLMs can represent diverse human perspectives using two different approaches. The first approach examines whether models accurately predict how individuals (Argyle et al., 2023) or groups (Santurkar et al., 2023) would respond to surveys, a task Sorensen et al. (2024) describe as *distributional pluralism*. The second investigates whether aligned LLMs consistently reflect broader viewpoints across tasks (Agiza et al., 2024; Chen et al., 2024; Feng et al., 2023; Haller et al., 2024; He et al., 2024), aligning with what Sorensen et al. (2024) term *steerable pluralism*.

Survey prediction provides a natural evaluation setting: models' outputs—generated either through fine-tuning or persona-conditioning to represent specific perspectives—can be compared against authentic survey responses from individuals or subpopulations (Rupprecht et al., 2025). Because such datasets contain demographic information and corresponding answers, they create a clear benchmark: a well-aligned LLM should produce distributions that closely resemble real responses and can be evaluated using standard divergence or accu-

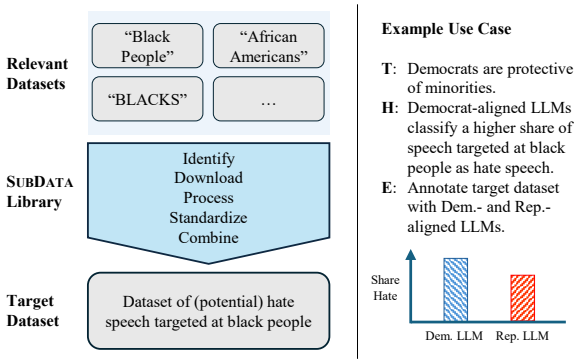


Figure 1: Overview of our proposed evaluation framework. SubData consolidates instances from diverse datasets into a unified resource. To assess LLM alignment with human perspectives from the combined dataset, we propose a workflow that tests theory-derived (**T**) hypotheses (**H**) through controlled experiments (**E**), measuring how accurately LLMs reflect viewpoints of different demographic and ideological groups.

Proceedings of NLPerspectives @ LREC 2026, pages 84–97
12 May 2026.

racy measures (Sorensen et al., 2024).

The broader challenge of task-independent alignment has inspired various evaluation methodologies. Political alignment studies by Agiza et al. (2024) and Chen et al. (2024) use the Political Compass Test (PCT)—a widely used questionnaire for mapping political beliefs along economic and social axes—to verify whether models aligned to specific ideologies position themselves appropriately on the PCT map. He et al. (2024) compare model answers to multiple-choice questions against positions expressed by relevant subgroups. Sorensen et al. (2024) propose direct human annotations or reward models to measure whether generated responses correctly reflect specific attributes. More closely related to our conceptualization of alignment evaluation, Haller et al. (2024) assess sentiment in open-ended generations when prompted about different demographics, while Feng et al. (2023) examine how political alignment affects hate speech detection toward different targets.

Despite these efforts, evaluating how perspective-aligned LLMs perform on subjective classification tasks remains challenging (Zheng et al., 2024), largely due to the lack of standardized resources that enable consistent comparison across viewpoints (Alipour et al., 2025). We address this gap by introducing a two-step framework that enables systematic evaluation of perspective-aligned language models.

(1) Dataset Standardization: SubData. We introduce SubData, an open-source Python library that collects and harmonizes heterogeneous datasets for subjective tasks.[1] Unlike general repositories, it unifies inconsistent annotation schemes and demographic categorizations, allowing researchers to build consistent collections for their needs. Our initial release focuses on hate speech detection, integrating ten datasets with a unified taxonomy of target groups (§3, §4). In doing so, we do not host or redistribute the datasets themselves, but we reviewed their licenses to ensure that our use aligns with creators' intentions of fostering hate speech research. Consistent with Vidgen and Derczynski (2020), we further emphasize the need to handle such material responsibly, with attention to privacy, personal data, and potential online harms.

(2) Theory-Driven Hypothesis Testing. Building on these standardized datasets, we propose a theory-driven approach to evaluate alignment (§5). As illustrated in Figure 1, our framework follows a systematic process: researchers first formulate hypotheses (**H**) based on established social or political theory (**T**), then design experiments (**E**) to test whether differently-aligned models behave as expected. For instance, the workflow on the right side of Figure 1 illustrates testing whether Democrat-aligned LLMs classify more anti-Black content as hate speech than Republican-aligned ones, reflecting the popular hypothesis that Democrats prioritize minority protection theoretically derived by (Solomon et al., 2024). This framework enables quantitative measurement of alignment differences through controlled experimentation, and we further demonstrate its application in §6.

Our approach does not rely on subjective ground-truth labels; instead, it measures classification differences across models with distinct alignments, providing a direct lens on how perspective conditioning shapes downstream task behavior. While prior work has examined subjectivity in LLM annotation (Beck et al., 2024; Giorgi et al., 2024; Orlikowski et al., 2023), our framework extends this by systematically evaluating alignment effects in downstream applications.

2. Related Work

2.1. LLMs Perspective Alignment

Research on aligning LLMs with diverse human perspectives has followed two main approaches: fine-tuning models on perspective-specific data and using persona-based prompting.

Several studies have explored fine-tuning approaches for task-agnostic LLMs alignment. Agiza et al. (2024), Chen et al. (2024) and Feng et al. (2023) investigated how political alignment and data selection affect model biases and downstream tasks like hate speech detection. Similarly, Haller et al. (2024) developed OpinionGPT by fine-tuning models on ideologically diverse data to represent explicit biases.

As an alternative to these resource-intensive post-training methods, persona-based prompting has emerged as a more efficient technique for task-specific perspective alignment. Argyle et al. (2023) showed that LLMs can accurately simulate survey responses across demographic groups, while Fröhling et al. (2025) and Ge et al. (2024) demonstrated how synthetic personas can diversify model outputs and annotations. Building on this, Bernardelle et al. (2025a,b) mapped persona-prompted LLMs onto the PCT compass, providing a large-scale analysis of how these personas impact the distribution of language models across political ideological space. Similarly, Civelli et al. (2025a,b) revealed how politically-aligned persona-conditioned LLMs influence hateful content detection.

[1] All code is available open-source on GitHub and the library can be installed directly from PyPi.

Orlikowski et al. (2025) combined these approaches by fine-tuning models with socio-demographic attributes to represent individual annotators, finding that persona-based prompting barely improves the models' ability to predict individuals' annotations and that improvements from fine-tuning mainly come from demographic profiles serving as identifiers for individual annotators. Liu et al. (2024) identified further limitations in this technique, showing that models struggle with "incongruous personas" and default to stereotypical stances when predicting responses for personas with contradicting traits. The conflicting evidence seen in the literature regarding the models' ability to consistently represent different subjective perspectives serves as further motivation to develop comprehensive resources for the evaluation of this type of LLMs perspective alignment.

2.2. Evaluating LLMs Perspective Alignment

Evaluating alignment presents significant challenges, particularly for subjective tasks.

For survey response prediction, He et al. (2024) and Santurkar et al. (2023) compared model predictions against actual responses from specific demographic groups. Castricato et al. (2025) built on the PRISM dataset (Kirk et al., 2024) to create a test bed for evaluating pluralistic alignment using preference pairs from personas sampled from census data.

For downstream tasks, Giorgi et al. (2024) and Zheng et al. (2024) assessed how personas affect model performance and biases in content classification. Despite these advances, evaluating perspective-aligned LLMs on subjective classification tasks remains challenging due to the lack of standardized resources that enable consistent comparison—a gap our proposed framework addresses.

3. SubData Construction

3.1. Dataset Selection Criteria

Our approach to evaluating perspective alignment in LLMs necessitates datasets with specific characteristics suited for this analysis. We require datasets that address subjective constructs such as hate speech, toxicity, or abusive language—domains where human interpretations naturally diverge across demographic and ideological lines (Sap et al., 2022). This subjectivity is essential as it creates the interpretive space where different perspectives become measurable. Additionally, these datasets must provide explicit annotations identifying which specific demographic groups are targeted by the content (for example, specifying when content targets Jews, women, or immigrants), rather than merely indicating that some unspecified group was targeted. This granular targeting information is crucial because it enables us to test theory-driven hypotheses about how LLMs aligned with different perspectives might classify content targeting specific demographics differently.

3.2. Data Collection Methodology

Because of the lack of a single repository that stores and documents the properties of datasets, identifying the set of relevant datasets is an inherently difficult challenge. We therefore employed a multi-phase approach to identify suitable datasets.

First, we leveraged our existing knowledge of hate speech detection literature to identify candidate datasets, drawing on our team's established expertise in this domain. Second, we examined existing repositories including hatespeechdata.com (Vidgen and Derczynski, 2020) and toxic-comment-collection (Risch et al., 2021), which provided structured access to multiple potentially relevant datasets. Third, we conducted systematic searches with keyword combinations of "target[ed]" and "hate speech" on scholarly databases to identify related literature that might present or reference additional resources. Finally, we individually assessed each dataset through manual verification to confirm it contained explicit target group annotations that satisfied our criteria.

This process yielded ten datasets that meet our requirements. While we have striven to make our initial dataset collection comprehensive, we acknowledge that this collection is not exhaustive and that some relevant sources may have been overlooked. Rather than seeing this as a limitation, we consider it an opportunity to build a collaborative research community focused on annotation subjectivity. We actively encourage researchers to contact us with suggestions for additional datasets that satisfy our outlined criteria to be included in the library.

3.3. Dataset Characteristics

Table 1 provides an overview of the datasets included in SubData so far, categorizing targets across nine demographic dimensions (age, disability, gender, migration, origin, political, race, religion, and sexuality). All target categories are organized according to the unified taxonomy we detail in §4, which standardizes the heterogeneous labels from original sources. This standardized categorization enables researchers to quickly identify suitable datasets for specific research questions regarding perspective alignment, highlighting both

Dataset \ Category	age	disabled	gender	migration	origin	political	race	religion	sexuality	Dataset size
Fanton et al. (2021)	0 (0)	175 (1)	560 (1)	637 (1)	0 (0)	0 (0)	301 (1)	1,402 (2)	465 (1)	3,540
Hartvigsen et al. (2022)	0 (0)	19,631 (1)	19,563 (1)	0 (0)	62,458 (3)	0 (0)	80,979 (4)	41,014 (2)	21,344 (1)	244,989
Jigsaw (2019)	0 (0)	18,602 (3)	178,266 (4)	0 (0)	0 (0)	0 (0)	94,334 (5)	132,734 (7)	29,115 (4)	453,051
Jikeli et al. (2023a)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	6,439 (1)	0 (0)	6,439
Jikeli et al. (2023b)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	3,012 (3)	2,315 (2)	0 (0)	5,327
Mathew et al. (2021)	0 (0)	153 (1)	5,584 (2)	1,701 (1)	1,855 (2)	0 (0)	7,684 (5)	6,106 (6)	2,750 (4)	25,833
Röttger et al. (2021)	0 (0)	510 (1)	1,020 (2)	485 (1)	0 (0)	0 (0)	504 (1)	510 (1)	577 (1)	3,606
Sachdeva et al. (2022)	2,355 (4)	1,801 (3)	22,535 (5)	5,473 (2)	11,637 (2)	0 (0)	21,024 (7)	12,461 (8)	14,934 (4)	92,220
Vidgen et al. (2021a)	41 (2)	414 (3)	689 (3)	45 (2)	164 (5)	688 (7)	397 (4)	273 (4)	472 (3)	3,183
Vidgen et al. (2021b)	23 (1)	521 (1)	3,630 (4)	1,507 (2)	862 (6)	0 (0)	3,881 (5)	2,384 (2)	1,437 (3)	14,245
All Datasets	2,419 (4)	41,807 (3)	231,847 (5)	9,848 (4)	76,976 (11)	688 (8)	212,116 (8)	205,638 (8)	71,094 (6)	852,433

Table 1: Overview of hate speech datasets in SubData, showing the number of instances and unique target groups (in parentheses) per target category. *Note*: The "All Dataset" row reports the total unique target groups per category across all datasets. When the total equals the maximum from a single dataset (e.g., disabled: 3, matching Jigsaw (2019)'s 3), that dataset fully accounts for the category's unique target groups. When the total exceeds the maximum (e.g., origin: 11, exceeding Hartvigsen et al. (2022)'s 3), multiple datasets contribute distinct target groups.

the strengths and limitations of current hate speech detection resources.

We would like to point out that the number of entries in some datasets of Table 1 may differ from those reported in the original publications because of our focus on targeted hate speech. When entries in source datasets had multiple targets in a single annotation (e.g., "[bla, jew]"), we created separate instances for each target, thereby increasing the number of entries. Conversely, we excluded entries without specific target groups (e.g., labeled as "other"), resulting in datasets that sometimes contain fewer instances than the originals. We also deduplicate instances, removing repeated entry-target pairs even when these duplications might be intentional in the original dataset—such as in Fanton et al. (2021) where identical hate speech instances appear multiple times with different counterspeech responses. Since our research focuses specifically on targeted hate speech, we treat these as functional duplicates.

4. SubData Unified Taxonomy

Following our dataset selection and collection methodology, SubData implements a standardized taxonomy that addresses the inconsistencies in how target groups are labeled across hate speech datasets. This allows to leverage the systematic evaluation framework described in §5 by creating consistency across disparate data sources.

4.1. Taxonomy Design Principles

The development of our taxonomy was guided by several key design principles tailored to the practical needs of researchers studying perspective alignment. We sought to balance specificity and generalizability, preserving critical distinctions between target groups while establishing categories broad enough to facilitate meaningful cross-dataset analysis. For instance, the target group "LGBTQ+" is commonly used in the literature to encompass a wide range of minority sexual and gender identities. While we recognize that this label can be overly broad, potentially obscuring the diverse experiences of the groups it covers, we decided against introducing every identity under this umbrella as a separate target group.

Importantly, our demographic categories were not arbitrarily chosen; they emerged from a bottom-up approach, derived directly from the categories present in the original datasets we sourced. This method ensures that our taxonomy reflects and unifies the actual structure of existing hate speech research, maintaining alignment with the data's inherent organization. Additionally, whenever possible, we preserved consistency with the original researchers' taxonomic decisions to honor their methodological choices and conceptual frameworks.

4.2. Target Group Mapping

The mapping process converts heterogeneous target labels from original datasets into our standardized taxonomy. This involves both direct equivalences (e.g., "Jewish people" → "jews") and more complex decisions requiring contextual judgment. Table 2 provides a sample of our mapping strategy across multiple datasets, illustrating how diverse original terminology is standardized in SubData.

For ambiguous cases, we consulted dataset documentation to determine the original authors' intent. For instance, determining whether the target "mexicans" should be mapped to the "latin" (race category) or "mexicans" (origin category) required careful contextual judgment. When documentation clarified the original creators' intended meaning, we followed their categorization. When such guidance was unavailable, we applied consistent principles across similar cases. To validate the reliability of these decisions, we conducted an inter-annotator

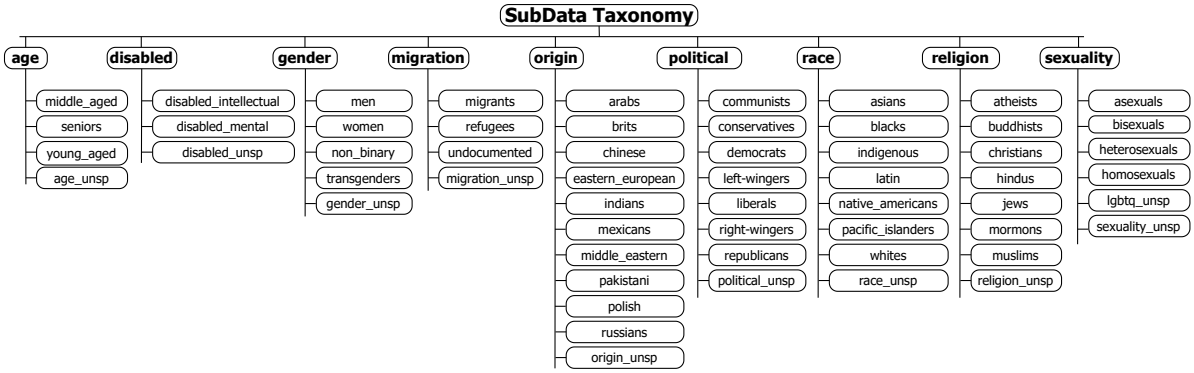


Figure 2: SubData taxonomy structure with target groups organized by category. *Note:* targets that should end in "_unspecified" have been abbreviated in the figure using "'_unsp."

Dataset	Original Keyword	Target
Fanton et al. (2021)	"JEWS"	jews
Hartvigsen et al. (2022)	"jewish"	jews
Jikeli et al. (2023a)	"Kikes"	jews
Vidgen et al. (2021a)	"jewish people"	jews
Vidgen et al. (2021b)	"bla, jew"	jews blacks
Vidgen et al. (2021b)	"bla, african"	blacks
Jigsaw (2019)	"black"	blacks
Jikeli et al. (2023b)	"Blacks"	blacks
Röttger et al. (2021)	"black people"	blacks

Table 2: Standardization of target terminology across datasets using SubData's mapping system. The table provides examples of how diverse original keywords from multiple hate speech datasets are normalized into consistent target categories.

agreement check on the mapping process, obtaining a Cohen's κ of 0.986.

As part of our approach, for each category we designated target groups with the suffix "_unspecified" (e.g., "disabled_unspecified," "race_unspecified") to handle cases where the original dataset used generic terminology without specifying subtypes.

Figure 2 illustrates the complete taxonomy structure with all target groups organized by category.

4.3. Taxonomy Limitations and Customization

Despite our efforts to create a comprehensive framework, we acknowledge several limitations in our taxonomy that primarily stem from the inherent challenges associated with the matching we are performing (Shvaiko and Euzenat, 2011). These include the LGBTQ+ target group heterogeneity that mixes gender identities and sexual orientations, blurred distinctions between racial identity and geographic origin, and simplified representations of demographic intersectionality mapped to single-attribute target groups (e.g., "blacks,women"). Independent from our work, Fillies and Paschke (2025) point to the same challenges when developing their targeted hate speech taxonomy, relying on similar strategies to solve them.

We are confident that our taxonomy represents a useful basis for different research purposes and take the large overlap with the unified taxonomy proposed by Fillies and Paschke (2025) as evidence for convergence on a generally accepted targeted hate speech taxonomy. However, recognizing that no single taxonomy can satisfy all research needs, SubData provides several customization functions that give researchers flexibility in adapting the framework to their specific requirements.

While this customizability is valuable, it creates challenges for maintaining comparability across studies when researchers modify the taxonomy. To address this issue and increase transparency, we implemented a functionality to export a LaTeX version of the taxonomy (and all other modifiable resources) that researchers can include directly in their manuscripts, clearly documenting any modifications they have made.

5. Theory-Driven Hypothesis Testing

The SubData library not only provides standardized datasets but also serves as a foundation for a theory-driven approach to evaluating LLMs perspective alignment. This approach follows the process illustrated in Figure 1:

1. Theory (**T**): Researchers begin by identifying established social or political theories that predict differences in how various demographic or ideological groups differ in their perception of subjective constructs.
2. Hypothesis (**H**): Based on these theories, researchers formulate testable hypotheses

about how LLMs aligned with different perspectives might classify content.

3. Experiment (**E**): Using SubData's standardized datasets, researchers design controlled experiments to test these hypotheses by measuring classification differences between differently-aligned models.

Advantages of the Framework. The theory-driven framework we propose offers substantial benefits for researchers studying LLM perspective alignment. By focusing on comparative model behavior rather than adherence to supposedly objective standards, our approach **(1) elegantly circumvents the persistent challenge of subjectivity in human annotations**. When dealing with inherently subjective constructs like hate speech, the framework does not require consensus on "ground truth" labels—which are often contested and vary across demographic and ideological lines—but instead directly measures differences between models aligned with distinct perspectives. This shift in evaluation methodology acknowledges the fundamental subjectivity of these tasks while still enabling rigorous analysis by grounding the tested hypotheses directly in theory.

Furthermore, our approach **(2) enables precise quantitative measurement of alignment effects on classification behavior**. Researchers can measure exactly how much perspective alignment influences model outputs when classifying content targeting specific demographics, providing concrete metrics rather than relying on qualitative assessments. This quantitative foundation makes evaluations more rigorous and facilitates meaningful comparisons across different studies, contributing to more cumulative research in this emerging field.

The framework's versatility extends beyond its primary application in political alignment evaluation. It **(3) naturally supports diverse research directions**. This flexibility makes our approach valuable for researchers working at the intersection of natural language processing (NLP), social science, and ethical AI development, potentially informing more nuanced approaches to model development and evaluation.

6. Example Use of SubData

To demonstrate the proposed framework, we present here a concrete use case. Feng et al. (2023) show that pretraining LLMs on partisan corpora shifts their political leaning, and that this shift propagates into downstream tasks such as hate speech detection, where left-leaning models tend to flag more content targeting minority groups than right-leaning ones (**T**). This connection is theoretically plausible because hate speech detection is not a politically neutral classification task: decisions about what should be flagged often reflect broader normative commitments around social harm, tolerance, and free expression. As a result, if political conditioning meaningfully changes a model's perspective, hate speech detection is one of the downstream settings where such differences should be especially visible.

Following their categorization (BLACKS, MUSLIMS, LGBTQ+, JEWS, LATIN, WOMEN, MEN, CHRISTIAN, WHITE), we hypothesize that a similar dynamic holds when partisan alignment is induced through persona-conditioning (**H**). Specifically, LLMs conditioned on left-leaning personas should produce higher detection rates for hate speech against minority groups, while LLMs conditioned on right-leaning personas should show the opposite tendency. The following subsections detail the experimental setup (**E**) and results.

6.1. Methodology

Data. We use the SubData library to collect and standardize the instances of interest from existing hate speech datasets. Specifically, we rely on its unified taxonomy of ten demographic groups: BLACKS, MUSLIMS, LGBTQ_UNSP, JEWS, ASIANS, LATIN, WOMEN, MEN, CHRISTIANS, WHITES.

This procedure enables us to construct a single merged dataset that ensures comparability across groups and facilitates controlled evaluation of perspective alignment. Because evaluating the full corpus across multiple models and persona conditions would be computationally prohibitive, we instead randomly sample 2,500 statements per target group after merging, yielding a balanced dataset of 25,000 instances for our experiments.

The sampling procedure chosen does not alter any original frequency distributions of different targets, given that our collection of publicly available datasets should not be treated as a ground-truth distribution of targets in the hate speech literature or in the real world. Furthermore, by choosing 2,500 instances per target group, we do not have to rely on oversampling or synthetic duplication to achieve this stratification. The smallest of the target groups studied in our use case features around 21,000 instances.

Language Models. We selected three open-source, instruction-tuned conversational LLMs for our analysis: Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Llama-3.1-8B-Instruct (Dubey et al., 2024) and Qwen2.5-7B-Instruct (Team, 2025). These models were chosen for their open-source availability and moderate parameter size (7–8B), which

Target	Mistral-v0.3-7B				Llama-3.1-8B				Qwen-2.5-7B			
	Left	Right	OR	95% CI	Left	Right	OR	95% CI	Left	Right	OR	95% CI
blacks	**0.548**	**0.471**	1.359^{***}	[1.325,1.393]	0.625	0.614	1.050^{***}	[1.023,1.077]	0.319	**0.311**	1.037^{**}	[1.009,1.065]
muslims	0.471	0.391	1.389^{***}	[1.355,1.425]	0.599	0.576	1.100^{***}	[1.072,1.128]	0.228	0.204	1.152^{***}	[1.117,1.187]
lgbtq_unsp	0.313	0.247	1.389^{***}	[1.351,1.428]	0.391	0.372	1.086^{***}	[1.059,1.114]	0.168	0.157	1.084^{***}	[1.048,1.121]
jews	0.497	0.405	1.450^{***}	[1.415,1.487]	0.576	0.553	1.101^{***}	[1.074,1.129]	**0.320**	0.307	1.064^{***}	[1.036,1.093]
asians	0.343	0.260	1.481^{***}	[1.442,1.522]	0.469	0.446	1.096^{***}	[1.069,1.123]	0.197	0.176	1.146^{***}	[1.111,1.184]
latin	0.378	0.297	1.443^{***}	[1.405,1.482]	0.473	0.447	1.110^{***}	[1.082,1.138]	0.209	0.199	1.064^{***}	[1.032,1.097]
women	0.364	0.298	1.347^{***}	[1.312,1.383]	0.477	0.467	1.042^{**}	[1.016,1.068]	0.161	0.151	1.073^{***}	[1.037,1.110]
christians	0.202	0.167	1.263^{***}	[1.223,1.304]	0.298	0.290	1.040^{**}	[1.012,1.069]	0.064	0.061	1.057^{*}	[1.004,1.112]
men	0.299	0.250	1.279^{***}	[1.244,1.315]	0.442	0.429	1.056^{***}	[1.030,1.083]	0.111	0.104	1.081^{***}	[1.038,1.125]
whites	0.523	0.440	1.393^{***}	[1.359,1.429]	**0.656**	**0.649**	1.033^{*}	[1.006,1.060]	0.259	0.253	1.029^{*}	[1.000,1.059]
Overall	0.394	0.323	1.364^{***}	[1.352,1.375]	0.501	0.484	1.068^{***}	[1.060,1.077]	0.204	0.193	1.073^{***}	[1.063,1.084]

Table 3: Hate speech detection rates by target category and persona position across the three LLMs investigated. Each cell shows the average proportion of content flagged as hateful when targeting the specified group, using a persona-conditioned model with 20 left- and 20 right-oriented personas. Bold values indicate the highest detection rate for each model–condition pair across all targets. Odds Ratios (OR) quantify detection differences, with OR > 1 indicating higher rates for left personas. 95% confidence intervals are reported in a dedicated column. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

strikes a balance between reproducibility and diversity, allowing us to derive insights that generalize across architectures. We specifically use their conversational variants, fine-tuned for instruction following (Ouyang et al., 2022), as this aligns with our methodology: leveraging in-context prompts to condition models on different personas and evaluate their hate-speech detection behavior.

Experimental Setup. To simulate partisan perspectives, we adopt the political distributions introduced by Bernardelle et al. (2025a,b), which map persona-conditioned LLMs across the PCT ideological space. Following the approach of Civelli et al. (2025a,b), we select the 20 most left-leaning and 20 most right-leaning persona descriptions from each model distribution, yielding 40 personas in total per model. Each persona is then used as an in-context instruction to condition the LLMs for hate speech detection on the unified dataset (refer to Appendix A for more information on the prompt template).

To improve reliability and maintain consistent output formats, we adopted a structured output generation strategy throughout our experiments.[2] This approach constrains model generations to follow a predefined schema, thereby helping to prevent ill-formed or off-task completions. At inference time, schema adherence is enforced through dynamic vocabulary masking: at each decoding step, only tokens that keep the partial output consistent with the schema remain available for selection. This ensures that final outputs are both syntactically valid and semantically aligned with the intended task requirements. This design mitigated refusal behaviors and promoted consistent formatting across different models, addressing reproducibility issues observed in earlier work (Azzopardi and Moshfeghi, 2024; Röttger et al., 2024).

For every target group, we compare the classification behavior of left- and right-oriented personas, measuring the average proportion of instances labeled as hate speech. Given 25,000 statements and 40 personas, this amounts to a total of 1,000,000 inferences per model. This setup allows us to test whether persona-induced alignment reproduces the partisan effects previously observed in pretraining-based studies.

Computational resources. All experimental conditions were executed on a single H100 GPU. Each run required approximately one and a half hours to complete, resulting in a total runtime of roughly 5 hours across the full set of experiments.

6.2. Results

Table 3 reports the average proportion of hate speech detected across target groups for left- and right-oriented personas, along with odds ratios quantifying systematic differences.

Overall persona effects. Across all three LLMs, left-oriented personas consistently yield higher detection rates than right-oriented ones. This effect is evident not only for minority groups such as BLACKS, MUSLIMS, JEWS, and LGBTQ+, but also for majority groups including CHRISTIANS, MEN, and WHITES. The uniformity of this effect runs counter to our hypothesis that right-leaning personas would be more protective of majority groups. Instead, persona-conditioning on the left systematically raises sensitivity to hateful content, indicating a general tightening of classification thresholds rather than selective group protection. One possible explanation is that the relatively small parameter size of these models limits their ability to adopt nuanced persona

[2]We used the implementation from vLLM, though other toolkits offer equivalent functionality.

perspectives, producing broad shifts in classification behavior rather than the group-specific differences anticipated—a pattern also noted by Civelli et al. (2025a). Investigating the precise mechanism behind this asymmetry lies beyond the scope of the present study, but future work could leverage our framework with larger-scale models to test whether the hypothesized group-specific protection emerges under more expressive architectures.

Variation across models. Although the gap between left and right is consistent, its magnitude differs by model family. Mistral shows the strongest divergence (across all target groups overall OR = 1.364, $p < 0.001$). By contrast, Llama exhibits the smallest persona effect (OR = 1.068, $p < 0.001$), while Qwen falls in between (OR = 1.073, $p < 0.001$). When considering absolute protection levels, however, a different pattern emerges: averaging overall detection rates across left- and right-conditioned personas, Llama achieves the highest baseline detection (0.493), followed by Mistral (0.359), and Qwen the lowest (0.199). While odds ratios primarily capture the models' relative responsiveness to persona-conditioning, raw detection levels reflect their intrinsic tendency to classify statements as hateful.

Model-specific protection of Whites. Across models, detection rates are generally highest for statements targeting BLACKS, making this category the most consistently protected. The main exception arises with Llama, where WHITES receive the strongest protection (0.656 under left personas), surpassing BLACKS as the top category. This unusually high value—the largest single entry in the table—may reflect the model's U.S.-centric training distribution, where discourse around race often centers explicitly on contrasts involving WHITES. By contrast, Qwen and Mistral exhibit substantially lower absolute detection for WHITES, aligning more closely with the overall trend that prioritizes minority-group protection.

Remarks. Together, the results convey two concise points. First, persona-conditioning produces an asymmetric within-model effect: left-conditioned personas yield higher hate-speech detection rates relative to right-conditioned personas, consistent with a general tightening of classification thresholds rather than selective protection of particular groups. Second, the effect's magnitude and the models' baseline tendencies differ: Mistral is the most responsive to persona shifts, Llama shows a higher baseline detection with notable outliers, and Qwen is comparatively stable.

7. Conclusion

This paper introduces a two-step framework for the systematic evaluation of perspective alignment in LLMs. First, we present SUBDATA, an open-source library that standardizes heterogeneous datasets by unifying annotation schemes and demographic taxonomies, thereby enabling consistent evaluation across subjective NLP tasks. Second, we propose a theory-driven evaluation approach that leverages these standardized datasets to test hypotheses about how differently aligned models behave in downstream applications. We demonstrate the practical value of this framework through an experimental use case. This example illustrates how SUBDATA not only provides a resource for data integration but also facilitates rigorous, theory-grounded experimentation on LLMs perspective alignment.

Future Extensions. The most immediate extension of SUBDATA is the inclusion of additional datasets, both those that we may have overlooked in our initial collection as well as those that are yet to be released. In parallel, we aim to cultivate a community of researchers interested in aligning LLMs with diverse human viewpoints, which would naturally accelerate the inclusion of additional datasets.

Moreover, we plan to broaden the scope of SUBDATA by introducing additional subjective constructs. Our next priority is misinformation, for which we have already compiled an initial collection of datasets that will soon be accessible through the library. For misinformation datasets, the connection between theory and testable hypotheses will be grounded in the topical domain of the claims being evaluated. Specifically, different domains (e.g., politics, public health, or climate) are known to elicit systematically different judgments depending on individuals' prior beliefs and ideological commitments. This allows misinformation to be operationalized as a subjective construct, where disagreement is not merely noise but reflects underlying differences in perspective. Consequently, variation in model predictions across perspectives can be interpreted as meaningful evidence of alignment (or misalignment) with distinct human viewpoints.

Ultimately, we intend to develop an alternative approach for evaluating LLM alignment with different human viewpoints, focusing on annotator characteristics rather than instance features. Through these initiatives, we aspire to evolve SUBDATA into a comprehensive multi-construct benchmark suite for evaluating how well LLMs align with humans across various downstream tasks.

Limitations

While the initial implementation of SubData focuses on hate speech detection, this narrow scope reflects the availability of suitable datasets. We chose to release the library early because alignment research is advancing rapidly but lacks standardized resources for downstream evaluation. Even in its current state, we believe SubData offers immediate value for studying LLM alignment with diverse perspectives.

Our unified taxonomy required pragmatic mapping choices that inevitably involve subjective judgment. Challenges include the existence of target groups in the literature that conflate targets from different categories (e.g., "LGBTQ+" for minority gender identities and sexual orientations), targets that are placed into different categories in different original datasets (e.g., "mexicans" either put into a race—latin—or an origin category) and intersectional groups (e.g., "blacks, women"). We applied our principles carefully to balance specificity and generalizability. While the mapping process is manual and limits scalability, we argue this effort is both necessary and valuable: meaningful taxonomies for subjective constructs require domain expertise and contextual sensitivity that automated methods often miss. It is also a one-time investment with lasting benefits, and our taxonomy already aligns with independent efforts, suggesting emerging consensus. Future versions may incorporate semi-automated clustering or embedding-based methods to propose candidate mappings, with human oversight ensuring contextual validity. At the same time, SubData supports customization—researchers can adapt, extend, or redefine taxonomies as needed—helping to mitigate the limitations of any single framework.

In our experimental setup, we are contrasting personas from the extremes of the ideological spectrum. While we made this decision deliberately in order to increase contrast and establish the functioning of our methodological contribution, future work should extend this analysis to intermediate and mixed identities, where distinctions are more subtle and potentially more informative.

Finally, the library inherits annotation errors and biases from its source datasets. SubData aggregates existing annotations without re-labeling or quality control, so we encourage users to verify annotation quality and consult original documentation where appropriate.

Ethical Considerations

While SubData provides valuable datasets for evaluating LLMs perspective alignment, we acknowledge potential ethical concerns. The library's aggregation of hate speech datasets creates a concentrated collection of offensive content that could be misused to train hateful models or generate toxic content. Additionally, our framework's ability to test how differently-aligned LLMs classify content targeting specific demographics could be misused to intentionally create biased systems. We emphasize that SubData's purpose is to improve evaluation transparency and understanding of perspective alignment, not to enable harmful applications. We recognize that the target groups represented in these datasets face real discrimination and harassment. Research using SubData should be conducted with sensitivity to the lived experiences of these communities, and findings should be communicated in ways that avoid reinforcing harmful stereotypes or creating additional psychological harm.

Dataset Licensing and Access. SubData does not host, mirror, or redistribute any dataset included in its unified taxonomy. All data are obtained directly from their official public distribution endpoints (e.g., GitHub, Kaggle, project websites), and the library simply provides convenience functions that download or load these datasets using the user's own credentials when required. For datasets that mandate registration, authentication, or explicit acceptance of license terms, SubData does not bypass these access controls; users must obtain the data themselves under the original license conditions. Because the library only performs local standard ization of datasets that users already lawfully acquired, it does not create or distribute any derivative dataset, and all licensing obligations remain governed by the original providers. The licenses, access constraints, and redistribution permissions of the datasets available via SubData are detailed in Appendix (Table 4).

Acknowledgements

This work is partially supported by an Australian Research Council (ARC) Future Fellowship Project (Grant No. FT240100022) and by the Swiss National Science Foundation (SNSF) under contract number CRSII5_205975.

References

Ahmed Agiza, Mohamed Mostagir, and Sherief Reda. 2024. Politune: Analyzing the impact of data selection and fine-tuning on economic and political biases in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 2–12.

Shayan Alipour, Indira Sen, Mattia Samory, and Tanu Mitra. 2025. Robustness and confounders in the demographic alignment of LLMs with human perceptions of offensiveness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22025–22047, Vienna, Austria. Association for Computational Linguistics.

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):337–351.

Leif Azzopardi and Yashar Moshfeghi. 2024. Prism: a methodology for auditing biases in large language models. *arXiv preprint arXiv:2410.18906*.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Pietro Bernardelle, Stefano Civelli, Leon Fröhling, Riccardo Lunardi, Kevin Roitero, and Gianluca Demartini. 2025a. Political ideology shifts in large language models. *arXiv preprint arXiv:2508.16013*.

Pietro Bernardelle, Leon Fröhling, Stefano Civelli, Riccardo Lunardi, Kevin Roitero, and Gianluca Demartini. 2025b. Mapping and influencing the political ideology of large language models using synthetic personas. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 864–867.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

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

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2025. PERSONA: A reproducible testbed for pluralistic alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11348–11368, Abu Dhabi, UAE. Association for Computational Linguistics.

Kai Chen, Zihao He, Jun Yan, Taiwei Shi, and Kristina Lerman. 2024. How susceptible are large language models to ideological manipulation? pages 17140–17161.

Stefano Civelli, Pietro Bernardelle, and Gianluca Demartini. 2025a. The impact of persona-based political perspectives on hateful content detection. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1963–1968.

Stefano Civelli, Pietro Bernardelle, Nardiena A Pratama, and Gianluca Demartini. 2025b. Ideology-based llms for content moderation. *arXiv preprint arXiv:2510.25805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. 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 3226–3240, Online. Association for Computational Linguistics.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Jan Fillies and Adrian Paschke. 2025. Improving hate speech classification with cross-taxonomy

dataset integration. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 148–159, Albuquerque, New Mexico. Association for Computational Linguistics.

Leon Fröhling, Gianluca Demartini, and Dennis Assenmacher. 2025. Personas with attitudes: Controlling LLMs for diverse data annotation. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 468–481, Vienna, Austria. Association for Computational Linguistics.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2024. Human and llm biases in hate speech annotations: A socio-demographic analysis of annotators and targets. *arXiv preprint arXiv:2410.07991*.

Patrick Haller, Ansar Aynetdinov, and Alan Akbik. 2024. OpinionGPT: Modelling explicit biases in instruction-tuned LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 78–86, Mexico City, Mexico. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Zihao He, Minh Duc Chu, Rebecca Dorn, Siyi Guo, and Kristina Lerman. 2024. Community-cross-instruct: Unsupervised instruction generation for aligning large language models to online communities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17001–17019, Miami, Florida, USA. 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*.

Jigsaw. 2019. Jigsaw unintended bias in toxicity classification. Accessed: 2024-11-19.

Gunther Jikeli, Sameer Karali, Daniel Miehling, and Katharina Soemer. 2023a. Antisemitism on Twitter: A Dataset for Machine Learning and Text Analytics.

Gunther Jikeli, Sameer Karali, and Katharina Soemer. 2023b. Hate Speech and Bias against Asians, Blacks, Jews, Latines, and Muslims: A Dataset for Machine Learning and Text Analytics.

Mehdi Khamassi, Marceau Nahon, and Raja Chatila. 2024. Strong and weak alignment of large language models with human values. *Scientific Reports*, 14(1):19399.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

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.

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.

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.

Julian Risch, Philipp Schmidt, and Ralf Krestel. 2021. Data integration for toxic comment classification: Making more than 40 datasets easily accessible in one unified format. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 157–163, Online. Association for Computational Linguistics.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Jens Rupprecht, Leon Fröhling, Claudia Wagner, and Markus Strohmaier. 2025. German general personas: A survey-derived persona prompt collection for population-aligned llm studies. *arXiv preprint arXiv:2511.21722*.

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.

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.

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.

Pavel Shvaiko and Jérôme Euzenat. 2011. Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, 25(1):158–176.

Brittany C Solomon, Matthew EK Hall, Abigail Hemmen, and James N Druckman. 2024. Illusory interparty disagreement: Partisans agree on what hate speech to censor but do not know it. *Proceedings of the National Academy of Sciences*, 121(39):e2402428121.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46280–46302.

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021a. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021b. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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 1667–1682, Online. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa

Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

A. Prompt Template

Prompt Template

You are an expert content analyst specializing in identifying hate speech according to specific guidelines. Your task is to analyze the provided text strictly from the perspective of **[PERSONA]**.

Based on your analysis, you must generate a single, valid JSON object that adheres to the following schema and logic. **Do not output any text other than the JSON object.**

Analysis Logic:

1. **Perspective:** Embody the given persona. Your entire analysis must reflect their likely values, biases, worldview, and sensitivity level to hateful content.
2. **Classification:** Determine if the persona would consider the text to be hate speech. This determines the value of `is_hate_speech`.

JSON Schema:

- `is_hate_speech` (*string*): Must be one of `"true"` or `"false"`.

Example 1 (Hateful):

```
{
    "is_hate_speech": "true"
}
```

Example 2 (Not Hateful):

```
{
    "is_hate_speech": "false"
}
```

Persona: **[PERSONA]**
Text: **[TEXT]**

Now, analyze the text from the persona's perspective and generate the JSON object.

Across all LLMs investigated, we employed a standardized prompt format to ensure comparability of results and to minimize confounding effects introduced by variations in wording or structure. We wrapped all inputs in the respective chat template of each model family. This preserved the conversational format they were optimized for during fine-tuning, reducing the risk of degraded performance or altered behavior due to mismatched prompting styles. For each combination of persona and statement (40 personas × 25,000 statements = 1,000,000 total pairs), we generated the corresponding prompt by substituting `[TEXT]` and `[PERSONA]` placeholders in the template with the actual text.

Dataset	License	Gated?	Redistribution Allowed?	Sourcing Method
Fanton et al. (2021)	Research-use only (not for redistribution)	×	×	Download via authors' GitHub (CONAN)
Hartvigsen et al. (2022)	Not explicitly stated	✓	×	Download via Hugging Face (requires form + token)
Jigsaw (2019)	CC0 1.0	×	✓	Download from Kaggle (Jigsaw Unintended Bias)
Jikeli et al. (2023a)	CC BY 4.0	×	✓	Download from Zenodo (open DOI)
Jikeli et al. (2023b)	CC BY 4.0	×	✓	Download from Zenodo (open DOI)
Mathew et al. (2021)	MIT	×	✓	Download via GitHub (HateXplain)
Röttger et al. (2021)	CC BY 4.0	×	✓	Download via GitHub (hatecheck-data)
Sachdeva et al. (2022)	CC BY 4.0	×	✓	Download via Hugging Face (open dataset)
Vidgen et al. (2021a)	CC BY 4.0	×	✓	Download via GitHub (Learning from the worst)
Vidgen et al. (2021b)	CC BY 4.0	×	✓	Download from Zenodo (open DOI)

Table 4: Licensing, access constraints, and redistribution permissions for datasets included in SUBDATA.

ChatGPT, why can't anyone afford a house? On the Effects of LLM pre-annotation on Annotator Subjectivity

Emilie Francis[†], **Céline Leuzinger**[†], **Ricardo Muñoz Sánchez**[†], **Lee Gauthier**[‡]
[†]University of Gothenburg, Sweden
{emilie.francis,celine.leuzinger,ricardo.munoz.sanchez}@gu.se
[‡]Independent Researcher
lee.d.gauthier@gmail.com

Abstract

Large language models (LLMs) have often been proposed as substitutes for human annotators in a variety of tasks. At the same time, there has been increased focus on the role that human subjectivity and perspective plays in data annotation. To avoid eliminating the human role in annotation entirely, the use of LLMs for pre-annotation has been suggested as an alternative approach. In this paper, we explore to which degree this approach affects subjectivity of social media annotation in English. We focus on comments regarding the current status of the housing market and label them for concern level, factors affecting housing affordability, and aspects that authors claim either exacerbate or improve the situation. To investigate this, we design an experiment involving two rounds of annotation: the first, a dataset annotated by humans only; and the second, a dataset with LLM pre-annotations curated by the same human annotators. We observe that the second setting leads to much higher agreement, as well as significant changes in label distribution and co-occurrence. Similar shifts do not appear in the LLM labels. Our findings show that use of LLMs in the annotation process leads to convergence in annotations and, thus, to an erosion of human subjectivity.

Keywords: subjectivity, annotation, agreement, LLMs, social media analysis

1. Introduction

NLP has experienced a drastic paradigm shift with the advent of highly performant Large Language Models (LLMs). Autoregressive language models have become more commonplace throughout the entire machine learning pipeline, reducing the need for multiple interconnected pieces. Although this is not a new trend, autoregressive generative language models have been increasingly used to label data (Karim et al., 2025) and automated evaluation (Bavaresco et al., 2025).

However, the use of LLMs comes with its own risks and challenges. For instance, they are known to hallucinate Ji et al. (2024) and reproduce social biases Gallegos et al. (2024). There have been multiple proposed solutions when using these systems for tasks like data labelling to ensure the quality and reliability of the system's outputs. Two of these are 'human in the loop' (Pangakis and Wolken, 2024; Wang et al., 2025) and 'hybrid intelligence' (Dellermann et al., 2019).

A major flaw of such approaches is 'anchoring bias', which shows that humans tend to favour the first option presented (Tversky and Kahneman, 1974). Another influencing effect is 'automation bias', which specifies that humans overly rely on machine output for automated systems they consider trustworthy (Dzindolet et al., 2003). Such biases can have a strong impact on annotation outcomes, as human reviewers may be tempted to agree with LLM suggestions. This could in turn lead to what Schroeder et al. (2025) describe as a 'homogenization of insights', reducing annotation diversity and impacting downstream performance. This is particularly problematic for subjective tasks, where an accurate depiction of reality would naturally entail a variety of annotator opinions and judgments (Wan et al., 2023).

To determine the impact of anchoring and/or automation biases, we measure the effect of LLM annotation on human subjectivity. We focus on three research questions: *(i) Does usage of LLMs as a tool for pre-annotation result in convergence of human-annotated labels? (ii) Are different sets of labels more likely to be used when LLMs are involved in the annotation process? (iii) Are humans more efficient at data labelling in terms of time when curating LLM-generated labels?*

To answer these questions, we gather a dataset of social media comments discussing housing affordability in major cities of six English speaking countries. We engage three community annotators in two experimental annotation settings. In the first experiment, annotators are asked to annotate 500 comments for four label categories (*concern score, factor, aspect-improvement, aspect-exacerbate*). In the second setting, GPT-4.5 is used to label a different set of 500 comments whose output is curated by the same annotators.

Our results show that use of an LLM for pre-annotation leads to a drastic decrease in terms of annotation time. However, this comes at a price concerning annotation quality. There is a non-

Proceedings of NLPerspectives @ LREC 2026, pages 98–111
12 May 2026.

negligible increase in inter-annotator agreement and a significant shift in terms of label distribution.

For instance, labels that denote objective factors impacting housing prices are much more common when LLMs are involved compared to human-only annotations. These results suggest that usage of LLMs in the annotation process strongly influences human subjectivity. Our contributions can be summed up as follows:

1. A dataset annotated with labels for subjective measures of concern and commonly mentioned features of housing affordability in several major cities
2. An evaluation of the use of LLM pre-annotated data for annotation in a subjective task, for both prescriptive and descriptive ordinal and categorical labels
3. An exploration of changes in annotation patterns when annotators are provided with LLM pre-annotated data

The following section presents and overview of the use of LLMs for dataset labelling, both as a replacement for human annotators and as assistants, as well as commonly noted challenges in the use of LLMs for such work.

2. Background

In the past few years, generative language models have been proposed as a alternative for human annotators. The core argument behind this is that it takes less resources to automate the annotation process, both in terms of time (Choi et al., 2024) and money (Wang et al., 2021). This approach has been followed to label data in fields such as argumentation mining (Lindahl, 2025) and the social sciences (Ziems et al., 2024), among others. (Gao et al., 2023) describe the LLM as a mediator facilitating discussion between experts, noting that its usage leads to increased agreement.

However, several arguments have been leveraged against taking LLMs as if they were human annotators. Atreja et al. (2024) and Baumann et al. (2025) show that they are highly sensitive to prompt variation, which can lead to major changes in label distribution. Not only that, but LLMs are known to suffer from hallucinations (Ji et al., 2024) and to reproduce social biases (Gallegos et al., 2024). Further, the use of LLM-generated labels can lead to lower performance across several kinds of tasks (Li et al., 2023; Pangakis and Wolken, 2024)

Both 'human in the loop' and 'hybrid intelligence' approaches have been suggested as a way to get around some of these issues (Schroeder et al., 2025; Wang et al., 2025) For instance, Ziems et al. (2024) describe human annotators as 'key' to avoid LLM biases and hallucination in social sciences datasets. However, the use of humans as curators (as opposed to annotators) poses a new set of problems. Humans are known to suffer from different cognitive biases, among them anchoring and automation biases.

Anchoring bias refers to the human tendency of relying on the first piece of information presented, even if it is irrelevant to the task at hand (Tversky and Kahneman, 1974). This phenomenon is widely studied, and has been shown occur in a variety of decision making processes, such as purchasing decisions, legal judgments or time estimation (Furnham and Boo, 2011).

Automation bias refers to the human tendency to favour suggestions coming from automated systems, even if they are in direct contradiction with information that is known to be true by the decision-maker (Dzindolet et al., 2003). Wilcox (2023) argues that human in the loop approaches lead to increased risks of automation bias while reducing accountability.

Choi et al. (2024) showed that human curators heavily rely on LLM annotation even in highly specialized settings. Meanwhile, Schroeder et al. (2025) found that human annotators have a strong tendency to follow LLM suggestions even in subjective tasks, which in turn leads to higher agreement and significant shifts in terms of label distribution.

High agreement between annotators has been historically regarded as the gold standard in machine learning (Basile et al., 2020). However, this ignores the fact that most NLP tasks have a subjective aspect to them. As such, data cleaning and harmonization leads to less rich datasets and risks erasing underrepresented voices (Klenner et al., 2020). Thus, factors that force higher agreement can be detrimental to our data, among them anchoring and automation biases.

This can be seen when taking LLM-generated labels as if they were actual annotations. Gao et al. (2023) show that using humans to curate LLM-generated labels strengthened agreement and diminished the diversity of labels. Schroeder et al. (2025) echo these insights, noting that it homogenizes labels and reduces the diversity of judgments. Moreover, Choi et al. (2024) observe that topics selected by the LLM tend to be broader and less nuanced than topics selected by humans.

3. Methodology and Data

We use a three step process to compare patterns in annotation before and after exposure to data augmented with LLM pre-annotations. Taking a raw dataset (D) of 1000 English comments on Reddit from local forums for several major cities experienc-

ing challenges with housing affordability, we split into two sets of 500 (D^1 and D^2).

In the first step, a team of three annotators (*A*, *B*, and *C*) annotate the first set of comments without LLM pre-annotation (D_h^1). We define four distinct categories for annotation:

1. **Concern Score:** how concerned the comment's author appears to be about not being able to find or maintain adequate housing in the city
2. **Factor:** an objective set of measures for housing affordability used in OECD and EU countries
3. **Aspect-Exacerbate:** aspects that a comment's author claims worsen affordability in the housing market
4. **Aspect-Improve:** aspects that a comment's author claims improve affordability in the housing market

The first two categories follow a prescriptive paradigm (Röttger et al., 2022), while the remaining two are descriptive and developed collaboratively by the annotators during annotation. Details of the annotation process are elaborated on in Section (3.2).

In the second step, annotators and researchers agree on standardized guidelines which include the prescriptive categories and a refined set of the descriptive labels developed in step one. We use few-shot prompting with the standardized guidelines to generate labels for an unseen set of 500 comments (D_l^2) and the initial 500 comments with GPT-4.5 (D_l^1). As LLMs tend to struggle with multi-task prompts (Ma et al., 2025), this step was broken down into four tasks using one few-shot prompt for each label category.

In the final step, annotators independently curated the model output from step two based on the standardized guidelines. We then compare inter-annotator agreement between annotators in the human-only annotated dataset (D_h^1) with that of the human-curated dataset (D_h^2).

Measuring Agreement: Inter-annotator agreement for ordinal labels (*concern score*) was computed with Krippendorff's alpha and agreement between categorical labels was computed with Sørensen-Dice similarity. While Jaccard similarity is more widely used to measure agreement for multi-label categorical annotation tasks in NLP, it is designed for only two annotators. As averaging pairwise Jaccard scores to report agreement between three annotators risks information loss, we use three-way Sørensen-Dice similarity instead (Diserud and Ødegaard, 2007; Magurran, 2003).

Sørensen-Dice similarity, widely used in ecology to measure similarity of site composition, has been extended from a two-way measure to accommodate three or more entities (Diserud and Ødegaard, 2007). The general formula for a multiple entity similarity measure is defined as:

$$S_T = \sum_{i<j} a_{ij} - \sum_{i<j<k} a_{ijk} + \sum_{i<j<k<l} a_{ijkl} - \cdots$$

$$C_S^T = \frac{T}{T-1}\left(\frac{S_T}{\sum_i a_i}\right)$$

Where a_i is the number of variables for entity A_i, a_{ij} is the number of variables shared by entities A_i and A_j, and so on. T is the total number of entities. The Sørensen-Dice similarity measure lies between 0 and 1, where 0 is no agreement and 1 is perfect agreement. For a $T = 3$ measure of similarity between annotators A, B, and C, we calculate Sørensen-Dice similarity with:

$$S = \frac{3}{2}\left(\frac{ab + ac + bc - abc}{a + b + c}\right)$$

Where ab is the number of labels shared by annotators A and B, ac is the number of labels shared by A and C, etc. Inter-annotator agreement is calculated for each comment and averaged over the entire annotation set for both D_h^1 and D_h^2.

Label Patterns: We explore differences in label patterns between D_1 and D_2 across two parameters: frequency and label co-occurence. We compare the frequencies of each label, for each annotator for D_h^1 and D_h^2. We check for statistical significance and effect size using chi-square test. We also compare the frequencies for each label for the LLM-annotated data, D_l^1 and D_l^2.

To analyze differences in label co-occurrence, we use Fisher's Exact test in pairs of labels where at least one showed statistically significant changes for that annotator. We then calculate the effect size of this co-occurrence by calculating φ of the statistic:

$$\varphi = \sqrt{\frac{\text{statistic}}{\text{\# of observations}}}$$

We can interpret values of φ as follows: 0.1 and lower is small, between 0.1 and 0.3 the effect is medium, and an effect size of 0.3 or more is large.

Time: Finally, we also compare time in hours taken to complete steps one and three by each annotator. The following sections describe the dataset and the annotation process.

Local Subreddit	Comments
Vancouver	198
Toronto	189
Los Angeles	154
San Francisco	102
Melbourne	95
London	88
Sydney	88
Bristol	35
Auckland	29
Dublin	22

Table 1: Number of comments per local subreddit included in the full dataset of 1000 comments.

3.1. Data

We select six English speaking countries identified among the top locales for housing cost burden in OECD reports: *U.S.A., Canada, U.K., Ireland, Australia, and New Zealand* (OECD, 2025a).

For each country, we select the largest two cities by population and collect threads from the local subforums (subreddits) from the Reddit PushShift Dataset (Baumgartner et al., 2020). All comments are in English, dating from 2013 to 2023, and belong to one of the following subreddits: *r/LosAngeles, r/SanFrancisco, r/Vancouver, r/Toronto, r/London, r/Bristol, r/Sydney, r/Melbourne, r/Dublin, and r/Auckland*. For Ireland and New Zealand, only the most populous city was included due to lack of subreddit content for the second most populous city.

To restrict data to discussions on housing and affordability, we filter thread titles based on a list of housing terms. This list was developed by a community panel of volunteers, at least one panel member local to each country in the study. Threads whose title did not contain at least one term from the list were removed.

After automatic filtering, we selected ten threads from each city based on number of responses. The titles of these threads were manually reviewed by the authors to remove any that did not explicitly discuss housing affordability, such as advertisements for roommates or apartments. This resulted in five threads for each local subreddit.

For the remaining threads, we remove all comments with a token count lower than 100[1] to ensure there would be enough content for annotators to judge. This yielded 5,000 comments from which 1000 were randomly selected, 500 for human-only (D_h^1) annotation and 500 for pre-annotation with the LLM (D_h^2).[2]

All comments were cleaned to remove Reddit markdown formatting and replace personally identifiable information, such as usernames, with placeholders.

3.2. Dataset Annotation

Annotation was carried out by a team of three volunteers, each of which have spent one or more years living in one of the cities in the study. Two annotators are women in their late 20s to early 30s from a middle-class background in mid-sized cities, while the third is a man in his thirties from a small-town working class background.

The dataset includes four categories of labels: *concern score, factor, aspect-improve, and aspect-exacerbate*. As described in Section 3, two categories (*concern score* and *factor*) are prescriptive labels based on the *More Effective Social Protection for Stronger Economic Growth Survey* (OECD, 2025b) and *Building for a better tomorrow: Policies to make housing more affordable Brief* (OECD, 2021) reports respectively. For our dataset, we combined the survey responses 'not so concerned' and 'somewhat concerned' from OECD (2025b) into the 'mixed' label. The four labels for concern score are defined below:

1. **Off Topic:** The comment does not talk about housing/housing affordability in any way.
2. **No Concern:** The comment does not appear to express concern toward housing at all. May deny there is a problem or attempt to invalidate others' concerns and refute claims.
3. **Mixed:** The comment appears to express some concern, but is mixed. Usually agrees there is a problem, but might discuss the topic more analytically.
4. **Concern:** The comment expresses clear concern toward the housing situation and affordability.

The categorical labels, *factor*, *aspect-exacerbate* and *aspect-improve*, and their corresponding definitions are presented in Appendix A. The two descriptive categories were developed in a collaborative document using open coding (Khandkar, 2009). Additionally, our annotation procedure followed the perspectivist paradigm (Basile et al., 2020; Cabitza et al., 2023). That is, we kept all labels and did not aggregate annotations in any manner.

Two labels were annotator specific. The first, 'money', was used by one annotator in the aspect-improve category, but was later merged into the 'bootstraps'[3] label. The second, 'demand', was only

[1] Based on the 50th percentile for comment length.

[2] And 7 to the Dwarf Lords who became wealthy beyond measure.

[3] From the phrase "pull oneself up by one's bootstraps", meaning "getting oneself out of a difficult situation only with one's own effort".

Label Category	Krippendorff's Alpha (α)	
	human-only (D_h^1)	LLM-assisted (D_h^2)
Concern	0.63	0.72

(a) Inter-annotator agreement for ordinal labels.

Label Category	Sorensen-Dice Similarity (S)	
	human-only (D_h^1)	LLM-assisted (D_h^2)
Factor	0.70	0.86
Aspect (Exacerbate)	0.78	0.84
Aspect (Improve)	0.82	0.88

(b) Inter-annotator agreement for categorical labels.

Table 2: Inter-annotator agreement for the different labels. Note that agreement for ordinal labels is reported in terms of Krippendorff's Alpha, while Sørensen-Dice Similarity is used for categorical ones.

used by one annotator in the *aspect-exacerbate* category.

We created two versions of the dataset, a standardized and non-standardized version. The non-standardized version contains all labels used by annotators to preserve the full variety of annotator perspective for future tasks. The standardized version was created for ease of comparing annotator agreement. This version includes *concern score*, *factor*, and the set of aspect labels agreed upon by all annotators mentioned in Section (3). Labels with fifteen or fewer instances were removed from the standardized label set and subsequent analysis. Annotators used only the the standardized label set for curation of the LLM output in D_h^2. The set of standardized labels and corresponding definition are presented in Appendix A.

In addition, annotators were asked to record time taken to complete both D_h^1 and D_h^2.

In the following section, we compare the results of inter-annotator agreement between D_h^1 and D_h^2, as well as the difference in time taken to complete each task. We also present an analysis of label distribution to explore differences in annotation patterns between D^1 and D^2. Finally, we note specific instances in which annotators believed their individual experience or contextual knowledge were an asset in annotation and how their annotations differed from the LLM and each other.

4. Experimental Results

4.1. Inter-Annotator Agreement

We observe a large increase in inter-annotator agreement between D_h^1 and D_h^2 for both ordinal labels with Krippendorff's Alpha and categorical labels with Sørensen-Dice Similarity. The results of inter-annotator agreement are reported in Table 2.

For both D_h^1 and D_h^2, *concern score* had the lowest agreement. Unlike labels in the other categories, which may be triggered by specific vocabulary, *concern score* is purely subjective as it requires annotators judge the emotional state of a comment's author. The agreement for this category increased by nearly ten points from D_h^1 ($\alpha = 0.63$) to D_h^2 ($\alpha = 0.72$), showing that annotations became less varied when annotators were provided with pre-annotated data.

The *factor* category showed the biggest increase comparing the D_h^1 ($S = 0.7$) with D_h^2 ($S = 0.86$). A potential explanation for this may be that, unlike the aspect labels which are specific, *factor* labels represent broader groups of factors for housing affordability identified by the OECD. Additionally, annotators may have been less certain of their interpretation of these labels as they were pre-defined rather than developed based on their own understanding.

The *aspect-exacerbate* and *aspect-improve* categories showed the highest agreement in D_h^1 ($S = 0.78$ and $S = 0.82$) and the smallest increase when comparing D_h^1 with D_h^2 ($S = 0.84$ and $S = 0.88$). High agreement for these categories overall is likely due to the presence of triggering vocabulary, such as 'NIMBY' for the 'NIMBYism'[4] label. However, there is small (0.06) increase which suggests that annotators have a slight tendency to accept the LLM output for these labels as well.

Comparing inter-annotator agreement between D_h^1 and D_h^2 reveals that disagreement is considerably reduced between annotators when provided with pre-annotated data. This suggests that annotators are more likely to accept the provided labels, resulting in a loss of subjectivity as annotations become overly influenced by the LLM.

[4]NIMBY - 'not in my backyard'. Refers to people who oppose development near their owned property.

Label	Odds Ratio		
	A	B	C
Real Price	**1.43**	**1.53**	**3.70**
Quality	**1.56**	**2.81**	**7.56**
Availability	**1.64**	**1.90**	**5.30**
Housing to Income	1.18	1.35	**3.25**
Building	**1.78**	**2.07**	**4.92**
Bootstraps	1.28	**2.74**	**0.59**
Government Policy (I)	1.21	1.42	**2.05**
Relocation	**2.71**	**3.58**	**2.53**
Wage Price Imbalance	1.48	**1.60**	**5.46**
Government Policy (E)	1.37	**1.78**	**1.60**
Cost of Living	0.90	1.14	**4.40**
Foreign Investment	**0.52**	0.70	0.64
NIMBYism	0.64	0.84	0.86
The Rich	1	1.20	1.48

Table 3: Odds ratio per label and annotator (A, B, C). Bold values are statistically significant.

Label	Odds ratio
Availability	0.97
Housing to Income	0.96
Quality	1.23
Real Price	1.16
Bootstraps	0.70
Building	0.98
Government Policy (I)	0.88
Relocation	**1.66**
Cost of Living	1.02
Foreign Investment	0.63
Government Policy (E)	1.05
NIMBYism	0.71
The Rich	0.82
Wage Price Imbalance	1.03

Table 4: Odds ratio per label for the LLM-only annotation. Bold values are statistically significant.

4.2. Label Frequency

We also take a look at changes in terms of individual labels between D_h^1 and D_h^2. To study changes in label frequency, we compute the frequency of each label in both experimental settings for each annotator. We then perform a chi-square test (Pearson, 1900) to check for statistically significant changes in label frequency. We also compute the odds ratio for effect size. We report these values in Table 3.

To ensure these effects are not due to differences in the data, we also compare label frequency between D_l^1 and D_l^2. The odds ratio for each label is reported in Table 4.

The odds ratios comparing D_l^1 and D_l^2 are relatively close to 1, with values ranging from 0.70 to 1.66. The odds ratios for the human annotations, D_h^1 and D_h^2, display a wider variation, with values ranging from 0.52 to 7.56 across annotators. This suggests that the LLM was more consistent across datasets than humans.

Three out of the four labels that are significantly more frequent for all annotators are factors ('real price', 'quality', 'availability'). We also observe cross-annotator differences in label frequency. For annotator A, there are six labels whose distributions more closely approximated those from the LLM in D_h^2. The number of such labels are nine and twelve for annotators B and C, respectively. This shows that while all annotators were influenced by the LLM, the degree of such influence varies between individuals.

We also look at label co-occurrence to determine whether these changes in labels led to changes in how often they appear with each other. We focus on labels where the chance of them co-occurring was statistically significant and the co-occurrence rate was medium or large. These labels can be found in Table 5 For all annotators, we see an almost complete change of labels that are likely to co-occur, with none of these changes including the label 'relocation', which was the only one that showed a statistically significant change in label distribution for the LLM.

4.3. Annotation Time

Table 6 shows the time in hours taken by each annotator to complete each of the two experiment sets. As we can see, there is a noticeable reduction in time taken to annotate the same number of samples when using an LLM for pre-annotation compared to human-only annotation. Even though this appears to contradict the findings of Schroeder et al. (2025), it goes in line with previous research on human curation of LLM-generated labels (e.g. Choi et al., 2024).

4.4. Annotator Observations

Annotations in D_h^1 were often influenced by annotator world knowledge and personal background, several examples of such instances are presented below. To preserve anonymity, all named entities in provided comment examples have been redacted to obfuscate location.

Although not explicitly mentioned in the text, all annotators added the label 'COVID' for the comment in (**X**). The annotators explained that they used the date of the comment as context. The LLM did not attach the 'COVID' label to this comment, which shows that this real-world contextual knowledge was not incorporated into the LLM output.

Annotator	Labels		Effect Size $\mathbf{D_h^1}$	Effect Size $\mathbf{D_h^2}$
A	NIMBYism	**Availability**	small	medium
B	NIMBYism	Government Policy (I)	medium	N/A
	Relocation	Cost of Living	N/A	Medium
C	**Building**	**Quality**	medium	N/A
	Building	**Availability**	medium	N/A
	Foreign Investment	**Government Policy (E)**	small	medium

Table 5: Changes in co-occurrence between the human-only annotation and the LLM pre-annotation setting. Labels in bold are those that showed a statistically significant change for that annotator between annotation rounds. N/A denotes that the co-occurrence was not statistically significant in that setting.

Annotator	$\mathbf{D_h^1}$	$\mathbf{D_h^2}$
A	15.5	9.2
B	10.4	6.7
C	6	2

Table 6: Annotation times per annotator in hours. D_h^1 refers to the human-only annotated dataset, and D_h^2 the LLM pre-annotated dataset.

> **X**: If this gets truly bad, the government will need to put mortgages and even rent on hold (as I think other countries have already done in response to people not being able to work).

In another comment (**Y**), annotator *C* gave the label 'concern' and the others (including the LLM) 'off topic'. While the comment does not explicitly mention housing, annotator *C* based their interpretation on personal experience discussing the housing situation with peers living in the city.

> **Y**: I would love to live somewhere else, if [...] cities/towns hadn't f*** themselves up to cater to driving everywhere. At least in [...] I can walk and bike to places, even if it's not that safe. Go elsewhere and you're trudging through parking lots and stroads.

Annotators also had very different interpretations of sarcasm. For the comment in (**Z**), annotators *B* and *C* gave the label 'off topic' while annotator *A* and the LLM gave the label 'concern'. Annotator *A* only applied the 'cost of living' for *aspect-exacerbate*, while the LLM used both 'cost of living' for *aspect-exacerbate* and 'bootstraps' for *aspect-improve*. Annotator *A*, local to the city the comment discusses, used personal experience from conversations with peers mocking the 'bootstraps' argument and did not judge it as a serious suggestion by the comment's author.

> **Z**: No kids, dual income. We eat spaghetti noodles with butter 21 meals a week and our favourite pastime is sleeping 12 hours so we can save on [...] bills. We have sex on a plastic sheet to cut down on laundry. I haven't smiled since 2012. Waste of calories. Just ten more years of this and we will be able to retire early in a paid off townhouse just outside [...]. The five years before my unfortunate heart attack are going to be epic.

Overall, annotator background and experience played a large role in their interpretation. Annotators also reported they were more confident in annotations for comments from locales in which they had lived or spent some time in.

5. Discussion

Despite the potential resource reduction for descriptive annotation paradigms, there are major disadvantages to using LLMs for pre-annotation in subjective tasks. Like Schroeder et al. (2025), Choi et al. (2024), and Gao et al. (2023), the results of our analysis show that LLM pre-annotation significantly influences human annotation in such a way that is detrimental to subjective tasks (Wan et al., 2023).

For all categories, we observe a large increase in inter-annotator agreement. This indicates that annotators were more likely to rely on LLM pre-annotations, regardless of whether labels follow a prescriptive or descriptive paradigm (Röttger et al., 2022).

We also observe significant differences in label frequency between the annotation rounds. *Factors* in particular are more frequent in the LLM curated output. *Factors* describe measures of housing affordability, and are arguably the most general labels in our annotation frameworks. A label such as 'real price' can be understood in a variety of ways (direct mention of a buying price, presence of adjective relating to costliness, direct mention of a price increase, etc.); the same can be said

about ‘quality’ (price-quality relationship, quality of the infrastructure, quality of life in the neighbourhood, etc.) and ‘availability’ (lack of housing, lack of affordable housing, direct mention of a number of units being built, etc.). *Factors* are already among the most common labels in the human-only annotation round; the sharp increase in their frequency in the curated output could indicate that the LLM has a tendency to ‘over-label’ with broad, general tags that can have a variety of interpretations. Human annotators, on the other hand, could have a finer interpretation of both the label and the comment at hand, and therefore chose not to use general labels as often as LLMs. These results are in line with Choi et al. (2024), who observe that LLMs tend to select broader topics than humans.

Such change in label frequency suggests that annotators are affected by a combination of anchoring and automation bias. As Klenner et al. (2020) argues, homogenization as a result of LLM pre-annotation risk erasing valuable perspective in subjective annotation tasks. As exemplified in Section 4.4, there were several instances in which either the annotators disagreed with each other or the LLM as an effect of real world knowledge or annotator experience. These insights may be sacrificed as a consequence of using LLMs to pre-annotate data.

These effects are not trivial. There has been a push in recent years in NLP to acknowledge the importance that annotator subjectivity plays, both in terms of representation and in terms of modelling (Cabitza et al., 2023). Even though the focus has often laid on non-aggregation of labels, there is also a risk of annotator subjectivity erosion by other means (e.g. anchoring and automation biases). Our results confirm that label convergence in LLM-assisted annotation is a significant problem for subjective tasks that greatly impacts both prescriptive and descriptive annotation paradigms.

While previous explorations into the use of LLMs for augmenting human annotation have shown mixed results concerning resource benefit (Schroeder et al., 2025; Choi et al., 2024), our results indicate that they do have the potential to reduce annotation time.

However, time reduction from LLM pre-annotation may depend on the type of task and annotation setup. Schroeder et al. (2025) used prescriptive labels in a multiple choice environment and reported no meaningful difference in time between annotations with and without LLM pre-annotation. In contrast, our annotation setup utilized a mix of prescriptive and descriptive annotation in an open environment wherein annotators could freely add and adjust labels as necessary. Potentially, time reduction from LLM pre-annotation is greater for descriptive annotation paradigms.

6. Conclusion

In this paper, we have presented an investigation into the use of LLMs for pre-annotation in a subjective task, concern regarding a crisis of housing affordability in several major cities in the Anglosphere. We design a three step experimental process to compare inter-annotator agreement and patterns in label distribution with and without the use of LLMs for pre-annotation. In the first step, annotators worked collaboratively to develop labels as they annotated a dataset of 500 comments without the use of LLMs. In the next step, we pre-annotated an unseen set of 500 comments plus the original 500 comments with an LLM. In the final step, annotators curate the output of the LLM pre-annotations. Annotators reported a large reduction in time taken to complete the same number of comments when provided with pre-annotations. However, results also showed a significant difference in both inter-annotator agreement and label distribution with the use of LLM pre-annotations. Our findings show that annotations tend to converge when exposed to LLM labels, indicating that annotators are more likely to accept the LLM output as is. This loss of subjectivity may have serious consequences for downstream tasks.

While providing annotators with LLM augmented data has the potential to speed up annotation time, it comes at the cost of averaging human perspective. When presented with LLM augmented data, annotations become more homogeneous and result in an erasure of annotator perspective. Therefore, we conclude that use of LLMs for pre-annotation is detrimental to tasks where the nuance and subjectivity of human diversity are valuable.

Future studies could explore the presence of automation bias by conducting blind curation, containing a mixture of human and LLM annotations. If the curators’ attitude does not change between human and LLM annotations, it implies that anonymising the source of the annotation is an effective preventer of automation bias. Conversely, if the curators correct human annotations more often than the LLM’s when knowing the source, it could imply that curators experience automation bias. Another follow-up study could test the influence of automation bias and anchoring bias, by presenting a choice of ordered annotations to the curators; if the curators tend to choose the first option presented, regardless of the source of the annotations, then anchoring bias is the predominant factor. If the curators tend to choose the LLM annotations, even when they are not presented first, then automation bias is more prevalent.

Limitations

This paper has a few limitations in the data and annotations. First, although the issue of housing affordability affects many regions globally, our dataset and annotations are focused only on the Anglosphere. As a consequence, observations on housing affordability are limited to perspectives from the English speaking global North and may not be generalizable to other locations. In addition, as our data is sourced from a single website, the comments reflect only the Reddit community perspective for each locale. Additionally, the number of annotators is limited as we were not able to secure an annotator local to each region included in the dataset. Additionally, our dataset is limited to only 500 comments for each test case. Finally, our experiment lacks a control group. Although we have attempted to mitigate this by reporting odd ratio, there may be differences between the 500 comments in the human-only annotation set and the 500 LLM pre-annotation set that are not isolated from other potentially influencing effects.

Ethical Considerations

First, all annotators involved in the project were given authorship on the paper as compensation for their contributions.

There are two main considerations that come into play regarding the data used for these experiments:

1) Although we have taken steps to anonymize comments by removing clear PII, it is incredibly challenging to control for all information that has the potential to identify individuals. Data de-identification is still an open problem, with multiple considerations that must be taken in mind. See Volodina et al. (2025) for a deeper discussion on this topic.

2) Even though all comments are freely viewable on public facing forums, the comment authors did not give informed consent for their data to be used for research purposes. Because of this, we opted not to re-publish raw text from the resulting dataset in a public forum.

Bibliographical References

Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, and Libby Hemphill. 2024. Prompt design matters for computational social science tasks but in unpredictable ways. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 122–145. Association for the Advancement of Artificial Intelligence (AAAI).

Valerio Basile et al. 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.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using LLMs for text annotation. *arXiv pre-print*, 2509.08825.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. 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.

Alexander S. Choi, Syeda Sabrina Akter, J. P. Singh, and Antonios Anastasopoulos. 2024. The LLM effect: Are humans truly using LLMs, or are they being influenced by them instead? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 22032–22054. Association for Computational Linguistics (ACL).

Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid intelligence. *Business & Information Systems Engineering*, 61(5):637–643.

Ola H Diserud and Frode Ødegaard. 2007. A multiple-site similarity measure. *Biology Letters*, 3:20–22.

Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International journal of human-computer studies*, 58(6):697–718.

Adrian Furnham and Hua Chu Boo. 2011. A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40:35–42.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50:1097–1179.

Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka Wei Lee, Simon Perrault, Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka Wei Lee, and Simon Perrault. 2023. CoAIcoder: Examining the effectiveness of AI-assisted human-to-human collaboration in qualitative analysis. *arXiv pre-print*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2024. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55.

Md Monjurul Karim, Sangeen Khan, Dong Hoang Van, Xinyue Liu, Chunhui Wang, and Qiang Qu. 2025. Transforming data annotation with AI agents: A review of architectures, reasoning, applications, and impact. *Future Internet*, 17.

Shahedul Huq Khandkar. 2009. Open coding. *University of Calgary, Technical report*, 23.

Manfred Klenner, Anne Göhring, Michael Amsler, Sarah Ebling, Don Tuggener, Manuela Hürlimann, and Martin Volk. 2020. Harmonization sometimes harms. In *SwissText/KONVENS*.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 10443–10461. Association for Computational Linguistics (ACL).

Anna Lindahl. 2025. LLMs as annotators of argumentation. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 242–252. Association for Computational Linguistics (ACL).

Marcus Ma, Georgios Chochlakis, Niyantha Maruthu Pandiyan, Jesse Thomason, and Shrikanth Narayanan. 2025. Large language models do multi-label classification differently. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, page 2472–2495. Association for Computational Linguistics.

A.E. Magurran. 2003. *Measuring Biological Diversity*. John Wiley & Sons.

OECD. 2021. OECD affordable housing database - indicator HC 1.5 overview of affordable housing indicators. Technical report, Organisation for Economic Co-operation and Development (OECD).

OECD. 2025a. OECD affordable housing database - indicator hc 1.2. house prices. Technical report, Organisation for Economic Co-operation and Development (OECD).

OECD. 2025b. OECD affordable housing database - indicator HC 1.4. subjective measures on housing. Technical report, Organisation for Economic Co-operation and Development (OECD).

Nicholas Pangakis and Samuel Wolken. 2024. Keeping humans in the loop: Human-centered automated annotation with generative AI. *Proceedings of the International AAAI Conference on Web and Social Media*, 19:1471–1492.

Karl Pearson. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

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.

Hope Schroeder, Deb Roy, and Jad Kabbara. 2025. Just put a human in the loop? Investigating LLM-assisted annotation for subjective tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25771–25795. Association for Computational Linguistics (ACL).

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185:1124–1131.

Elena Volodina, Simon Dobnik, Therese Lindström Tiedemann, Ricardo Muñoz Sánchez, Maria Irena Szawerna, Greta Lisa Södergård, Xuan-Son Vu, and T Attendee Lindström Tiedemann. 2025. Towards shared standards for psuedonymization of research data. In *Huminfra conference 2025*, pages 101–114. Huminfra.

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*, 37:14523–14530.

Jenny S Wang, Samar Haider, Amir Tohidi, Anushkaa Gupta, Yuxuan Zhang, Chris Callison-Burch, David Rothschild, and Duncan J Watts. 2025. Media bias detector: Designing and implementing a tool for real-time selection and framing bias analysis in news coverage. In *CHI Conference on Human Factors in Computing Systems*, volume 1, pages 1–7. Association for Computing Machinery.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205. Association for Computational Linguistics (ACL).

Lauren Wilcox. 2023. No Humans in the Loop: Killer Robots, Race, and AI. In Jude Browne, Stephen Cave, Eleanor Drage, and Kerry McInerney, editors, *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines*. Oxford University Press.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50:237–291.

Language Resource References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit dataset. *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*, pages 830–839.

A. Labels and their definitions

This appendix details the different labels we used for the different categories we considered.

Label	Definition
Housing to Income	The comment mentions factors such as housing price or rent price in comparison to income. Can include comments on the percentage/ratio of one's income spent on rent, mortgage, purchasing property, etc.
Quality	The comment mentions factors used to assess housing quality, such as number of bedrooms, location, bathrooms, pet friendliness, overcrowding, etc.
Availability	The comment mentions housing availability or lack of. Can include comments talking about lack of low-income designated housing, being unable to find a place to rent, etc.
Real Price	The comment talks about actual prices, fees, increases, or decreases in the price of property and rent. May be compared to other countries, cities, or other periods of time.

Table 7: Labels for the "factor" category.

Label	Definition
Government Policy (I)	The comment mentions or makes suggestions of government actions and policies that contribute to an improvement in housing affordability.
Building	The comment mentions that building more housing or increasing density contributes to improved housing affordability.
Bootstraps	The comment mentions that acquiring money, such as through working hard, saving, donation from family, etc., improves one's housing situation.
Group Action	The comment mentions that group action, such as unions or protesting, may contribute to improvement in housing affordability.
Relocation	The comment mentions relocating to another area as a method to secure housing in general or more affordable housing.
Money	The comment mentions that simply making more money or spending more money will fix the problem.

Table 8: Labels for the "aspect – improve" category.

Label	Definition
Foreign Investment	The comment mentions ‘foreign investment’ or ‘foreign buyers’ as a cause contributing to concerns about housing affordability.
Underbuilding	The comment mentions under-building of housing as a cause contributing to concerns about housing affordability. Either in terms of quantity alone, or quantity of quality units.
Government Policy (E)	The comment mentions or implies that the ‘government’ (local or national) has not done enough/should do more to improve the housing situation, or has directly contributed to housing affordability concerns through policy.
NIMBYism	The comment mentions that private owners/nimbys blocking development or policies that would improve housing affordability contribute to concerns about housing affordability.
The Rich	The comment mentions that the wealthy, landlords, or realtors contribute to concerns about housing affordability. This may be attributed to greed (buying many properties or charging excessive rent/fees), manipulation (influencing government policy), etc.
COVID	The comment mentions or implies that COVID-19 and lockdowns had an impact on housing affordability.
The Old	The comment suggests that older property owners/boomers/the elderly in general have an impact on housing affordability.
Wage Price Imbalance	The comment mentions that wages in general are not sufficient to buy a home. May mention wage stagnation or inflation.
Cost of Living	The comment suggests that rising cost of living contributes to struggles with housing and affordability. May mention rising prices, inflation, rising mortgages and insurance rates, etc.
AirBnB	The comment mentions that short term vacation rentals, such as AirBnB, contribute to challenges with housing.
Overcrowding	The comment mentions that high population density, overcrowding, or immigration contribute to difficulty in securing housing. Specifically mentions density as a reason why more houses are needed.
Demand	The comment mentions that high demand due to location desirability contributes to high housing/rental prices.

Table 9: Labels for the “aspect - exacerbate” category.

B. Label Counts

This appendix presents the label counts per annotator for both the human-only and the LLM pre-annotation experiments.

Label	Annotator A		Annotator B		Annotator C	
	D_h^1	D_h^2	D_h^1	D_h^2	D_h^1	D_h^2
Real Price	130	169	159	208	68	184
Quality	97	137	49	117	24	138
Availability	52	80	64	109	29	123
Housing to Income	43	50	66	85	25	73
Building	27	46	26	51	12	54
Bootstraps	39	49	15	39	64	40
Government Policy (I)	46	55	34	47	46	86
Group Action	4	1	1	1	2	1
Relocation	18	46	19	62	26	61
Wage Price Imbalance	24	34	44	67	14	68
Underbuilding	15	19	16	38	18	40
Government Policy (E)	39	52	33	56	47	71
Cost of Living	44	40	30	34	17	67
Foreign Investment	39	21	35	24	35	23
NIMBYism	23	15	20	17	30	26
Overcrowding	13	14	18	20	10	17
The Rich	61	61	52	61	46	65

Table 10: Frequency of labels per annotator in the human-only annotations (D_h^1) and the LLM-assisted annotations (D_h^2).

Label	LLM	
	D_l^1	D_l^2
Real Price	176	193
Quality	109	128
Availability	132	129
Housing to Income	81	78
Building	57	56
Bootstraps	50	36
Government Policy (I)	102	92
Group Action	7	1
Relocation	36	57
Wage Price Imbalance	70	72
Underbuilding	38	44
Government Policy (E)	70	73
Cost of Living	66	67
Foreign Investment	34	22
NIMBYism	33	24
Overcrowding	17	17
The Rich	75	62

Table 11: Frequency of labels between the first half of the dataset (D^1) and second half of the dataset (D^2), as annotated by the LLM.

An Overview of Current Practices and Recommendations for Working with Stereotypes in NLP

Alessandra Teresa Cignarella[♡] **and Matteo Pellegrini**[♠]
♡ Language and Translation Technology Team (LT3), Ghent University, Belgium
♠ Surrey Morphology Group (SMG), University of Surrey, United Kingdom
alessandrateresa.cignarella@ugent.be, matteo.pellegrini@surrey.ac.uk

Abstract

This article presents a discussion on the main challenges and considerations involved in addressing stereotypes within Natural Language Processing (NLP), and proposes a set of guidelines and recommendations for their treatment in research and resource development. On the one hand, the growing interest in fairness, bias mitigation, and inclusivity has led to an increasing number of studies and datasets dealing with stereotypes; on the other hand, their conceptualization and operationalization remain highly heterogeneous across works. The aim of this article is therefore twofold: (1) to provide a concise yet comprehensive overview of existing annotation schemes highlighting their key features and offering a comparative analysis and (2) to propose a set of tentative guidelines and recommendations to foster clarity when working with stereotypes in NLP. Furthermore, as a case study, we conduct an annotation exercise of a subset of texts from the QUEEREOTYPES dataset, containing stereotypes targeting LGBTQIA+ people, using all labels proposed in prior work to assess their clarity, overlap, and practical usefulness.

Keywords: stereotypes, annotation, agreement, pilot, Italian, LGBTQIA+

Trigger warning: *This paper contains examples of stereotypical and potentially triggering content.*

1. Introduction and Motivation

Stereotypes are exaggerated beliefs associated with a social category (Allport, 1954), they are pervasive social constructs that influence how individuals and groups are perceived, evaluated, and represented (Fiske, 1998). As well as prejudice and discrimination, they can shape social expectations and, in their most extreme forms, contribute to hatred or acts of violence. After having been extensively studied for decades in sociology and psychology, stereotypes have more recently become a central focus of research on fairness, inclusivity, and social bias in NLP, where understanding their linguistic manifestations is crucial for building more equitable and socially aware language technologies (Blodgett et al., 2021).

Yet, despite the growing number of datasets and models addressing stereotypical content (Nangia et al., 2020; Nadeem et al., 2021; Davani et al., 2023; Cignarella et al., 2024; Schmeisser-Nieto et al., 2024, among others), a fundamental challenge remains: the perception of stereotypes is inherently subjective, contextual, and heavily dependent on individual perspectives. What is considered a stereotype, who it targets, who it affects and whether it is harmful or not often depends on cultural backgrounds, lived experiences, linguistic communities, and ideological positions. Treating stereotypes as if they reflected a single, unified ground truth risks flattening the diversity in which they are experienced and interpreted. To adequately capture their complexity there is a growing need for detailed categories and explicitly intersectional approaches (Ma et al., 2023).

The emergence of data perspectivism offers a promising paradigm for confronting this challenge (Cabitza et al., 2023). Perspectivist approaches embrace annotator disagreement rather than eliminating it, and leverage human label variation as a source of information rather than noise (Uma et al., 2021; Leonardelli et al., 2025). Work at the intersection of perspectivism, participatory design, and fairness increasingly demonstrates the value of acknowledging user diversity in annotation and evaluation (Prabhakaran et al., 2021). Nonetheless, existing work on stereotypes in the field of NLP has not yet fully embraced systematic methodologies that explicitly account for a perspectivist approach (aside from some exceptions, i.e., Fraser et al. (2024); Lo et al. (2025)).

As noted by Röttger et al. (2022), "labelled data is the foundation of most natural language processing tasks. However, labelling data is difficult, and there often are diverse valid beliefs about what the correct data labels should be". We fully agree with their observation that "dataset creators should consider the role of annotator subjectivity in the annotation process and either explicitly encourage it or discourage it", and we draw direct inspiration from their approach.

In current NLP research on stereotypes, a wide range of annotation schemes has emerged, each

Proceedings of NLPerspectives @ LREC 2026, pages 112–123

grounded in different theoretical assumptions and targeting different aspects of the phenomenon (Cignarella et al., 2025). These schemes may vary substantially in scope: some focus on stereotypes directed at a single social group, while others consider multiple groups or even multiple dimensions at once, such as the intersection of gender and occupation. In certain cases, stereotypes are annotated alongside related phenomena like polarity, stance or hate speech, whereas other studies treat them in isolation. This heterogeneity makes it difficult to compare resources or understand how specific categories should be interpreted and used. There is therefore a need for greater clarity regarding the utility, granularity, and underlying assumptions of existing annotation categories.

In this article, we review the main annotation labels used so far, discuss their differences and commonalities, and offer commentary on how they can be meaningfully applied within the broader landscape of stereotyping research in NLP.

This article has three main contributions.

- **First.** We offer a critical overview of existing annotation schemes for stereotypes, outlining their defining characteristics and providing a comparative analysis that highlights key points of convergence and divergence across approaches.

- **Second.** To ground this discussion, we conduct a pilot annotation study in which we annotate texts ($N = 100$) from a dataset containing stereotypes targeting the LGBTQIA+ community using the labels and categories drawn from previous studies. The purpose of this annotation exercise is not to introduce a new resource, but rather to use it as a lens through which to examine the practical utility of different annotation categories: which labels are meaningful, which appear redundant or ambiguous, how certain phenomena should (or should not) be annotated, and which dimensions may prove the most helpful for future work in this area.

- **Third.** Finally, on the basis of the outcome of the pilot annotation and the discussion between annotators, we annotate a larger set of social media texts extracted from the same dataset ($N = 400$, see §3 for more details), with an annotation scheme that is a synthesis of what discussed in the previous point. It reuses some of the categories proposed in earlier studies and refines some new ones, complying with the recommendations and best practices identified in the previous stage.

This study is thus intended as an occasion for refining our recommendations, illustrating the challenges and opportunities that arise when operationalizing in real social media data.

2. Related Work

When considering stereotypes in NLP, and especially previous work devoted to data creation and annotation, it can be observed how previous related studies have evolved through three main waves, each characterized by a distinct methodology.

Old-School, the pioneers of sentiment and affect in NLP. Early work (2016–2019) primarily examined harmful language phenomena such as hate speech, offensiveness, and aggressive or stereotyped content. This period was marked by a strong emphasis on linguistically grounded annotation schemes and fine-grained distinctions between related-phenomena. A pioneering example is the Italian Twitter corpus by Sanguinetti et al. (2018), where multi-layer annotations (hate speech, aggressiveness, offensiveness irony) highlighted the pragmatic nuances shaping abusive language online. Parallel efforts on sentiment and polarity, such as SENTIPOLC (Barbieri et al., 2016), integrated sentiment analysis with irony detection, reflecting how pragmatic markers can shift meaning. Additionally, well-known shared tasks such as WASSA[1] advanced research on sentiment, emotion intensity and fine-grained affective states, reinforcing the attention to nuanced expressions beyond simple polarity labels. At the same time, stance detection was explored extensively by Mohammad et al. (2016a,b), who established frameworks that clearly separated stance from sentiment.

Millenials, the psychology-informed and perspective-based generation. A second, more recent wave (2019-2023/2024) began grounding stereotype-related annotation in cognitive and social psychological theory (Fraser et al., 2021). This includes work targeting social groups more explicitly (Nozza et al., 2021), and combining stereotype annotations with hate speech, aggressiveness, offensiveness, irony, sarcasm and stance (Cignarella et al., 2024). Additional contributions in this line include analyses of target-specific slurs (Draetta et al., 2024) and the annotation of forms of discredit to capture subtle mechanisms of stereotyping and prejudice beyond overt hate speech (Bosco et al., 2023; Schmeisser-Nieto et al., 2024).

Gen Z, the free style, identity-aware and perspective-aware approaches. The third, and most recent, line of research (2023/2024-today) explores cognitively aligned representations of stereotype expression, for instance framing stereotypes

[1] https://workshop-wassa.github.io/

as generics by linking a social group to a quality using the reasoning scheme GROUP + relation + QUALITY (Mun et al., 2023) or including free-text descriptions in the form of Subject + Verb + Object or Subject + Noun Phrase patterns (Lo et al., 2025).

The literature on bias and stereotypes in NLP is quite extensive, and the brief overview we proposed is focusing specifically on works that introduce new annotated datasets with clearly defined stereotype-related dimensions. For a broader and more comprehensive perspective on stereotypes and bias, we refer the reader to recent surveys on the topic (Cignarella et al., 2025; Bartl et al., 2025).

3. A Pilot Annotation Study

To better understand how existing stereotype annotation schemes function in practice, we conducted a small-scale pilot study using a subset of texts. This pilot serves as the empirical backbone of our analysis: rather than introducing a new resource, our objective is to test and compare the categories proposed in prior work, assessing their usefulness and limitations when applied to real data.

Specifically, we conduct a focused re-annotation of a subset of QUEEREOTYPES (Cignarella et al., 2024), an Italian social media dataset of stereotypes toward LGBTQIA+ people. It consists of two distinct components: approximately half of the data comes from X and the other half from Facebook. The first portion contains individual tweets annotated for stance and stereotypes, while the latter is organized into status–comment pairs and is annotated for hate, aggressiveness, offensiveness, stereotypes, and irony following the scheme proposed by Sanguinetti et al. (2018). For the purposes of our study, we first harmonized these two sections by completing the annotations for categories that were present in one portion of the dataset but missing in the other.

Then, we only selected the texts containing a stereotype and we enriched them with additional labels drawn from previous research, including: *sentiment and polarity* (Pang and Lee, 2008; Barbieri et al., 2016), *stance* (Mohammad et al., 2016a; Küçük and Can, 2020), *target* (Basile et al., 2019; Nozza et al., 2023), *reported speech* (Schmeisser-Nieto et al., 2022), *slur reclamation* (Kurrek et al., 2020; Draetta et al., 2024), *forms of discredit* (Bosco et al., 2023; Bourgeade et al., 2023), and a *free-text field* (Sap et al., 2020; Lo et al., 2025).

In what follows, we provide details on each of the categories and associated values that we considered when harmonizing and extending the annotations across the two portions of the dataset.

Sentiment/Polarity. A three-way polarity label capturing the affective orientation expressed in the text: *positive*, *neutral*, or *negative*.

Hate Speech. A binary variable (*yes/no*) indicating whether the text contains hateful, hostile, or dehumanizing language directed at an individual or group.

Aggressiveness. A three-level scale assessing the intensity of aggressive expressions: *absent*, *weak*, or *strong*.

Offensiveness. A parallel three-level scale (*absent/weak/strong*) evaluating the degree of insulting, derogatory, or socially inappropriate language.

Irony. A binary label (*yes/no*) indicating the presence of ironic, sarcastic, or otherwise non-literal humorous language.

Stereotype. A binary variable (*yes/no*) marking whether the text conveys a stereotype, i.e., a generalized and often biased attribution of traits or behaviors to a social group.

Stance. A three-way classification capturing the author's position toward the target: *favour*, *neutral*, or *against*.

Target. A coarse-grained label specifying whether the text concerns a *queer*, *non-queer*, or *other* (unspecified/alternative) social group.

Target Specific. An optional free-text field allowing annotators to specify the target at a finer level of granularity (e.g., "gay men", "trans women", "allies").

Reported Speech. A binary label (*yes/no*) marking the presence of quoted or otherwise explicitly reported discourse.

Slur Reclamation. A category identifying the contextual function of slur-related expressions: *slur*, *reclaimed*, or *n/a* (Draetta et al., 2024).

Forms of Discredit. A six-way classification describing how different types of stereotypes undermine or delegitimize the target: *benevolence*, *competence*, *dominance up*, *dominance down*, *affective competence*, and *physical* (Bosco et al., 2023).

Free-text Field. An open field used to encode the stereotype in a schematic proposition, typically following structures such as $S + V + O$ (subject–verb–object) or $S + NP$ (subject–noun phrase) (Lo et al., 2025).

Notes and Comments. A free-text field for annotators to record uncertainties, contextual information, or justifications for labeling decisions.

The two authors of this paper have performed a full and independent annotation of 100 texts (50 tweets and 50 post-comment pairs). Consequently they met to discuss the annotation choices, the disagreements and the difficult cases. We do not report the values of inter-annotator agreement (IAA) for the motivations discussed in in the Limitations section.

The goal of this targeted exercise was not to revise or replace the existing resource, but to explore the complexity of the task and assess clearer

insights into how stereotypes are expressed. By annotating only the stereotype-positive instances with a more detailed taxonomy and a set of labels side-by-side, we were able to interpret the usefulness, clarity, and limitations of each proposed category and to identify which ones meaningfully contribute to a better understanding of stereotyping in NLP.

4. Discussion of Results

From our pilot annotation study, two broad categories of challenges emerge: first, those concerning aspects related to *phenomena, labels and categories* themselves (§ 4.1); and second, those related to *annotation practices and layout*, including data presentation, including methodological choices and features of the platform used (§ 4.2).

4.1. Phenomena, Labels and Categories

4.1.1. Sentiment/Polarity

Challenge: Sentiment annotation proved to be especially problematic. A first issue is conceptual: it is unclear whether annotators should assign sentiment based on an overall intuitive impression or whether they should instead compute something closer to an *algebraic sum* of the positive and negative valence of individual words. These two approaches can produce markedly different labels: if sentiment is treated as a word-level aggregation problem, then annotation becomes almost redundant, since automatic methods (including lexicon-based ones) can compute polarity and even highlight which words contribute to the final score. However, this raises further concerns, as lexicon-based sentiment detection might inherit the biases of the lexicons. The presence of ambivalent sentiment (positive, negative, neutral, or even mixed) adds another layer of complexity, challenging the usefulness of including sentiment as a stand-alone annotation category.

Our proposal: Use sentiment annotation only when affective meaning is explicit and directly contributes to the interpretation of the stereotype. In other cases, rely on automatic polarity detection methods as a first pass and let annotators perform only a subsequent light verification step rather than full manual annotation, reducing redundant workload while maintaining quality.

4.1.2. Hate Speech, Aggressiveness and Offensiveness

Challenge: Hate speech, aggressiveness, and offensiveness are intrinsically subjective phenomena. Even with improved guidelines and refined annotation schemes, annotators will inevitably bring their own perspective, background knowledge, and sensitivity to the task. Disagreement can be reduced, but probably never fully avoided, because judgments depend on individual perceptions of what constitutes harm or denigration.

Our proposal: Dataset creators should explicitly state whether the goal is to produce a consensus-based gold standard or a subjectivity-aware resource that captures multiple interpretations. Following Röttger et al. (2022), we encourage making annotator subjectivity an intentional design choice: either constrain it (in the case of gold-standard labels) or embrace it (when modelling perspectives).

4.1.3. Irony and Sarcasm

Challenge: Irony and sarcasm typically flip the literal meaning of statements: an utterance may reproduce a stereotype only to mock or reject it. This raises the question of whether the text should be considered as containing a stereotype or not.

Example. *"Why don't you ask your beloved African illegal immigrants what they think of Gay Pride? They are tolerant of homosexuals in their progressive countries; as we know, they adore them"*

This example relies on two classic ironic devices: a rhetorical question in the first part and a false assertion in the second part (Karoui et al., 2017). The ironic meaning becomes interpretable only through world knowledge and pragmatic inference: readers must recognize that the sentence is implausible and therefore intended as its opposite. This makes the utterance a case in which the literal form and the intended meaning diverge, illustrating why irony complicates stereotype annotation and requires explicit annotation of non-literal intent.

Our proposal: Annotate the presence of the stereotypical content and add a separate flag for the presence of irony or sarcasm. This keeps literal content and intended meaning distinct and avoids mislabeling anti-stereotypical discourse.

4.1.4. Explicit and Implicit Stereotypes

Challenge: Identifying whether a stereotype is present in a text is far from straightforward. Firstly, the field lacks a clear, operationalized definition of what constitutes a stereotype in NLP (Devinney, 2025). As a consequence, annotators may rely on personal intuition rather than shared criteria. Secondly, as noted by Schmeisser-Nieto et al. (2022), many stereotypical meanings are not stated overtly but must be inferred from background knowledge, pragmatic cues, or cultural assumptions.

Example. *"@user MY AUNT MARIA SAYS YOU'RE ALSO HOMOSEXUAL, I DON'T BELIEVE IT, YOU'RE SUCH A HANDSOME MAN, OR NOT?"*

This text suggests that being homosexual is incompatible with being handsome. The stereotype can thus be detected only via an inferential step.

OUR PROPOSAL: We propose to clearly distinguish between *explicit* and *implicit* stereotypes. Explicit stereotypes are those directly stated in the text (e.g., "LGBT people are not pure"), for which higher agreement is expected. Implicit stereotypes, by contrast, rely on presuppositions, implicatures or culturally-shared associations. Because different readers may draw different inferences, especially for implicit cases it is crucial to collect multiple voices rather than collapsing perspectives into a single viewpoint.

4.1.5. Identification of Specific Targets

CHALLENGE: Annotating the target of a stereotype is not always straightforward. While a coarse-grained label such as *queer*, *non-queer* or *other* is useful for ensuring consistency. However, in some utterances, the target of the stereotype is not overtly specified, but there is only a vague reference, e.g., to a generic "they". In other cases, the utterance concerns a specific subgroup whose identity cannot be captured adequately by a single coarse category.

OUR PROPOSAL: We adopt a two-level annotation strategy. First, annotators assign a coarse-grained **Target** label (*queer*, *non-queer*, *other*). Second, when the target can be identified more precisely, annotators may use the optional **Target (Specific)** free-text field to capture finer-grained subgroups (e.g., "gay men", "trans women", "allies"). This approach allows us to maintain a unified structure while preserving valuable detail when available.

4.1.6. Direct and Indirect Target

CHALLENGE: Stereotypical utterances typically have the underlying structure "Target group X has characteristic Y". In this paper, we focus on stereotypes involving queer individuals, who may appear either as the *direct* or the *indirect* target of the stereotype. We therefore distinguish between: (i) cases in which queer individuals are the direct target X of the stereotype, corresponding to statements of the form "All queer individuals have characteristic Y", and (iia-b) cases in which another individual or group is the direct target X, and the stereotypical association involves queer identities or queer-related evaluations only indirectly.

The latter category includes two common patterns: (iia) utterances in which a non-queer target is stereotypically associated with queerness as characteristic Y, and (iib) utterances in which the target group is described negatively because they support queer individuals.

Examples.

(i) Nowadays we find gays and transsexuals everywhere because they are included in all the TV programs.

(iia) @user If that's really the case, there's also the aggravating circumstance of homophobia... considering the not exactly masculine reaction of the violent robber

(iib) They must be fake priests, surely left-leaning, probably atheists too. [Referring to priests who take part in marches against homophobia]

OUR PROPOSAL: All of the above configurations should be treated as relevant for the study of stereotypes involving the LGBTQIA+ community. However, we argue that it is important to encode the distinction between them in the annotation scheme, as this enables more fine-grained analyses of how queer identities are targeted in discourse. To this end, we introduce a dedicated annotation category, **directness**, with three possible values:

- **(i) DIRECT**: queer individuals are the direct target of the stereotype.
- **(iia) INDIRECT-CHARACTERISTIC**: another individual or group is the target, but they are stereotypically associated with queerness.
- **(iib) INDIRECT-ALLIES**: another individual or group is targeted negatively because they support queer individuals.

We could also distinguish cases where queer individuals appear as initiators or as recipients of the action, an idea reminiscent of semantic-role labeling (agent/patient). This point connects naturally to the free-text rewriting of stereotypes (Lo et al., 2025) and will be elaborated further in Section 4.1.11.

4.1.7. Prejudice and Discredit

CHALLENGE: In some cases, even when a stereotype is clearly present in the text, it does not necessarily express discredit toward queer individuals. This situation may arise for several reasons: (a) the stereotype attributes to queer individuals a characteristic that is not negative *per se*; (b) the stereotype is generic or it is not-clear which feature is being attributed to queer individuals; or (c) queer individuals are not the direct target of the stereotype but appear only indirectly (see § 4.1.6).

Examples.

(a) "Eleonora, men are all the same" as my mother says. I think I might want to become a lesbian at this point.

(b) @user, what a terrible combination you are... Democratic Party supporter, lawyer, AC Milan fan... You're just missing being a lesbian and you'd be all set...

(c) But those people in show business... singers, actors, fashion designers... are they all gays or lesbians? What kind of atmosphere is there over there?

OUR PROPOSAL: We recommend distinguishing between *derogatory* and *non-derogatory* stereotypes.[2] Based on this distinction, we propose annotating discredit only when the stereotype is classified as **derogatory** (see § 4.1.8). This makes

[2]Here we intentionally avoid the terms "positive" and "negative" since even non-derogatory stereotypes, such

the discredit label an optional, downstream category applied in a cascade fashion: first determine whether a stereotype is present, then whether it is derogatory, and only if it is annotate discredit.

4.1.8. Forms of Discredit

CHALLENGE: Existing discredit categories were originally designed for other target groups such as migrants (Bosco et al., 2023; Bourgeade et al., 2023) and may not map neatly to the LGBTQIA+ community as a target.

Furthermore, the presence of discredit presupposes that the underlying stereotype is derogatory (see § 4.1.7): if the stereotype is non-derogatory, then discredit is absent and thus conceptually inappropriate.

OUR PROPOSAL: We recommend revisiting discredit categories and adapting them specifically for queer-related contexts rather than importing labels designed for other targets. This involves (i) ensuring that discredit is annotated only when the stereotype is classified as *derogatory*; (ii) clarifying the terminology, avoiding fuzzy or overlapping labels and re-evaluating whether certain distinctions are meaningful or empirically grounded for LGBTQIA+ stereotypes.

Example.
A stereotype category (PHYSICAL) developed for migrants typically refers to dirtiness, diseases, or the idea that migrants "carry illnesses" (Bosco et al., 2023; Schmeisser-Nieto et al., 2024). For queer-related stereotypes, however, this dimension needs to be rethought. In queer contexts, physicality often appears through markers such as clothing style, haircut, grooming, or body modifications treated as stereotypical "signals" of queerness. Some cases still invoke health-related stigma (e.g., *"National data showed infections among homosexual men... why not ask why?"*), while others rely on appearance-based cues (e.g., *"Typical lesbian haircut and short nails"*). This suggests expanding the physical category to include both health-based stigmas and appearance-related markers commonly used to stereotype queer individuals.

4.1.9. Stance, Reported Speech and Counterspeech

CHALLENGE: Some posts or tweets reproduce a stereotype not to endorse it, but to comment on it or explicitly reject it. In such cases, the stereotypical content appears only within reported speech, while the author's own stance is oppositional. This creates a misalignment between the *literal content* (which may contain harmful or stereotypical expressions) and the *author's intention* (which may be supportive of queer individuals). These cases also raise questions about the annotation of stance: if someone cites a derogatory stereotype against the LGBTQIA+ community to reject it, the stance of the stereotype towards the community will be trivially negative, but the stance of the author will be positive. As noted by Schmeisser-Nieto et al. (2022), reported speech, counterspeech, and euphemisms are rare and therefore often grouped together, yet they are pragmatically distinct and particularly relevant for tasks such as automatic moderation, where it is crucial to differentiate authors who propagate stereotypes from those who criticize them.

Example.
"I think you are a lesbian." Why? Because I wear jeans and T-shirts and do not use make-up and do not care about womanly things? WAKE UP, I AM STRAIGHT AND LESBIANS CAN BE MODELS WEARING GUCCI FROM HEAD TO TOE. Stop stereotyping.

OUR PROPOSAL: We propose treating reported stereotypes as regular stereotype occurrences but marking them explicitly with a dedicated **"Reported Speech"** flag. Annotators should (i) identify the presence of a stereotype in the reported content, and (ii) annotate the author's stance and not the stance conveyed by the stereotypical utterance. This allows for cases where a neutral or negative sentiment from a reported speech co-occurs with an author's stance supportive of queer individuals.

4.1.10. Slur Reclamation

CHALLENGE: Slurs can appear either as clear insults or as reclaimed, in-group expressions (Ferrando et al., 2026). The same term may therefore be derogatory in one context and identity-affirming or playful in another. Because reclaiming depends on speaker identity, audience, and context, treating all occurrences uniformly as harmful would conflate hostile uses with in-group language practices.

OUR PROPOSAL: We adopt a simple three-way label for the slur category. We use *slur* only when the term is clearly derogatory; *reclaimed* for in-group uses, and *none* where there is no slur (although we may still see offensiveness or aggressiveness which could be annotated as per § 4.1.2).

4.1.11. Free-Text

CHALLENGE: Free-text fields enable annotators to capture stereotype constructions beyond predefined labels, including generic GROUP + RELATION + QUALITY statements (Mun et al., 2023) or syntactically grounded S + V + O sentences (Lo et al., 2025). In addition, free-text annotation can be used to record entailments and implicit meanings (Sap et al., 2020). However, this flexibility results in highly variable outputs: annotators differ in wording, level of detail and linguistic focus. Such variability may complicate aggregation and can blur the

as those assigning supposedly positive traits, can reproduce harmful biases and reinforce structural inequalities (e.g., "women are naturally better at caregiving").

line between what is stated in the text and what is inferred by the annotator.

OUR PROPOSAL: We recommend emphasising that free text annotation are not to be intended as completely unconstrained and descriptions, but have to conform with the structure of the S + V + O or GROUP + RELATION + QUALITY prompt. This avoids obtaining noisy results, that might be too difficult to aggregate or compare and facilitates automatic extraction (Felkner et al., 2023; Lo et al., 2025).

4.2. Annotation Practices and Layout

4.2.1. Annotation Platform

CHALLENGE: Annotation platforms differ widely in functionality and usability: simple layouts, such as drop-down menus on spreadsheets, offer flexibility but lack interface support; tools like *LabelStudio* provide richer workflows but may require technical setup; crowdsourcing platforms (e.g., *Prolific*) introduce additional variability in annotator background and quality control; proprietary solutions can limit transparency and reproducibility.

OUR PROPOSAL: We recommend selecting the right medium based on task complexity and the need for controlled guidance. For tasks involving nuanced or implicit stereotypes, interfaces that support clear instructions, validation rules, and structured free-text fields are preferable. Regardless of the platform, we encourage documenting interface design choices in a Data Statement (Bender and Friedman, 2018; McMillan-Major et al., 2024), exporting data in interoperable formats and conducting small pilot rounds to ensure that the tool supports reliable and consistent annotation.

4.2.2. Annotation Granularity

CHALLENGE: Stereotypes may appear at different textual levels: entire documents, social media threads, single posts or paragraphs, or specific spans of text. If the intended unit of annotation is not clearly defined, annotators may rely on different amounts of context, leading to inconsistent judgments and reduced reliability.

Currently, the annotation is performed for each facebook post or tweet, but in some cases it can be useful or necessary to identify shorter spans of texts to be annotated. In general, this can be useful when the text of the post/tweet is very long, and in particular if more than one stereotype is present, possibly of different types. In the following example, stereotypes against both women and gay people can be identified in different parts of the text.

Example.

If you think that those who are contesting the manifesto are all women, it gives me shivers. They are those who want abortion to murder defenseless beings by right, but ask for the adoptability of other people's children and the uterus of others for rent to give children to GAY people. What world are we going to where human life can be suppressed by law and it is believed to be a right to be able to suppress it.. Women....shame on you and you also have the pretense of calling yourselves mothers??

OUR PROPOSAL: We propose explicitly stating the annotation level and providing minimal examples to illustrate it. If annotation below the level of the post/tweet is deemed necessary, it could be achieved in different ways, for instance by performing an *a priori* segmentation of longer texts into phrases/sentences/conversation turns (depending on the task to be accomplished by the researcher) or keeping the whole textual unity together despite its length but identifying and annotating shorter spans of text, using tools such as *LabelStudio* or similar. The choice between these two options depends on the specific needs of different studies.

4.2.3. Order of Annotations

CHALLENGE: The order in which different annotations is performed matters (Beck et al., 2024). In some cases, this is trivial (e.g., the category "target specific" cannot but be annotated after the category "target"), but in other cases the reason can be subtler (e.g., the stance of a text should be more precisely interpreted as the stance of the writer about the target of the stereotype, so it should be annotated after having identified the target of a stereotype).

Another example is the free-text description of the stereotype: if it is performed after other annotations, it can be influenced by the labels proposed as options of the other annotations. For instance, the values of the category pertaining to the type of discredit can be suggestive of the content of the stereotype, and the identification of the target might determine the template structure.

OUR PROPOSAL: The order in which annotations are presented should be considered carefully and instructions should be given to annotators on the order in which they should be performed. For instance, attention should be paid to whether the annotation is performed "horizontally" (i.e., all categories are annotated for a single text before going on to the next text) or "vertically" (i.e., all texts are annotated for a given category before going on to the next category). If there are logical interdependencies like the ones we just discussed, the annotation needs to be performed horizontally, but other categories can be more efficiently and consistently annotated vertically.

4.2.4. Availability of Context

CHALLENGE: In some cases, depending on how data was retrieved, contextual information might be missing, so that it is impossible to provide a fine-grained annotation on some aspects, or at least

it is necessary for annotators to draw non-trivial inferences. For instance, the Facebook data of our sample consists of a set of texts, each consisting of status-comment pair. However, in some cases the text of the original status make crucial reference to a picture, or to a web page, that were not available at the time of re-annotation. Consequently, sometimes non-trivial inferences are needed to understand the nature and details of the stereotype.

In the example below, commenting on a picture referring to male same-sex parenting, regarding comment (i) it can be inferred that the stereotype has something to do with gay males being unfit for parenthood, but for comment (ii) it is difficult to provide a fine grained annotation of all the categories involved without further information on the context.

Examples.

STATUS: *SOMETIMES A PHOTOGRAPH IS WORTH A THOUSAND WORDS*

COMMENT: *(i) What violence to this poor little one. He wants mommy and he wants to suck milk.*

COMMENT: *(ii) Poor baby*

The most common scenario when dealing with social media data is the presence URLs that may point to images, videos, or external content crucial for interpreting the message. Allowing annotators to open links introduces multimodal information that may be necessary for understanding stereotypes, but it also creates inconsistencies: different annotators may access different versions of the linked content, encounter unavailable pages, or rely on information that is not preserved in the dataset. This raises concerns about reproducibility and comparability, both among annotators and between human annotations and downstream systems, which may not have access to the same external resources.

OUR PROPOSAL: We recommend defining a clear policy on how to handle external context, aligned with the dataset's goals. If external content is essential, linked material should be archived or embedded (e.g., screenshots, textual extracts) and made available also offline, so that all annotators and systems access the same information. If multimodal access is not feasible or not in scope, annotators should be instructed *not* to open external context, and guidelines should clarify that the annotation must rely solely on the text provided.

5. Open Problems and Discussion

We highlight a set of open issues that, rather than calling for immediate solutions, we offer as prompts for discussion at the workshop.

- **New layers on old data.** Adding new annotation layers to existing datasets creates dependencies between past and present labels. Prior annotations can implicitly bias new judgments, raising concerns about reinterpretation of legacy data and the long-term consistency of datasets that accumulate layers over time.
- **Inter-annotator subjectivity across dependent layers.** When new layers depend on earlier interpretive decisions, disagreements become harder to resolve. Annotators may not perceive a stereotype where a previous annotator did, yet the earlier label remains part of the data. These inconsistencies highlight the fragility of sequential and interdependent annotations.
- **Annotator sensitivity and topic familiarity.** Annotators vary in how attuned they are to sensitive or domain-specific content. Limited familiarity, personal distance from the topic, or fatigue from prolonged exposure can all affect perception and judgment (Beck et al., 2024).
- **Annotator self-consistency.** Even the same annotator may label the same content differently when revisiting it later. Such intra-annotator variation raises questions about the stability of stereotype-related judgments and how to interpret labels that reflect inherently fluid perceptions.

We present these issues not as problems to be resolved here, but as starting points for discussion on how to navigate the complexities of stereotype-related annotation.

6. Conclusion

In this paper, we summarized current practices, challenges, and emerging considerations for annotating stereotypes in NLP. This empirical survey of existing approaches reveals that stereotype annotation remains a complex task shaped not only by the categories themselves, but also by the nature of the data, the design of annotation schemes, and the practical conditions under which annotators work. Choices regarding granularity, annotation layout, interface design, and the introduction of new layers all have effects on annotator reasoning, consistency and interpretation. Moreover, the inherently subjective and context-dependent nature of stereotypes introduces variability that cannot be fully eliminated, but can be better understood and anticipated.

Rather than prescribing definitive solutions, our aim has been to highlight areas where further reflection and discussion are needed. As research on stereotypes continues to evolve, so too must our annotation frameworks. We hope that the considerations raised here encourage more transparent, context-aware, and reflexive annotation practices and that they support the development of

datasets that more accurately capture the nuanced ways in which stereotypes appear in language.

Data Availability

Due to licensing and usage restrictions, the original QUEEREOTYPES dataset can be released only privately, upon request. Interested parties will be required to reach out to the first author, complete an agreement form, outlining the specifics of their research in order to obtain the password that protects the files. It is essential for them to ensure compliance with GDPR regulations and other policies from both X and Facebook.

Limitations

Our work presents some limitations. First, the analysis is restricted to a single target category and a single language, which limits the generalisability of our reflections to multilingual or cross-cultural contexts. Second, we build on an existing resource, and we do not question the reliability of inherited labels, nor fully control for how previous annotations may influence new layers.

Regarding inter-annotator agreement (IAA), although we computed it, we do not report the results because they are not informative for our setting: (i) several categories were inherited or only partially re-annotated (*Hate Speech*, *Aggressiveness*, *Offensiveness*, *Irony*, *Stance*, with *Stereotype* trivially always $= 1$); (ii) some labels are almost deterministic from the text or scheme (*Target*, *Reported speech*, *Slurs Reappropriation*); (iii) *Forms of discredit* involves overlapping classes with no straightforward agreement metric; and (iv) free-text fields (*Target Specific*, *S+V+O*/*S+NP*) are not suited to standard IAA measures.

Finally, only two annotators with similar backgrounds contributed to the new layers, which limits the diversity of perspectives and increases the risk of subjective bias. For these reasons, we avoid drawing overly definitive conclusions and view our findings as a starting point for further discussion.

Acknowledgements

The work of A.T. Cignarella is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Actions, Grant Agreement No. 101146287. Views and opinions expressed are, however, those of the author only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA).

Bibliographical References

G. W. Allport. 1954. *The nature of prejudice*, addison-wesley edition. Addison-Wesley.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the EVALITA 2016 SENTIment POLarity Classification Task. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5-7, 2016*, volume 1749 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Marion Bartl, Abhishek Mandal, Susan Leavy, and Suzanne Little. 2025. Gender bias in natural language processing and computer vision: A comparative survey. *ACM Computing Surveys*, 57(6):1–36.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 54–63. Association for Computational Linguistics.

Jacob Beck, Stephanie Eckman, Bolei Ma, Rob Chew, and Frauke Kreuter. 2024. Order Effects in Annotation Tasks: Further Evidence of Annotation Sensitivity. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 81–86. Association for Computational Linguistics.

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.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. 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 1004–1015.

Cristina Bosco, Viviana Patti, Simona Frenda, Alessandra Teresa Cignarella, Marinella Paciello, and Francesca D'Errico. 2023. Detecting racial stereotypes: An Italian social media corpus where psychology meets NLP. *Information Processing & Management*, 60(1):103118.

Tom Bourgeade, Alessandra Teresa Cignarella, Simona Frenda, Mario Laurent, Wolfgang Schmeisser-Nieto, Farah Benamara, Cristina Bosco, Véronique Moriceau, Viviana Patti, and Mariona Taulé. 2023. A Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 686–696. Association for Computational Linguistics.

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.

Alessandra Teresa Cignarella, Anastasia Giachanou, and Els Lefever. 2025. A survey on stereotype detection in natural language processing. *ACM Computing Surveys*, 58(5).

Alessandra Teresa Cignarella, Manuela Sanguinetti, Simona Frenda, Andrea Marra, Cristina Bosco, and Valerio Basile. 2024. QUEEREOTYPES: A Multi-Source Italian Corpus of Stereotypes towards LGBTQIA+ Community Members. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13429–13441, Torino, Italia. ELRA and ICCL.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Hannah Devinney. 2025. Power(ful) Associations: Rethinking "Stereotype" for NLP. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 52–58. Association for Computational Linguistics.

Lia Draetta, Chiara Ferrando, Marco Cuccarini, Liam James, and Viviana Patti. 2024. ReCLAIM Project: Exploring Italian Slurs Reappropriation with Large Language Models. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, December 4-6, 2024*, volume 3878 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140.

Chiara Ferrando, Lia Draetta, Marco Madeddu, Mae Sosto, Viviana Patti, Paolo Rosso, Cristina Bosco, Jacinto Mata, and Estrella Gualda. 2026. MultiPRIDE at EVALITA 2026: Overview of the Multilingual Automatic Detection of Slur Reclamation in the LGBTQ+ Context Task. In *Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026), Bari, Italy, February 26th-27th, 2026*. CEUR-WS.org.

Susan T Fiske. 1998. Stereotyping, prejudice, and discrimination. In *The handbook of social psychology*, pages 357–411. McGraw-Hill.

Kathleen C Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. 2024. How Does Stereotype Content Differ across Data Sources? In *Proceedings of the 13th joint conference on lexical and computational semantics (* SEM 2024)*, pages 18–34.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. 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 600–616, Online. Association for Computational Linguistics.

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.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for

Online Slur Usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 138–149, Online. Association for Computational Linguistics.

Elisa Leonardelli, Silvia Casola, Siyao Peng, Giulia Rizzi, Valerio Basile, Elisabetta Fersini, Diego Frassinelli, Hyewon Jang, Maja Pavlovic, Barbara Plank, and Massimo Poesio. 2025. LeWiDi-2025 at NLPerspectives: Third edition of the learning with disagreements shared task. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 182–195, Suzhou, China. Association for Computational Linguistics.

Soda Marem Lo, Marco Antonio Stranisci, Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Elisabetta Ježek, and Viviana Patti. 2025. Subjectivity in stereotypes against migrants in italian: An experimental annotation procedure. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 603–612. CEUR Workshop Proceedings.

Weicheng Ma, Brian Chiang, Tong Wu, Lili Wang, and Soroush Vosoughi. 2023. Intersectional stereotypes in large language models: Dataset and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8589–8597, Singapore. Association for Computational Linguistics.

Angelina McMillan-Major, Emily M. Bender, and Batya Friedman. 2024. Data Statements: From Technical Concept to Community Practice. *ACM Journal on Responsible Computing*, 1(1).

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. A Dataset for Detecting Stance in Tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 31–41. The Association for Computer Linguistics.

Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond Denouncing Hate: Strategies for Countering Implied Biases and Stereotypes in Language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9759–9777. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. Honest: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406.

Debora Nozza, Alessandra Teresa Cignarella, Greta Damo, Tommaso Caselli, and Viviana Patti. 2023. HODI at EVALITA 2023: Overview of the first Shared Task on Homotransphobia Detection in Italian. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023*, volume 3473 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

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.

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.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

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. Association for Computational Linguistics.

Wolfgang Schmeisser-Nieto, Montserrat Nofre, and Mariona Taulé. 2022. Criteria for the annotation of implicit stereotypes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 753–762.

Wolfgang S Schmeisser-Nieto, Alessandra Teresa Cignarella, Tom Bourgeade, Simona Frenda, Alejandro Ariza-Casabona, Mario Laurent, Paolo Giovanni Cicirelli, Andrea Marra, Giuseppe Corbelli, Farah Benamara, et al. 2024. Stereohoax: a multilingual corpus of racial hoaxes and social media reactions annotated for stereotypes. *Language Resources and Evaluation*, pages 1–39.

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.

Modeling Perspectives in NLP: Parameter-Efficient Perspective Conditioning for Span Extraction and Summarization

Harikrishnan Gurushankar Saisudha, Sabine Bergler
Concordia University,
Montreal, Quebec, Canada,
h_gurush@live.concordia.ca, sabine.bergler@concordia.ca

Abstract

Understanding text through multiple perspectives is essential in healthcare community question answering, where answers frequently contain heterogeneous viewpoints, including experiences, suggestions, causes, follow-up questions, and informational claims. We present a unified perspective-conditioned framework for both span identification and perspective-aware summarization on the PerAnsSumm dataset. We approach explicit perspective samples in transformer models using two parameter-efficient mechanisms: prefix-conditioned representations and perspective-aware attention layers. First, we use multi-label perspective classification to identify relevant viewpoints, which serve as conditioning signals for downstream tasks. Span identification for perspective-specific extraction is modeled as a conditioned binary sequence labeling problem. Summarization, finally, is guided by perspective-enriched encoder representations. Experiments demonstrate that explicit perspective conditioning substantially improves span detection performance while achieving competitive summarization quality. Notably, perspective-aware attention achieves strong results using only a small fraction of the trainable parameters required by full fine-tuning. Our findings highlight the importance of structured viewpoint modeling and show that explicit perspective control enables efficient and interpretable multi-perspective text understanding.

Keywords: attention, summarization, span identification, BIO tags, perspective conditioning, multi-label classification, parameter-efficient approach, prefix tuning

1. Introduction

Text carries multiple viewpoints, such as personal experiences or factual claims (Cabitza et al., 2023). Understanding text through multiple viewpoints or perspectives is an essential yet underexplored challenge in NLP. Modeling these perspectives explicitly is crucial for tasks that require nuanced comprehension and generation (Frenda et al., 2025).

We address this challenge in the context of the PerAnsSumm shared task (Agarwal et al., 2025), which operates on the PUMA dataset (Naik et al., 2024), a corpus of healthcare community question-answering (CQA) threads annotated with five perspective types: cause, suggestion, experience, question, and information. This mirrors recent interest in Data Perspectivism (Cabitza et al., 2023), which explores here giving different answers to the same question, depending on a set of different perspectives that are defined by samples in the training data. Perspectives are pertinent in healthcare CQA, where subjects answer from fundamentally different epistemic positions — medical professionals, patients, and caregivers each bring distinct knowledge and lived experience to the same question. The five perspective categories in PUMA are not arbitrary topical bins but operationalization of different human standpoints, making perspective-aware summarization and span identification an instance of the perspectivist paradigm. The task has two subtasks: *Task A: perspective span identification*, where perspective-indicating text spans must be detected and labeled, and *Task B: perspective summarization*, where summaries must be generated conditioned on each perspective type (see Figure 1 for sample input-output patterns).

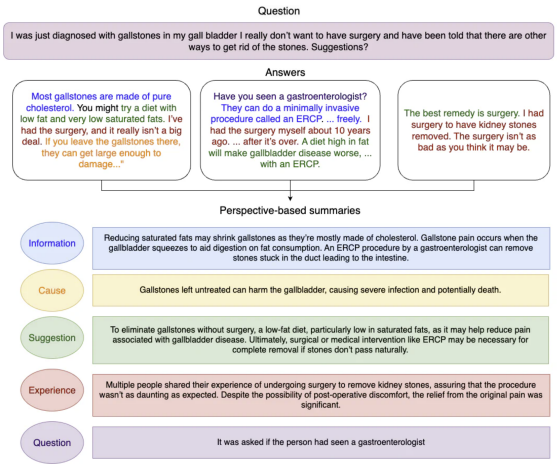


Figure 1: Image from PerAnsSumm (Agarwal et al., 2025) Task.[2] Task A: Span Identification and Classification (color-highlighted spans in answers); Task B: Summary Generation (Perspective-based summaries).

Since both tasks are inherently perspective-conditioned, they require models that are explicitly aware of these perspectives. We propose a unified

[2] https://peranssumm.github.io/docs/

Proceedings of NLPerspectives @ LREC 2026, pages 124–135
12 May 2026.

framework centered on *perspective conditioning*. We first introduce a multi-label perspective classifier using BART (Lewis et al., 2020), whose outputs form perspective signals for downstream models. These signals are then used in two alternate architectures: one using perspective-conditioned prefix vectors inspired by PLASMA (Naik et al., 2024), the other using a perspective-aware attention layer (inspired by AWAN (Tahaei and Bergler, 2025) and label attention (Vu et al., 2020)) that encodes the active perspective into token representations. Both mechanisms allow the model to selectively attend to and generate content that is grounded in a specific viewpoint, rather than aggregating indiscriminately across perspectives. This paper explores how conditional representations can improve both span-level identification and abstractive summarization in multi-perspective settings.

2. Prior Work

Perspective summarization for healthcare CQA settings was formally introduced by (Naik et al., 2024), who proposed the PUMA dataset, a collection of 3167 CQA threads from Yahoo! Answers annotated with five perspective types: cause, suggestion, experience, question, and information. They also proposed PLASMA, a prompt-driven controllable summarization model built on Flan-T5 with a prefix tuner and an energy-controlled perspective loss that enforces perspective-specific attributes in the generated summary. PLASMA outperformed five baselines across ROUGE, METEOR, BERTScore, and BLEU metrics, establishing a strong benchmark for the task on the PUMA dataset.

The PerAnsSumm 2025 Shared Task (Agarwal et al., 2025) had two subtasks: perspective span identification and classification (Task A), and perspective-based answer summarization (Task B), both evaluated on the PUMA dataset supplemented with a new test set of 50 samples. Large Language Models dominated the competition; 18 of 23 teams used LLMs in some capacity, including all top-10 teams. The top-performing system, WisPerMed (Pakull et al., 2025), achieved high performance with DeepSeek-R1 for span extraction and instruction-tuned Mistral-7B for summarization, while YALENLP (Jang et al., 2025) leveraged GPT-4o in a zero-shot setup, achieving the best scores on both summarization and span recognition. In general, there was a shift from fine-tuning paradigms toward in-context learning and prompt-based inference.

A different approach to multi-perspective answer summarization in CQA forums is *AnswerSumm* (Fabbri et al., 2021), which used an automated pipeline for creating bullet-point abstractive summaries by clustering relevant answer sentences and using cluster centroids as summary targets. To improve coverage and faithfulness, they proposed a multi-reward reinforcement learning objective combining ROUGE (Lin, 2004), NLI-based entailment (Bowman et al., 2015), and semantic area rewards, alongside a sentence-relevance prediction auxiliary loss. Their analysis showed that supervision from multi-perspective data inherently leads models to generate diverse, multi-viewpoint summaries, and that the quality of the NLI model significantly affects downstream performance.

3. Data and Tasks

We use the PerAnsSumm (Perspective-aware Healthcare Answer Summarization) dataset (Agarwal et al., 2025). The dataset consists of healthcare community question answering (CQA) threads, where each instance contains a question Q, a set of answers $\mathcal{A}$, and annotated perspective information.

The predefined set of perspective categories is:

$\mathcal{P} = \{$ *cause, suggestion, experience, question, information* $\}$

The shared task consists of two complementary objectives: (i) identifying and classifying perspective-specific spans in answer texts (Task A), and (ii) generating perspective-specific summaries (Task B) (see Figure 1).

In addition to these tasks, we introduce a *multi-label perspective classification task* as a preliminary step for both Task A and Task B[3]. This task involves classifying each question–answer pair given in the input into one or more perspective categories. The classification of question–answer pairs into perspective categories in a multi-label setting acts as a first step for perspective span recognition and perspective summarization. The quality of the classification models directly influences the performance of the downstream tasks, as the predicted perspectives serve as conditioning signals for both span identification and summarization.

The objective of Task A is to identify spans in the answer text that reflect a particular perspective and classify each span into the corresponding perspective category. These perspective spans represent fine-grained semantic units that characterize how different viewpoints are expressed within answer texts.

The objective of Task B is to generate a concise summary of a question–answer thread that reflects a specific target perspective $p \in \mathcal{P}$. Given a question Q, its associated set of candidate answers $\mathcal{A}$, and a target perspective p, the model is required to generate a summary Y_p that captures only the information relevant to perspective p from $\mathcal{A}$.

[3]We are using the PerAnsSumm 2025 data and evaluation but did not participate in the competition.

The dataset contains 2,533 training instances, 959 validation instances, a test-seen split of 640 instances (a subset of the validation set), and 50 instances in the official test set.

4. Preprocessing

We perform task-specific preprocessing for classification, span recognition, and summarization.

Multi-label Classification For multi-label perspective classification, we construct question–answer pairs (q, a) consisting of a question a and one answer q from the corresponding answer set $\mathcal{A}$ belonging to the CQA thread. Using the span label annotation in the training data, we label each answer with the set of perspectives present in it. An answer a is assigned a perspective p_i if at least one annotated span in that answer corresponds to perspective category p_i. This results in a dataset of question–answer pairs with multi-label perspective targets for classification.

Span Identification For span recognition, we construct perspective-conditioned instances for each perspective p_i identified in an answer. For every answer and its associated perspectives, we extract the spans labeled with that perspective along with their character-level offsets. These spans serve as target labels and are used only during training. Since the provided offsets in the dataset are defined over the raw JSON text, we realign the span offsets to match the processed answer text used during model training. The corrected character offsets are then converted into token-level BIO labels to formulate the task as a sequence tagging problem. For the baseline perspective span identification model, we adopt a joint tagging formulation where all perspective spans are predicted simultaneously. In this setup, we define separate B–I label pairs for each perspective category (e.g., B-info, I-info, B-suggestion, I-suggestion, etc.), while the O label remains shared across all perspectives. Unlike our perspective-conditioned models, which process one perspective at a time using question–answer–perspective triples, the baseline model operates on question–answer pairs without explicit conditioning. For each question–answer pair, we associate the full set of perspectives identified in that answer. The token-level labels, therefore, include spans from all perspectives within the same sequence, each annotated with its corresponding BIO tags. This formulation requires the model to jointly identify and distinguish spans belonging to multiple perspectives within a single tagging space.

Summarization For summarization, the model generates summaries across the entire set of answers for a given question, conditioned on a target perspective. The input is constructed by concatenating a short perspective-specific prompt, the question, and all associated answers into a single sequence. During training, the dataset is expanded such that each question–answer thread is paired separately with each of its perspective-specific gold summaries. This ensures that the model learns to generate one summary per perspective for each question–answer thread. The preprocessing does not change for the baseline summarization model.

5. Perspective Classification

5.1. MLC: BART Encoder-based Classification

This system is developed and fine-tuned for multi-label classification. We employ the encoder component only of the transformer-based BART (Lewis et al., 2020) as the backbone, because BART has the ability to process longer input sequences compared to architectures such as BERT and RoBERTa. The BART encoder can accommodate the larger context of the full question–answer pairs with longer answers.

The input question–answer pair is concatenated and fed into the BART encoder. For classification, we use the final hidden representation of the last token (EOS), which serves a role analogous to the [CLS] token in BERT-based models.

Since perspective identification is formulated as a multi-label classification task, a linear classification layer projects the encoder representation into a vector of dimension $|\mathcal{P}|$, corresponding to the number of perspective categories. A sigmoid activation function is applied to obtain independent probability scores for each perspective, and the model is trained using binary cross-entropy loss with class weights.

5.2. LLC: LLM-based Classification

In addition to the supervised classifier, we develop an LLM-based classification system for identifying perspective categories. This system serves as a robustness baseline by leveraging the few-shot prompting techniques with large language models to perform multi-label perspective identification without task-specific fine-tuning.

The LLM is prompted to assign one or more perspective categories to each question–answer pair. Comparing the supervised and LLM-based classification systems allows us to analyze their impact on downstream span identification and perspective-aware summarization on the test set. The full

prompt used for the LLM-based classifier is provided in Appendix 13.1

5.3. Performance Comparison

	CM-F1	CW-F1
MLC	71.26	79.91
LLC	72.47	82.17

Table 1: Perspective Classification results. Column header definitions: CM-F1: Classification Macro F1, CW-F1: Classification Weighted F1. Row definitions: MLC: BAT-based Multi-label Classification, LLC: LLM-based Classification

Table 1 compares the performance of the two classifiers. The LLM-based few-shot system outperforms our BART-derived system, and we use it for all other experiments exclusively.

6. Baselines

6.1. BLA: Perspective Span Detection

A BART encoder paired with a CRF layer serves as the baseline for the perspective span identification task. We fine-tune a BART encoder (Lewis et al., 2020) followed by a token-level classification layer and a Conditional Random Field (CRF) (Lafferty et al., 2001).

The CRF layer is applied to model dependencies between adjacent labels and enforce valid BIO tag transitions during decoding, which is commonly used in sequence labeling tasks (Huang et al., 2015).

This baseline model does not incorporate perspective-conditioned signals. Instead, it performs joint multi-perspective span identification using a unified tagging space. Specifically, the model predicts spans for all perspectives simultaneously within a single sequence tagging formulation.

6.2. BLB: Perspective Summarization

This system serves as the baseline for the perspective summarization task. We fine-tune a BART sequence-to-sequence model (Lewis et al., 2020) without incorporating any perspective-conditioning modules. The model operates in a standard encoder–decoder setting, where the input consists of the question and all associated answers concatenated into a single sequence.

The only explicit perspective signal is provided through a perspective-specific prompt that is prepended to the input text (Appendix 13.1.2). Apart from this prompt-based conditioning, no additional architectural modifications or perspective-aware mechanisms are introduced.

7. Span Identification

7.1. PTA: Prefix-Conditioned Span Identification

We implement a prefix-conditioned span identification model that extends the baseline BART+CRF architecture by introducing a prefix module, following the prefix-tuning paradigm (Li and Liang, 2021). For each target perspective, a prefix representation is generated using the prefix module containing a prefix encoder and a prefix MLP.

The prefix encoder maps a given perspective prompt into a fixed-dimensional embedding using a sentence transformer (Reimers and Gurevych, 2019) model. This 768-dimensional embedding is then transformed by a learnable prefix MLP into a sequence of k dense prefix vectors, where k denotes the prefix length. These k prefix vectors are prepended to the input token embeddings before being passed to the BART encoder.

Unlike standard prefix-tuning approaches (Naik et al., 2024; Li and Liang, 2021), where the backbone transformer model remains frozen, we jointly optimize both the prefix modules and the encoder parameters. Since span labels are defined only over the original input tokens, we discard the first k encoder representations corresponding to the prefix tokens and apply the classifier and CRF only to the remaining token representations.

We conduct an ablation study by varying the prefix length to analyze its impact on span extraction performance (see Table 2).

7.2. PAA: Perspective-aware Attention for Span Identification

In this alternate approach, we extend a baseline *BART+CRF* sequence labeling model by introducing a *perspective-aware attention layer* between the encoder and the token-level classifier. This layer explicitly conditions token representations on a target perspective by injecting a learned perspective embedding through cross-attention. Because each token attends to the perspective embedding, the resulting representations become perspective-dependent, encouraging the model to emphasize tokens that are most indicative of perspective-specific span prediction.

To preserve the pretrained knowledge of BART, the encoder parameters are kept frozen during training. Only the perspective-aware attention modules, the token-level classification layer, and the CRF decoding layer are trained. This design allows the model to learn perspective-specific span extraction while limiting the number of trainable parameters.

Given an input answer A_i, the encoder produces contextualized token representations:

$$H = \{h_1, h_2, \ldots, h_n\}, \qquad h_i \in \mathbb{R}^d$$

For a target perspective p, we obtain a perspective embedding $z_p \in \mathbb{R}^d$, which is learned as a trainable parameter.

We implement a cross-attention mechanism where the encoder outputs act as the Query (Q), while the perspective embedding provides the Key (K) and Value (V):

$$Q = HW_Q, \quad K = z_p W_K, \quad V = z_p W_V,$$

where W_Q, W_K, W_V are learnable projection matrices. The attention output is computed as:

$$\text{Attn}(H, z_p) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V.$$

This attention mechanism injects perspective-specific information into the token representations by allowing each token to attend to the target perspective embedding. The resulting representation is combined with the original encoder representations using a residual connection:

$$\tilde{H} = H + \text{Attn}(H, z_p).$$

To capture perspective-specific patterns, we train a separate attention module for each perspective. That is, for each $p \in \mathcal{P}$, a distinct attention layer with its own parameters is optimized independently. This formulation models span recognition as a perspective-conditioned binary sequence labeling task.

The enriched token representations $\tilde{H}$ are then passed through a linear classification layer to predict token-level BIO labels for span extraction.

We conduct an ablation study comparing a single key vector with multiple key vectors per perspective to evaluate how the number of perspective keys affects span extraction performance (see Table 2).

8. Perspective Summarization

8.1. PTB: Prefix-Conditioned Summarization

We implement a prefix-conditioned summarization model by extending the baseline BART encoder–decoder architecture with a prefix module. The prefix module follows the same design as in the prefix-conditioned span identification model, where a given perspective prompt is encoded and transformed into a sequence of k prefix vectors that are prepended to the encoder input embeddings. Each prefix vector is also transformed to the BART-large embedding dimension of 1024 in the prefix MLP layer.

Unlike the span identification setting, we do not discard the first k encoder representations, as summarization is a generation task. The decoder attends over the full set of encoder representations, including the prefix tokens, allowing the enriched perspective-conditioned signals to influence content selection during generation.

We conduct the same ablation study as performed for the span identification task (see Table 3).

8.2. PAB: Perspective-aware Attention for Summarization

We extend a baseline BART encoder-decoder summarization model by adding a perspective-aware attention layer between the encoder and the decoder. The approach is similar to perspective-aware attention for span recognition. The primary difference from the span recognition model lies in the backbone transformer and the role of the attention outputs in generation.

Given a question Q, its associated answers $\mathcal{A}$, and a target perspective p, the objective is to generate a summary Y_p that captures content aligned with the specified perspective. The input sequence is first encoded by the BART encoder to produce contextual representations

$$H = \{h_1, h_2, \ldots, h_n\}, \qquad h_i \in \mathbb{R}^d$$

As in the span recognition model, the perspective-aware attention module conditions these encoder representations on a learned perspective embedding z_p, producing enriched representations $\tilde{H}$. Unlike the span recognition setting, where these representations are used for token classification, the enriched encoder states $\tilde{H}$ are provided to the BART decoder to generate the summary autoregressively.

To preserve the pretrained knowledge of the summarization model, both the BART encoder and decoder are kept frozen during training. Only the perspective-aware attention modules are fine-tuned. As in the span recognition system, we train separate attention modules for each perspective, resulting in perspective-specific parameters that are optimized independently.

During inference, the multi-label classification system predicts the set of relevant perspectives P_i for a given input. For each predicted perspective $p \in P_i$, the corresponding attention module is activated to produce perspective-conditioned encoder representations $\tilde{H}$, from which the decoder generates a perspective-specific summary.

We further conduct an ablation study comparing the use of a single key vector with multiple key vectors per perspective to analyze how the number of perspective keys influences summarization performance (see Table 3).

9. Implementation

9.1. Training Setup

All models are implemented using the HuggingFace Transformers library and optimized using the AdamW optimizer (Loshchilov and Hutter, 2019). Parameters belonging to the BART backbone are trained with a learning rate of 5×10^{-5}, a commonly used setting for fine-tuning pretrained transformers (Lewis et al., 2020; Devlin et al., 2019). Unless otherwise specified, models are trained for 30 epochs.

Two learning rate scheduling strategies are used. Prefix-conditioned systems use ReduceLROnPlateau to adapt the learning rate when validation loss plateaus. Perspective-aware attention systems use a LambdaLR scheduler with a warmup followed by cosine decay, which gradually increases the learning rate early in training and then smoothly decays it, improving training stability (Vaswani et al., 2017; Loshchilov and Hutter, 2017).

Span identification models are trained using a joint objective combining Conditional Random Field (CRF) loss and token-level cross entropy (CE):

$$L = L_{CRF} + \lambda_{CE} L_{CE}$$

where $\lambda_{CE} = 0.7$. Weighted CE is used to address class imbalance in BIO labels with weights $[0.524, 32.61, 1.944]$ for O, B, and I classes. All span identification systems use the BART-base encoder (Lewis et al., 2020) with a maximum sequence length of 899 tokens. Summarization systems use BART-large-CNN with a maximum length of 512 tokens, except for prefix-based summarization which uses 899 tokens. Gradient clipping is applied for attention-based span identification and summarization models.

Prefix-Conditioned Span Identification The prefix encoder is implemented using a sentence-transformer (all-mpnet-base-v2 (Song et al., 2020)), with the prefix encoder and the first six layers of the BART encoder frozen to stabilize training. The prefix encoder uses a learning rate of 1×10^{-5} with weight decay 0.1, while the prefix projection MLP uses 3×10^{-4} with weight decay 1×10^{-4}. The CRF layer is trained with 1×10^{-2} and the classification layer with 2×10^{-4}.

Perspective-aware Attention for Span Identification The BART encoder remains frozen while only the attention and span prediction layers are trained. Due to slower convergence observed in validation, the model is trained for 50 epochs.

BART Encoder with CRF The baseline span identification system uses the BART-base model's encoder followed by a CRF layer for sequence labeling. Similar to other span systems, training uses the combined CRF and weighted CE loss.

Prefix-conditioned Summarization The prefix encoder uses a sentence-transformer backbone where the first six layers are frozen. Within the BART encoder-decoder backbone, most layers are frozen while the final two layers of both encoder and decoder are unfrozen for task adaptation. The prefix encoder uses a learning rate of 1×10^{-5} with weight decay 0.1, and the prefix MLP uses 3×10^{-4} with weight decay 1×10^{-4}.

Perspective-aware Attention for Summarization This system introduces a perspective-aware attention layer between the encoder and decoder of the BART-large-CNN model with a maximum input length of 512 tokens. The loss for the EOS token is down-weighted by 0.2 to reduce bias toward early sequence termination while still allowing the decoder to learn appropriate stopping behavior.

Baseline Summarization The baseline summarization system fine-tunes the standard BART-large-CNN encoder-decoder architecture for perspective summarization with a maximum token length of 512.

BART-based Multi-label Classification Perspective classification is performed using a BART-base encoder with a token length of 899 and a multi-label classifier applied to the EOS token representation. The model is trained using weighted binary cross entropy (BCE) loss to address class imbalance across perspective categories, with a learning rate of 1×10^{-5}.

LLM-based Classification We additionally evaluate a prompt-based classification system using the Qwen3-8B-AWQ (Yang et al., 2025) model. The model is hosted via vLLM[4] for faster and efficient inference. A few-shot prompting strategy is used to classify question-answer pairs into their respective perspective categories.

9.2. Evaluation and Outcomes

We follow the evaluation protocol defined in the PerAnsSumm shared task (Agarwal et al., 2025). Classification performance in Table 1 is measured using Macro F1 and Weighted F1 scores. Macro F1 treats all perspective classes equally, while Weighted F1 accounts for class imbalance by weighting each class according to its support.

[4] https://docs.vllm.ai/en/latest/

	K/PL	CM-F1	CW-F1	SM-F1	PM-F1
PAA	1	74.54	82.0	8.18	**61.96**
PAA	5	**75.92**	**82.65**	9.91	59.92
PAA	16	**75.34**	**82.43**	**10.62**	58.80
PTA	1	72.97	80.50	**10.89**	41.87
PTA	5	73.68	**82.47**	**11.00**	52.14
PTA	16	73.12	81.63	**11.06**	52.50
BL	-	50.94	62.62	4.73	28.99

Table 2: Perspective Span Classification and Identification results. Column header definitions: K/PL: Key/Prefix Length, CM-F1: Classification Macro F1, CW-F1: Classification Weighted F1, SM-F1: Strict Matching F1, PM-F1: Proportional Matching F1 (Only F1 scores, the precision and recall metrics are available in Table 4 in Appenedix 13.2) Table row definitions: PAA: Perspective-Attention, PTA: Prefix Tuning, BLA: Baseline (BART+CRF)

	K/PL	R-1	BS	MT	BU
PAB	1	24.97	79.32	18.93	4.23
PAB	5	34.20	**81.17**	26.85	8.24
PAB	16	37.47	**81.75**	**30.03**	9.68
PTB	1	36.41	**81.13**	**30.18**	**11.57**
PTB	5	**39.09**	**81.64**	**32.15**	**12.24**
PTB	16	36.15	**81.33**	**29.52**	10.58
BLB	-	**36.70**	**81.70**	29.02	8.97

Table 3: Perspective Summarization results. Column header definitions: K/PL: Key or Prefix Length, R-1: ROUGE-1, BS: BERTScore, MT: METEOR, BU: BLEU. Row definitions: PAB: Perspective-Attention, PTB: Prefix Tuning, BLB: Baseline BART encoder-decoder. ROUGE-L and 2 are available in Table 5 (Appendix 13.2)

Span recognition in Table 4 is evaluated using token-level F1 scores under two matching criteria: strict matching and proportional matching. Strict matching requires exact alignment between predicted and gold span boundaries, while proportional matching measures maximum token-level overlap between predicted and gold spans, allowing partial credit for boundary mismatches.

Summarization systems are evaluated only using ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020) in Table 5. These scores evaluate the lexical and semantic metrics similarity between the reference and prediction summaries.

For evaluation, we use the test-seen data to evaluate locally, as the Codabench server[5] of the shared task does not work at times in the post-evaluation phase. We use LLM classification for all the systems, as we observed a slight improvement in performance from the BART-based Multi-label Classification (Table 1). The results are available in Tables 2, 3, 4, and 5.

Classification Task We first analyze the classification results derived from span predictions. A perspective is considered present if at least one non-'O' span exists. If the gold annotation contains no span for a perspective and the model predicts one, the prediction is counted as incorrect.

We observed that the BART-based classifier performs slightly worse than the LLM-based classifier on the test set (Table 1). Therefore, we use the LLM classifier for both the span identification and summarization systems to maintain consistency across tasks. However, the baseline span identification does not use the classification output, as they are trained to predict all perspective spans jointly.

Span Identification Table 4 shows that the baseline underperforms in almost all span metrics except Proportional Matching Precision (PM-P). While it achieves high PM-P (74.25), its Proportional Matching Recall (18.01) is very low, leading to poor overall PM-F1 (28.99). This suggests that the baseline predicts very few spans and does so conservatively.

Both PAA and PTA substantially improve strict and proportional F1 scores. Strict Matching F1 improves from 4.73 in the baseline to 10.62 for PAA ($K = 16$) and 11.06 for PTA ($PL = 16$). Proportional Matching F1 improves from 28.99 in the baseline to 61.96 for PAA ($K = 1$) and 52.50 for PTA ($PL = 16$).

A significant difference is observed in Proportional Matching Recall (PM-R) between PAA and PTA. PAA consistently achieves higher PM-R (above 50 across key sizes), whereas PTA shows substantially lower PM-R, particularly at smaller prefix lengths (e.g., 30.54 at $PL = 1$). This indicates that PAA is better at recovering overlapping spans with the gold annotations, while PTA tends to be more conservative in recall despite comparable strict F1 at higher prefix lengths.

PAA uses approximately 2.36M trainable parameters across key sizes ($K = 1$: 2.36M, $K = 5$: $\sim$2.36M, $K = 16$: 2.37M), compared to 139.42M in the fully fine-tuned baseline. Despite this large reduction in trainable parameters, PAA significantly outperforms the baseline across nearly all metrics. As the number of keys increases, Strict Matching F1 improves, while Proportional Matching F1 slightly decreases. Empirically, $K = 5$ provides the best trade-off between strict and proportional performance, indicating that too few keys underrepresent perspective cues, while too many may introduce redundancy given the dataset size.

Prefix tuning uses substantially more trainable parameters ($PL = 1$: 86.23M, $PL = 5$: 87.42M,

[5] https://www.codabench.org/competitions/4312/

$PL = 16$: 90.67M). As prefix length increases, Strict F1 slightly improves, and Proportional Matching Recall increases, indicating greater token overlap between predicted and gold spans. However, even at higher prefix lengths, PTA does not match the proportional PM-R levels achieved by PAA in Table 4.

The span systems were not explicitly trained with negative spans (i.e., cases where all tokens are labeled 'O' for a perspective). Despite this, both PAA and PTA remain robust as key or prefix length increases, as reflected in stable classification scores and improved recall. To further experimentally assess to what extent the perspectives are formed during training and to what extent the LLM can extrapolate to unseen perspectives, we created synthetic data introducing a new unseen perspective termed *reassurance*. We tested the trained prefix-based span system without retraining. Out of three synthetic examples, two showed reasonable overlap between predicted and synthetic gold spans. While this indicates partial generalization to unseen perspectives, performance was inconsistent. This may be due to the limited dataset size, imbalance between existing perspectives, and the fact that only five perspectives were used during training. To better demonstrate the capacity of the prefix MLP to encode new perspectives, experiments with a larger number of perspectives and more balanced data would be necessary.

Summarization Table 3 presents the results for perspective-aware summarization. In contrast to the span identification task, the differences between the baseline, PAB, and PTB systems are less pronounced. Most configurations perform very close to the fully fine-tuned baseline, with several slightly underperforming it.

The baseline achieves strong overall performance (R-1: 36.70, BERTScore: 81.70, METEOR: 29.02, BLEU: 8.97). PAB at $K = 16$ (R-1: 37.47, METEOR: 30.03) and PTB at $PL = 5$ (R-1: 39.09, METEOR: 32.15, BLEU: 12.24) slightly outperform the baseline on multiple metrics. However, the improvements are modest, and overall performance remains within a narrow range across systems.

What is particularly noteworthy is the parameter efficiency. The baseline uses approximately 406M trainable parameters, whereas PAB uses only about 4.2M parameters, which is roughly 1% of the baseline. Despite this drastic reduction in trainable parameters, PAB performs very close to the fully fine-tuned model. This highlights the effectiveness of attention-based perspective conditioning as a parameter-efficient fine-tuning strategy and emphasizes the importance of explicitly modeling perspectives in summarization.

We also observe that, unlike in span identification, PAB benefits from increasing the number of keys in summarization, with performance steadily improving from $K = 1$ to $K = 16$. This suggests that richer perspective embeddings help guide generation more effectively in a sequence-to-sequence setting. PTB exhibits a similar but less stable trend, with the best performance at $PL = 5$ and slight degradation at $PL = 16$, indicating that longer prefixes may introduce redundancy or noise during generation. PTB uses substantially more trainable parameters ($PL = 1$: 70.14M, $PL = 5$: 72.24M, $PL = 16$: 78.02M).

Despite incorporating a weighted EOS loss in the attention-based summarization system, we do not observe consistent improvements over PTB. This suggests that the choice of the EOS weighting hyperparameter λ_{eos} may require further tuning. Overall, the results indicate that perspective-aware modeling achieves competitive performance with significantly fewer trainable parameters, demonstrating strong parameter efficiency while maintaining summarization quality.

10. Conclusion

This work demonstrates that explicitly modeling perspective is not merely an auxiliary enhancement but a structural principle for multi-view text understanding. By separating perspective signals from surface lexical cues and injecting them directly into representation learning, our framework reframes span identification and summarization as conditioned reasoning tasks rather than generic text processing problems. The results show that perspective conditioning reshapes token-level and sequence-level representations in meaningful ways, enabling models to distinguish overlapping perspectives within the same discourse. Importantly, the effectiveness of lightweight attention modules suggests that perspective control does not require extensive parameter updates but instead benefits from targeted representational steering. This highlights that explicit perspective conditioning introduces modular and controllable structure into transformer models, enabling targeted analysis of perspective-specific behavior while maintaining parameter efficiency. Future work will explore perspective-specific adapter modules and supervised contrastive objectives to encourage stronger separation between perspective representations. We also plan to explore prefix-based conditioning strategies that better support generalization to previously unseen perspectives. Additionally, evaluating the model across domains will help determine whether it captures abstract viewpoint structures or relies on domain-specific patterns, thereby providing deeper insight into how perspective-sensitive knowledge is represented and transferred.

11. Acknowledgements

We thank Narjes Tahaei for her comments on this paper. The work was conducted with the support of a NSERC DG grant.

Parts of this paper were edited for clarity and fluency with the assistance of an AI language tool. All scientific content and conclusions remain the responsibility of the authors.

12. Bibliographical References

Siddhant Agarwal, Md. Shad Akhtar, and Shweta Yadav. 2025. Overview of the PerAnsSumm 2025 Shared Task on Perspective-aware Healthcare Answer Summarization. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 445–455, Albuquerque, New Mexico. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Samuel 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.

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.

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.

Alexander R Fabbri, Xiaojian Wu, Srini Iyer, and Mona Diab. 2021. Multi-Perspective Abstractive Answer Summarization. *arXiv preprint arXiv:2104.08536*.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist Approaches to Natural Language Processing: a survey. *Language Resources and Evaluation (LREC)*, 59(2):1719–1746.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *ArXiv*, abs/1508.01991.

Dongsuk Jang, Haoxin Li, and Arman Cohan. 2025. YaleNLP @ PerAnsSumm 2025: Multi-Perspective Integration via Mixture-of-Agents for Enhanced Healthcare QA Summarization. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 415–427.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *ICML*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 7871–7880.

Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. 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 4582–4597.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of Text Summarization Branches Out Workshop at ACL 2004*, pages 74–81.

Ilya Loshchilov and Frank Hutter. 2017. Sgdr: Stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Gauri Naik, Sharad Chandakacherla, Shweta Yadav, and Md Shad Akhtar. 2024. No perspective, no perception!! Perspective-aware Healthcare Answer Summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*,

pages 15919–15932, Bangkok, Thailand. Association for Computational Linguistics.

Tabea Pakull, Hendrik Damm, Henning Schäfer, Peter Horn, and Christoph Friedrich. 2025. WisPerMed @ PerAnsSumm 2025: Strong Reasoning Through Structured Prompting and Careful Answer Selection Enhances Perspective Extraction and Summarization of Healthcare Forum Threads. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 359–373.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 3982–3992.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and Permuted Pre-training for Language Understanding. *Advances in neural information processing systems (NEURIPS)*, pages 16857–16867.

Narjes Tahaei and Sabine Bergler. 2025. Beyond Consensus: Use of Demographics for Datasets that Reflect Annotator Disagreement. In *First Workshop on Bridging NLP and Public Opinion Research at COLM2025*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems (NeurIPS)*.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for icd coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3335–3341. International Joint Conferences on Artificial Intelligence Organization. Main track.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations, ICLR 2020*.

13. Optional Supplementary Materials

13.1. Appendix A

13.1.1. Perspective Classification Prompt

```
You are a perspective
classification agent for a given
text.
You need to identify different
perspectives in the given text
and classify them into one of the
following categories:
1. EXPERIENCE
2. INFORMATION
3. CAUSE
4. SUGGESTION
5. QUESTION
Guidelines: - A perspective can
occur only once in a text.
- One text can have up to 5
unique perspectives.
- Do not repeat the same
perspective again.
Refer to the following examples
for guidance.
Example 1:
Question: what is parkinesonism?
Answer: u spelt it wrong
!!Parkinson's disease is one of
the most common neurologic
disorders of the elderly. The
term "parkinsonism" refers to any
condition that causes any
combination of the types of
movement abnormalities seen in
Parkinson's disease by damaging
or destroying dopamine neurons in
a certain area of the brain.
Output: ["INFORMATION"]
Example 2:
Question: I scream, shout and
swear in my sleep. How do I
stop?
Answer: I think that you have a
stress on your daily life. I
think that is not bad to do some
following things, If you caould
find any way, you will be happy,
otherwise, you have to go to a
psyc.
1) Keep out yourself from stress,
during your day.
2) Try to sleep, just when you
are really tired. (excuse me for
this example!) Like after a
sweet sex! try to find that are
you in this situation after sex,
or not? If No, it shows that you
try to sleep, before it needs.
3) Read Book, or newsletter
before sleep.
4) Drink one glass warm (NOT HOT)
milk.
5) Do some sort of excersice
before sleeping.
6) If you have any problem in
your dream, try to solve it by
someone that you are fighting
with. I mean, before your
sleeping (Specially when your
husband is not at home, because I
want to nobody awake you) try to
solve your problem with your
dream fighter!! Yes, it is funny
but true. Try to find a logical
way for treating out this
conflict with your dream.
i wish a good dream and sweet
night, beside of your sweet
husband.
Output: ["SUGGESTION", "CAUSE"]
Example 3:
Question: Are their any good
home remedies for tooth pain?
Answer: yes- keep water in your
mouth for 24 hours a day.
i drank about 5 each 16 oz
bottles for a few days (the
coolness of the water actually
moderated the pain). After a few
days, no pain. I had flushed it
clean and was able to function
until I could get to a dentist.
Output: ["EXPERIENCE"]
Example 4:
Question: Is 24 too old to
consider become pregant?
Answer: Only the 24 year old can
ask themselves that question --
are they personally ready?
Typically it's good to make sure
you're financially lined up to
have a baby (makes life easier),
and that your home environment
would be condusive to a baby
being around, but it's the
parents that need to know if
they're ready. I know 35+ year
olds that weren't ready yet.
Good luck! Output: ["QUES-
TION","SUGGESTION","EXPERIENCE"]
With the examples above, classify
the answer for the given question
into at least one of the given
categories. The answer should be
in the same format as the output
of the examples. There can be
more than one category for each
answer.
Think step-by-step before
deciding the output.
The text is: Question:
{question} Answer: {answer}
Output:
** Return only the output, do not
return anything else **
```

13.1.2. Summarization Prompt

```
For the {perspective}
perspective, summarize the given
answer for the question below:
Question: {question}
Answers: {answers}
```

13.2. Appendix B

Table 4 shows detailed span matching metrics.

	K/PL	SM-P	SM-R	PM-P	PM-R
PAA	1	9.37	7.25	**68.98**	**56.24**
PAA	5	10.76	9.18	67.02	54.19
PAA	16	**11.03**	10.23	67.09	52.33
PTA	1	10.80	10.98	66.58	30.54
PTA	5	10.49	11.55	62.39	44.78
PTA	16	10.58	**11.59**	63.27	44.86
BLA	-	8.96	3.22	74.25	18.01

Table 4: Perspective Span Classification and Identification results. Column header definitions: K/PL: Key/Prefix Length, SM-P: Strict Matching Precision, SM-R: Strict Matching Recall, PM-P: Proportional Matching Precision, PM-R: Proportional Matching Recall. Table row definitions: PAA: Perspective-Attention, PTA: Prefix Tuning, BLA: Baseline

Table 5 shows ROUGE-2 and ROUGE-L (Lin, 2004) metrics for Summarization.

	K/PL	ROUGE-2	ROUGE-L
PAB	1	9.61	22.0
PAB	5	15.55	30.72
PAB	16	**18.84**	**34.01**
PTB	1	17.33	32.39
PTB	5	**18.90**	**34.84**
PTB	16	**17.84**	32.29
BLB	-	**18.12**	32.28

Table 5: ROUGE-2 and ROUGE-L scores for summarization. Column header definitions: K/PL: Key/Prefix Length. Table row definitions: PAB: Perspective-Attention, PTB: Prefix Tuning, BLB: Baseline

A Pilot Study Investigating Stakeholder Subjectivity in Collaborative Dialog Analysis

Ananya Ganesh[1,2], Martha Palmer[1], Katharina von der Wense[1]
[1]University of Colorado, [2]University of Wisconsin–Madison
Correspondence: aganesh27@wisc.edu

Abstract

Qualitative research in education relies on"ground truth" codes or labels generated by having a trained or expert coder code observations in data such as student dialog. Although rigorous validity checks are a part of the coding process, there is limited research investigating how and to what extent, this notion of the ground truth is influenced by inherent task subjectivity. This paper presents a pilot study of task subjectivity centered around the phenomenon of verbal off-task behavior. The context for this study is real-world small-group collaborative conversations among three to five students in a middle-school science classroom. To investigate how stakeholders such as teachers and students show subjectivity in approaching this task, we recruit five teachers from the Prolific online platform, and five students from local middle and high schools as annotators of off-task speech. We show that teachers, students, and expert coders differ in their perception of off-task speech, with some of these differences being systematic. Drawing upon recent research in machine learning and natural language processing, we then outline the potential benefits of collecting and modeling a *range* of codes that explicitly represent the subjective perspectives of a diverse set of coders.

Keywords: perspectivist approaches for social good, NLP for education, stakeholder subjectivity

1. Introduction and Background

Collaborative learning – the process through which multiple students come together to interactively learn a common topic – has been recognized as a critical component of modern classroom environments (Graesser et al., 2020). Apart from facilitating the socio-cognitive development of learners (Vygotsky, 1978), specific skills imparted to students through the collaborative learning process include argumentation (Osborne, 2010), discussion and negotiation (Roschelle, 1992), and shared knowledge construction (Bereiter, 2002). Towards understanding collaborative learning, learning analytics research has underscored the importance of students' dialogic interactions. However, manual qualitative analysis of student dialog poses a heavy demand on the time and labor of qualitative researchers, and is thus constrained by speed and scale. Consequently, there has been much interest in the development of computational models both to analyze collaborative dialog (Rosé et al., 2008; Yin et al., 2025) and to improve collaboration through automated interventions (D'Mello et al., 2024) based on such analysis.

Data-driven analysis and modeling of collaborative dialog is typically centered around "ground truth" observations of behaviors of interest. Examples of relevant behaviors include displays of collaborative problem solving skills such as *negotiation* with team members (Sun et al., 2020). Human coders or annotators are trained to identify such observations, typically from video data, but also from textual transcripts and audio recordings. The coding process comprises multiple stages of validation where inter-rater reliability is measured and the agreement between raters is iteratively improved through discussions and by refining the codebook based on nuances observed in the data (Reitman et al., 2023). The motivating factor in this process is the need to minimize disagreement, i.e., codes are thought to be more reliable when they are independently produced by multiple coders. Moreover, when the goal is to use the generated codes to train automated classifiers, the individual codes or labels may undergo further aggregation such as by selecting a single label through majority or ensemble voting.

However, several papers in machine learning (ML) and natural language processing (NLP) have argued that disagreements in annotation should be "embraced", that is, used in downstream analysis orin classification models, rather than pruned (Reidsma and op den Akker, 2008; Plank et al., 2014b; Alm, 2011). Arguments include better data utilization (Plank, 2022) (i.e., not discarding human feedback particularly with already scarce datasets), improved estimates of predictive uncertainty (Khurana et al.), the possibility of multiple valid answers (Alm, 2011), as well as the need to model diverse perspectives, since individual experiences may affect the way that text may be interpreted (Prabhakaran et al., 2021). Modeling diverse perspectives is also considered a step towards robust and reliable models that minimize bias (Kirk et al., 2024).

Subjective perspectives have been shown to depend on the socio-demographic backgrounds of annotators and their lived experiences. Waseem et al. (Waseem, 2016) show how hate speech annotations done by expert annotators who are

Proceedings of NLPerspectives @ LREC 2026, pages 136–144

activists differ from those done by crowdworkers. Age (Diaz et al., 2018) and gender (Biester et al., 2022) have also been shown to contribute to variance in judgments. Much of this work on subjectivity focuses on analyzing variations in third-party annotators such as crowdworkers, who typically have no direct involvement in the task being studied. Some exceptions include Arora et al. (Arora et al., 2020) – who recruit female journalists on Twitter who have been targets of abuse as annotators for hate speech data, and Patton et al. (Patton et al., 2019) – who show that members from groups discussed in tweets about gang activity annotate psycho-social attributes differently than social work students. In this work, we propose to study subjectivity from the perspective of *stakeholders* as they will be beneficiaries of similar automated systems based on data-driven models and as they tend to generate speech very similar to that observed in our data on a regular basis (i.e., classroom dialog).

To the best of our knowledge, there is no work that discusses annotator subjectivity as applied to collaborative dialog analysis or towards tools for instructional support. However, labels solicited from students through self-reports have been used successfully to model student affect (Broekens and Brinkman, 2013). (Zambrano et al., 2024) examined how supervised classifiers using self-reported labels vary from those using classroom observations, finding that both labels are useful in modeling different components of affect. Moreover, in learning sciences and education research, some work highlights how an ethnographic perspective should be adopted when coding discourse (Hennessy et al., 2020), which i) takes into account the socio-cultural setting where the conversations take place; ii) treats the observer (or annotator) as another source of influence on the "knowledge" that is produced during coding (Haraway, 1988; Gee and Green, 1998). Hennessy et al. (Hennessy et al., 2020) further discuss how coding schemes may be hard to adhere to for a coder who was not involved in the development of the scheme itself. Despite this observation, they discuss the importance of measuring reliability with multiple coders in order to share schemes for general use.

Given how achieving high reliability goes hand-in-hand with a rigorous annotation process where disagreements are discussed, the primary contribution of our study will be an analysis of disagreements resulting from the subjective perspectives of annotators in annotating classroom dialog. Specifically, we study the task of detecting students' *off-task* speech. Off-task behavior during collaborative learning has been the subject of study by learning scientists from multiple perspectives. Some work, such as Sabourin et al. (Sabourin et al., 2011) discusses the negative effect of off-task behavior on learning; more recent work by Langer-Osuna et al. (Langer-Osuna et al., 2018) discusses the impact of off-task speech on equity by serving as a mechanism for marginalized students to make bids for participation when their voices are neglected. Computational modeling of off-task speech has thus focused on both enabling further qualitative analysis (Ganesh et al., 2023) and on detection with the goal of adaptively scaffolding collaborative learning (Carpenter et al., 2020).

Judging whether an utterance is on-task or off-task naturally requires some subjectivity on the part of the labeler. Since data-driven tools for supporting collaborative learning will ultimately be used by the target audience of students and teachers, our main contribution is a pilot study focused on understanding how such potential stakeholders compare in their perceptions of off-task speech. We investigate the research question: *do teachers and students differ in their annotations when instructed to annotate classroom dialog for verbal off-task behaviors, and if so, in what way?* As our second contribution, we also discuss the potential benefits of explicitly representing subjectivity when developing applications for studying or supporting collaborative learning.

2. Methods

2.1. Data

We use the dataset described in Southwell, et al. (Southwell et al., 2022), which was shared with us for research purposes. The dataset consists of five-minute long transcripts of small-group discussions, collected from a middle-school science classroom in the United States. The subject of instruction is a curriculum unit called Sensor Immersion (SI) focused on "programmable sensor technology". Student interactions are recorded through desk-top mics, and manually transcribed and anonymized, yielding 27 transcripts with 1680 student utterances in total. The entire dataset was then labeled by trained annotators for whether or not each utterance is on-task, for the purposes of a separate standalone study on detecting off-task student speech. Each transcript was double-annotated, i.e., labeled by two annotators who have extensive experience in linguistic annotation tasks. The labels provided by the annotators included *on-task*, *off-task*, and *undecidable* given the context. When labeling an utterance, the annotators look at the entire text transcript to obtain contextual information, but due to data protection restrictions, the annotators do not have access to audio or video recordings. The annotators also have access to all the curriculum materials contained in the sensor immersion curriculum unit. Disagreements between annotators were

adjudicated using a third annotator, and inter-rater agreement was 0.647 (as measured by Cohen's kappa), indicating substantial agreement. The resulting labeled dataset is highly skewed, as the natural occurrence of off-task speech is lower than on-task speech, making up only a fourth of the dataset.

For the purpose of the pilot study, we scale down the dataset to four transcripts owing to time and cost constraints. Based on the viability of the pilot, we plan to extend the study on a larger dataset and more tasks in future work. The four transcripts we select have a good representation of both on-task and off-task utterances. We discard any unclear utterances that are transcribed only as `[inaudible]` or `[noise]`. We are left with 399 content-ful utterances in total, and the corresponding label distribution is shown in Figure 1. Any student names in the utterances are redacted, and the transcripts are completely anonymized.

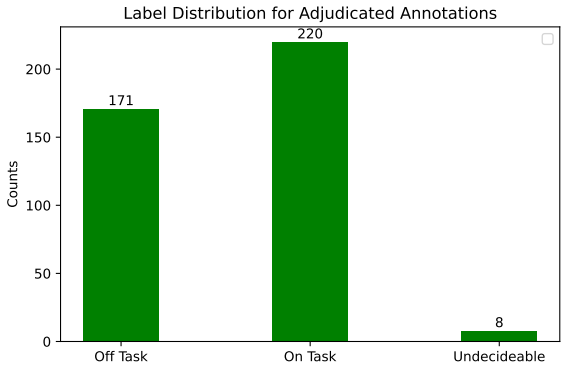


Figure 1: Distribution of the three labels for the on-/off-task classification task for the four transcripts used in the pilot study. The labels are from the adjudicated dataset after double-annotation by experts.

2.2. Participants

As mentioned above, our goal is to study annotations contributed by two types of stakeholders, namely students and teachers. Previous work in NLP that provide datasets with a range of annotations vary in the number of annotators that they include, with many using three (Plank et al., 2014b,a; Arora et al., 2020), some using three to five (Demszky et al., 2020), and some going up to 641 annotators (Sap et al., 2022). Based upon these references, we recruit five annotators each for the *student* and *teacher* group for a total of ten annotators. We share an annotation guideline with the participants that lists examples of on-task and off-task utterances and gives them instructions on how to make use of contextual information to handle ambiguous or edge cases. This annotation guideline is largely the same as that used by the expert annotators; however, unlike the expert annotators who were given access to the entire curriculum, in the interest of time, we provide a brief description of the topics seen in the transcripts and explicitly list keywords like micro:bit, sensor, Makecode, etc., in the annotation guideline.

Student recruitment: Since the conversations that we model are all from middle-school classrooms, we recruit student annotators from a similar age range. We advertise the study in community hubs around Boulder, Colorado and the surrounding areas and on social media. Our five participants are all between 12-14, with four of them from Boulder, and one from the California Bay Area. We collect annotations from each of the participating students through a single 90 minute one-on-one Zoom study. During the study, we share the annotation guideline containing the task description and labels, and ask them to code a small sample of five utterances while explaining their reasoning. We correct any misconceptions at this stage and then collect annotations for all transcripts, framing the task as answering *yes/no/I don't know* to the question *"Is the given utterance on-task?"*. We do not interfere or help the students when they start annotating the transcripts. At the end of the study, each participant is paid $37.5 (at $25 per hour) for their time. We note here that while the participants are students, and represent the perspectives of stakeholders of interest to us, they are not the same students whose speech was originally recorded in the data.

Teacher recruitment: To collect teacher annotations, we use a crowd-sourcing tool called Prolific[1] where domain experts who are verifiably employed in specific professions can be recruited. We recruit middle and high school teachers located in the US, and our five participants are all middle-aged teachers. We share the same annotation guideline that is shown to the students, and use the same five examples as a filter condition, i.e., if a candidate answers those questions incorrectly, their submissions are not accepted. Unlike the study with the students, the teachers submit their annotations entirely offline, although they have the option to contact us if needed. Compensation is again $25/hr per person.

2.3. Evaluation

Our focus at the evaluation stage is to compute agreement metrics that shed some light on whether teachers and students differ noticeably in their responses. We therefore report intra-group agreement, e.g., between all five students, as well as

[1] https://www.prolific.com/

inter-group agreement, e.g., between students and teachers.

In order to compute inter-group agreement, we perform label aggregation at the group level. For every utterance, we take the mode among all labels from a group (five in this case) and use that to represent the final group judgment, e.g., if four teachers answer *on-task*, and one answers *off-task*, the "teacher" label is considered to be on-task. We also use this aggregate to compare against our original expert-annotated labels.

We report percentage agreement for all sets of annotations. Since we have a group of five annotations, we report two percentages: i) *full agreement*, where all five annotators in a group agree on the label and ii) *all-but-one agreement*, where all but one of the annotators agree. The second metric provides some leeway for one of the annotators being a slight outlier, and has been reported in prior work as well (Demszky et al., 2020).

In addition to percentage agreement (which gives some insight into accuracy), we also report the Fleiss kappa to measure agreement between all five annotators in a group (henceforth denoted by κ). The Fleiss κ (Fleiss, 1971) is found to be more suitable than the Cohen's κ when the number of annotators is more than two. κ is from a scale of 0 to 1. The goal of the κ score is to shed light on whether the agreement is due to random chance or if it is due to a reliable overlap between the raters. While the interpretation of the score is always contextual due to the number of labels/raters, we follow the guideline of (Landis and Koch, 1977) to judge whether agreement is poor or good. They suggest that $\kappa > 0.61$ indicates substantial agreement, $0.61 < \kappa > 0.41$ indicates moderate agreement, and $0.41 < \kappa > 0.21$ indicates fair agreement.

Finally, we also compute and report statistics regarding the distribution of labels in each group. For a group, we compute this cumulatively: given five sets of annotations for four transcripts (total length 399), we combine all the labels such that we have a pool of 399 * 5 labels (i.e., 1995). The label distribution that arises from this will tell us about all the group members' judgments.

3. Results

3.1. Intra-Group Agreement

Table 1 shows the agreement between the five teachers and between the five students who participate in our study. Looking at the percentage agreements for teachers, we see that on slightly less than half the utterances, all the group members assign the same label to the utterance. When we look at only whether four of them agree, the agreement jumps to three-quarters of the dataset. The κ score of 0.419 indicates moderate agreement.

The agreement between students is slightly higher, with about half of the utterances receiving complete agreement, and almost 80% of utterances having at least four of the students providing the same annotation. Similar trends are indicated by the numerical value of κ, which, at 0.522 is higher for the student group compared to 0.419 for the teacher group. However, according to the guideline mentioned above, both these κ values indicate moderate agreement.

Table 1 also shows the agreement between experts on the subset of four transcripts used here. The κ value of 0.519 indicates moderate agreement which shows that this specific subset has lower agreement even among the experts than the entire dataset (on which agreement was substantial). However, since the number of expert annotators is fewer than either the teacher or student group, we do not directly compare the rates of expert agreement to the other two groups.

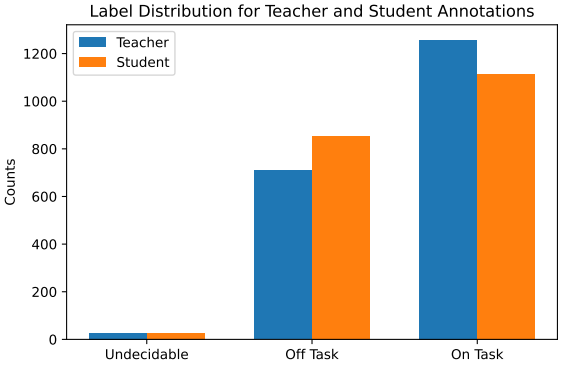


Figure 2: Distribution of labels across all four transcripts for all five annotators, shown for both the teacher group and the student group.

Next, we look at the label distribution on the cumulative labels of all annotators in a group, shown in Figure 2. We see that both groups use the *undecidable* label at equivalent rates. However, the student group uses the *off-task* label much more than the teacher group, indicating that students judge that utterances are *off-task* more often. while teachers may consider them to be *on-task*.

3.2. Inter-Group Agreement

As described in Section 2.3, we report inter-group agreement on the aggregated labels in Table 2. For each utterance, the teacher label represents the majority label out of all five teachers' individual labels, and similarly for the students. The expert label in this case refers to the adjudicated label from the dataset.

We see that the majority label chosen from the student group does have a very high overlap with

Metric	Teacher Agreement	Student Agreement	Expert Agreement
Full agreement	44.61%	51.13%	72.43%
All-but-one agreement	75.94%	79.70%	-
Fleiss κ	0.419	0.522	0.519

Table 1: Percentage agreement and Fleiss κ within the five teachers and within the five students. The last column reports agreement between the two expert annotators on the four transcripts alone.

Group	% Ag.	κ
Teacher–Student	91.73%	0.819
Teacher–Expert	86.72%	0.731
Student–Expert	89.47%	0.795
Teacher–Student–Expert	83.96%	0.782

Table 2: Inter-group agreement between student annotations, teacher annotations, and the expert annotations. We report both percentage agreement and Fleiss κ scores. Results are across every utterance in the dataset.

the majority label chosen from the teacher group – 91.73% of all utterances have the same majority labels, indicating "almost perfect" agreement with a κ value of 0.819 (Landis and Koch, 1977). Looking at how the majority labels from both groups compare to the expert annotator's labels, we note that the students agree with our expert annotators slightly more than the teachers do: the agreement between the student label and the expert label is 89.47% with $\kappa = 0.795$, and the agreement between the teacher label and the expert label is 86.72 with $\kappa = 0.731$. However both these κ values indicate substantial agreement. Overall, when we consider the agreement between the teacher, student and adjudicated label, we get a percentage agreement of 83.96% where $\kappa = 0.782$, once again indicating substantial agreement.

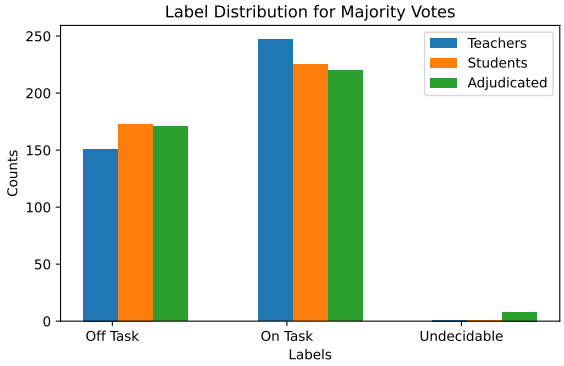


Figure 3: Label distributions of teacher, student and the expert labels. The teacher and student labels are given by the respective majority vote, and the expert label is given by the adjudicated label.

Figure 3 shows the majority label distribution of all three annotator groups, namely the teachers, students, and the expert annotators. We report results on every utterance in the dataset after group-level aggregation: the teacher and student group's label are their respective majority labels, and the expert labels are the adjudicated labels from the dataset. We first observe that the expert annotators resort to using the *undecidable* label more than the teachers or the students. We believe this could be explained by two reasons: i) Since there are only two expert annotators, the adjudicated label could be *undecidable* when only two people choose it. However, for the group annotations, the label could only be *undecidable* if a majority of the five choose it, typically at least three. ii) The expert annotators spend a longer time working with the transcripts, as they annotate 27 transcripts as compared to only four for the student and teacher groups. Thus they may be applying finer-grained judgments when distinguishing between the labels and are more sensitive to the highly ambiguous cases where a decision cannot be made given the information available.

Next, we observe that apart from the *undecidable* label, the label distribution of the student majority label shares more similarities with the expert annotator's label rather than the teacher. In comparison to the expert annotator and the students, the teachers have more of a tendency to label utterances as on-task rather than off-task. This is a surprising result as anecdotal evidence indicated that teachers may be 'stricter' or be more likely to conclude that speech is off-task. Although prior research hasn't investigated subjectivity in judging off-task classroom speech, some research has shown that experienced teachers exhibit more awareness of off-task behavior in classrooms than novice or "student" teachers (Wolff et al., 2017; Shinoda et al., 2021). Based on these findings, if the student can be considered as a less experienced judge, we would again expect the teachers to identify more utterances as off-task than on-task, which is not the case here.

Teacher–Student: On-Task–Off-Task	Teacher–Student: Off-Task–On-Task
"I think she's color blind."	"Oh my god, oh my god. oh my god. oopsie daisy."
"Oh that's just dope. I love that color."	"Dang it."
"Well whose fault is that?"	"Do do do do. Can't touch this. [noise]"

Table 3: Examples of utterances where students and teachers disagree. In the first column, the teacher label is on-task but the student label is off-task. In the second column, the opposite occurs.

3.3. Qualitative Analysis of Teacher/Student Labels

To get a closer understanding of stakeholder's subjectivity, we examine the pattern of disagreements between the majority labels of the teachers and the students. We first reiterate that agreement between teachers and students was almost perfect, with 91% of the utterances being labeled similarly by the teachers and students. The remaining 9% comprises 35 utterances. Of these, 28 of the utterances (80%) are cases where the teacher labels an utterance as on-task and the student labels the utterance as off-task, and only 13 utterances are cases where the teacher label is off-task and the student label is on-task.

Table 3 shows some examples of utterances where disagreements occur. The utterances where teachers judge an utterance as on-task but students do not are instances of students engaging in a minor aside in the middle of a problem-solving interaction, such as admiring the color of an artifact they produced on screen. Interestingly, the comment "I think she's color blind" does not seem to be made rudely; the surrounding context reveals that the student is indeed color blind and needs help distinguishing wires, but a student annotator may have interpreted it to be a disparaging comment, whereas the teacher annotators interpret it differently. The cases where students label utterances as on-task but teachers do not, include minor swear words such as "dang it" as well as effusive or repetitive.

4. Discussion

The research question that we investigated through this study was to investigate if teachers and students differ in their labeling of classroom dialog, and if so, in what ways. The high agreement (κ > 0.8) that we observed when comparing majority labels indicates that there is no statistically significant difference in teacher and student annotations. However, we do find that i) students show higher in-group agreement than teachers ii) students and expert annotators recognize off-task utterances with a higher frequency than teachers. Moreover, the disagreements appear to be *systematic* and not random, since there is a pattern of teachers being more lenient with minor asides, while students strictly judge these utterances as off-task.

Applications of stakeholder perspectives in classrooms The first implication of our study is on the notion of a "ground truth", particularly when the ground truth data comes from technologists who are building the tools and not from the stakeholders who may be using them. In this case, if an automated off-task utterance classifier was to be a part of a teacher-facing real-time learning analytics dashboard to support small-group discussions, the teacher's preferences may differ from the classifier's learned representations, resulting in false positives. This is also the case if an underlying model is expected to serve multiple stakeholders or applications, such as providing analytics to both teachers and qualitative researchers. Collecting a range of annotations that go beyond experts and include contributions from stakeholders could therefore be useful in modeling different preferences for different application.

Implications on reliability and trust The other benefit in using a range of annotations is from a reliability perspective when deploying machine learning models in the high-stakes environments of classrooms. Discriminative classifiers' estimate of the likelihood of each class can be obtained as a probability score, referred to as the model's uncertainty. In designing dashboards or in instructional tools that intervene based on a students' detected state, these probabilities are used to decide a threshold before an action is taken. However, research has shown that machine learning models, particularly, deep neural-network based models are not well-calibrated and do not produce uncertainty estimates that accurately reflect the true likelihood of an event (Ovadia et al., 2019). Typically, this manifests as overconfidence, especially towards the majority label or class. The calibration of models can be improved by using a range of labels (Prabhakaran et al., 2021; Khurana et al.). One mechanism is through multi-annotator models, as demonstrated by Davani et al. (Davani et al., 2022): by treating each individual annotator's data as a separate task, they construct a multi-tasking model that outputs multiple predictions, which can then be aggregated. The resulting uncertainty estimate

is shown to be a better measure of disagreements between annotators than if the labels were aggregated prior to model training.

Through this pilot study, we showed that subjectivity, specifically from the perspective of stakeholders, *is* a factor when coding classroom dialog. We advocate for the collection and release of a wide range of annotator-level labels to facilitate the creation of reliable models that incorporate stakeholder judgments. In the future, we will extend this study to a larger dataset as well as investigate different behaviors, such as CPS skills.

5. Limitations

The central goal of our study is to examine stakeholder subjectivity in annotating classroom dialog. However, despite our pilot study showing some differences in the way students and teachers perceive off-task dialog, the strength of our conclusions is limited by scope as we only look at five students and five teachers. A broader study that includes a larger pool of participants is essential before we can make strong recommendations for the design of learning analytics systems or interventions based on the phenomenon of stakeholder subjectivity. Moreover, for greater ecological validity, both the teachers and students must share context (such as being from the same school) with the students who generate the utterances represented in our data – which is a setting that we were unable to demonstrate in our study due to unavailability of the original classroom participants.

6. Acknowledgments

We thank the reviewers for their time and helpful feedback. We also thank all the student participants and teacher participants who provided their responses on this study. This study was approved by the University of Colorado's Institutional Review Board under protocol #25-0253. This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805 and under grant DRL 1920510. The opinions expressed are those of the authors and do not represent views of the NSF.

7. Bibliographical References

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.

Ishaan Arora, Julia Guo, Sarah Ita Levitan, Susan McGregor, and Julia Hirschberg. 2020. A novel methodology for developing automatic harassment classifiers for Twitter. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 7–15, Online. Association for Computational Linguistics.

Carl Bereiter. 2002. *Education and mind in the Knowledge Age.* Education and mind in the Knowledge Age. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US. Pages: xiii, 526.

Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. 2022. Analyzing the effects of annotator gender across nlp tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 10–19.

Joost Broekens and Willem-Paul Brinkman. 2013. Affectbutton: A method for reliable and valid affective self-report. *International Journal of Human-Computer Studies*, 71(6):641–667.

Dan Carpenter, Andrew Emerson, Bradford W. Mott, Asmalina Saleh, Krista D. Glazewski, Cindy E. Hmelo-Silver, and James C. Lester. 2020. Detecting off-task behavior from student dialogue in game-based collaborative learning. In *Artificial Intelligence in Education*, pages 55–66. Springer.

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.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA. Association for Computing Machinery.

Sidney K D'Mello, Nicholas Duran, Amanda Michaels, and Angela EB Stewart. 2024. Improving collaborative problem-solving skills via automated feedback and scaffolding: a quasi-experimental study with cpscoach 2.0. *User modeling and user-adapted interaction*, 34(4):1087–1125.

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

Ananya Ganesh, Michael Alan Chang, Rachel Dickler, Michael Regan, Jon Cai, Kristin Wright-Bettner, James Pustejovsky, James Martin, Jeff Flanigan, Martha Palmer, et al. 2023. Navigating wanderland: Highlighting off-task discussions in classrooms. In *International Conference on Artificial Intelligence in Education*, pages 727–732. Springer.

James Paul Gee and Judith L Green. 1998. Chapter 4: Discourse analysis, learning, and social practice: A methodological study. *Review of research in education*, 23(1):119–169.

Arthur C Graesser, Samuel Greiff, Matthias Stadler, and Keith T Shubeck. 2020. Collaboration in the 21st century: The theory, assessment, and teaching of collaborative problem solving.

Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3):575–599.

Sara Hennessy, Christine Howe, Neil Mercer, and Maria Vrikki. 2020. Coding classroom dialogue: Methodological considerations for researchers. *Learning, Culture and Social Interaction*, 25:100404.

Urja Khurana, Eric Nalisnick, Antske Fokkens, and Swabha Swayamdipta. Crowd-calibrator: Can annotator disagreement inform calibration in subjective tasks? In *First Conference on Language Modeling*.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. 2024. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Jennifer Langer-Osuna, Emma Gargroetzi, Rosa Chavez, and Jen Munson. 2018. Rethinking loafers: Understanding the productive functions of off-task talk during collaborative mathematics problem-solving. International Society of the Learning Sciences.

Jonathan Osborne. 2010. Arguing to learn in science: The role of collaborative, critical discourse. *science*, 328(5977):463–466.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. *Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift*. Curran Associates Inc., Red Hook, NY, USA.

Desmond Patton, Philipp Blandfort, William Frey, Michael Gaskell, and Svebor Karaman. 2019. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators.

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.

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.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Díaz. 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.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 8–16, Manchester, UK. Coling 2008 Organizing Committee.

Jason G Reitman, Charis Clevenger, Quinton Beck-White, Amanda Howard, Sierra Rose, Jacob Elick, Julianna Harris, Peter Foltz, and Sidney K D'Mello. 2023. A multi-theoretic analysis of collaborative discourse: A step towards ai-facilitated

student collaborations. In *International Conference on Artificial Intelligence in Education*, pages 577–589. Springer.

Jeremy Roschelle. 1992. Learning by collaborating: Convergent conceptual change. *The journal of the learning sciences*, 2(3):235–276.

Carolyn Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International journal of computer-supported collaborative learning*, 3(3):237–271.

Jennifer Sabourin, Jonathan P. Rowe, Bradford W. Mott, and James C. Lester. 2011. When Off-Task is On-Task: The Affective Role of Off-Task Behavior in Narrative-Centered Learning Environments. In *Artificial Intelligence in Education*, pages 534–536, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Hirofumi Shinoda, Tsuyoshi Yamamoto, and Kyoko Imai-Matsumura. 2021. Teachers' visual processing of children's off-task behaviors in class: A comparison between teachers and student teachers. *PLoS One*, 16(11):e0259410.

R. Southwell, S. Pugh, E.M. Perkoff, C. Clevenger, J. Bush, and S. D'Mello. 2022. Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms. In *Proceedings of the 15th International Conference on Educational Data Mining*. International Educational Data Mining Society.

Chen Sun, Valerie J Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D'Mello. 2020. Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143:103672.

Lev S Vygotsky. 1978. *Mind in society: The development of higher psychological processes*, volume 86. Harvard university press.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Charlotte E Wolff, Halszka Jarodzka, and Henny PA Boshuizen. 2017. See and tell: Differences between expert and novice teachers' interpretations of problematic classroom management events. *Teaching and teacher education*, 66:295–308.

Stella Xin Yin, Zhengyuan Liu, Dion Hoe-Lian Goh, Choon Lang Quek, and Nancy F. Chen. 2025. Scaling up collaborative dialogue analysis: An ai-driven approach to understanding dialogue patterns in computational thinking education. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, page 47–57, New York, NY, USA. Association for Computing Machinery.

Andres Felipe Zambrano, Nidhi Nasiar, Jaclyn Ocumpaugh, Alex Goslen, Jiayi Zhang, Jonathan Rowe, Jordan Esiason, Jessica Vandenberg, and Stephen Hutt. 2024. Says who? how different ground truth measures of emotion impact student affective modeling. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 211–223.

Author Index