Proposal: From One-Fit-All to Perspective Aware Modeling

Leixin Zhang
University of Twente
l.zhang-5@utwente.nl

Abstract

Variation in human annotation and human perspectives has drawn increasing attention in natural language processing research. Disagreement observed in data annotation challenges the conventional assumption of a single "ground truth" and uniform models trained on aggregated annotations, which tend to overlook minority viewpoints and individual perspectives. This proposal investigates three directions of perspective-oriented research: First, annotation formats that better capture the granularity and uncertainty of individual judgments; Second, annotation modeling that leverages socio-demographic features to better represent and predict underrepresented or minority perspectives; Third, personalized text generation that tailors outputs to individual users' preferences and communicative styles. The proposed tasks aim to advance natural language processing research towards more faithfully reflecting the diversity of human interpretation, enhancing both inclusiveness and fairness in language technologies.

1 Introduction

Understanding human perspectives and designing systems that cater to individual needs are critical goals in natural language processing (NLP) research. However, traditional approaches often rely on aggregated annotations in datasets and treat them as a singular ground truth for model training (Braylan and Lease, 2020; Qing et al., 2014).

In recent years, the assumption of a "single ground truth" has been increasingly challenged by researchers (Plank, 2022; Cabitza et al., 2023; Sap et al., 2022; Frenda et al., 2024), drawing attention to the limitations of conventional data construction and modeling practices in capturing the full spectrum of human perspectives. Beyond NLP research, similar concerns have arisen in related fields, such as the legal domain (Braun and Matthes, 2024; Xu et al., 2023), the medical domain (Miñarro-Giménez et al., 2018), and music annotation (Koops et al., 2019).

Growing evidence suggests that annotator perspectives are shaped by complex, context-dependent factors, including individual beliefs, their demographic backgrounds, context information, text ambiguity or interpretive uncertainty. Studies (Braun, 2024) also highlighted that human annotators frequently provide different but equally valid labels, challenging the assumption that there is always a single correct answer. This shift calls for a deeper investigation into annotation variation and human perspectives research in all stages: annotation (Plank, 2022), modeling (Uma et al., 2021; Mostafazadeh Davani et al., 2022; Mokhberian et al., 2024) and evaluation frameworks (Basile et al., 2021; Rizzi et al., 2024) in order to improve the inclusiveness and models' alignment of human perspectives.

This proposal aims to advance perspective-aware approaches in NLP by providing insights into annotation methodologies that better capture the complexity of human perspectives and improve modeling efficiency (Section 3), evaluating the influence of socio-demographic factors on annotation variation modeling (Section 4), and exploring methods to leverage persona information for personalized textual generation (Section 5). Three tasks are illustrated in Figure 1.

Annotation Format: This task explores different formats of annotation types in representing perspectives: binary labels vs. continuous or Likert scale values. We assess whether continuous values or Likert scales, rather than binary labels, better capture the uncertainty of annotators' tendencies and perspectives. The research outcome aims to improve annotation practices and derive more refined annotation methods for capturing the subtleties of diverse annotator perspectives.

Perspective Annotation Modeling: This task in-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1016–1025

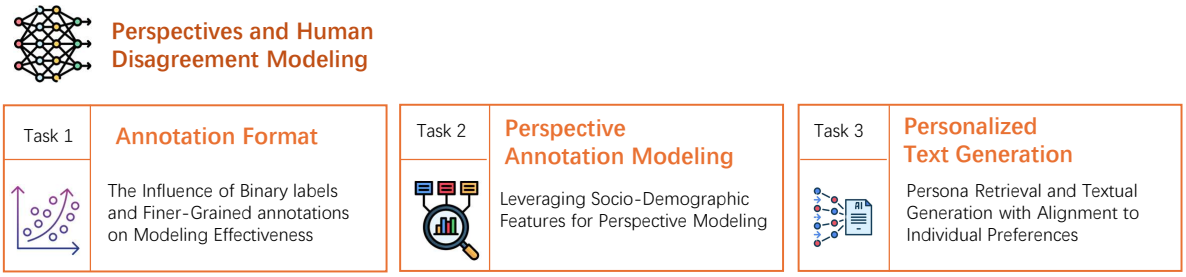

Figure 1: Proposed Tasks of Perspective Aware Modeling

vestigates the extent to which socio-demographic features can account for annotator perspectives or variation in humans' annotation patterns. We examine the effectiveness of predicting an individual's annotations based on their socio-demographic attributes in application domains that have not yet been explored.

Personalized Generation: This task explores persona-based modeling and personalized textual generation that reflect users' preferences and communication styles. We incorporate structured persona information, such as socio-demographic features, sentiment orientation, and linguistic complexity as additional signals for text generation. The objective is to produce responses or texts that are not only contextually appropriate but also tailored in terms of individual preference.

2 Related Studies

Recent studies have increasingly recognized the presence of human disagreement and diverse perspectives in annotation tasks. Various terms have been used to describe this phenomenon, including subjectivity (Reidsma and Carletta, 2008), human uncertainty (Peterson et al., 2019), perspectivism or perspectivist (Cabitza et al., 2023; Frenda et al., 2024), human label variation (Plank, 2022) and pluralism (Sorensen et al.; Feng et al., 2024). Moreover, an increasing number of studies have released datasets (Wang et al., 2023; Kumar et al., 2021; Frenda et al., 2023; Passonneau et al., 2012; Dumitrache et al., 2018) annotated by multiple individuals, in contrast with the single label from the traditional majority-vote aggregation or score averaging.

Prior research (Plank et al., 2014; Sheng et al., 2008; Guan et al., 2018; Fornaciari et al., 2021; Xu et al., 2024; Casola et al., 2023) has demonstrated that incorporating labels from multiple annotators can enhance model performance by improving the model's generalization ability. Methods include the cost-sensitive approach, where the loss of each instance is weighted based on label distribution (Plank et al., 2014; Sheng et al., 2008), as well as soft-loss approaches (Peterson et al., 2019; Lalor et al., 2017; Uma et al., 2020; Fornaciari et al., 2021). Furthermore, researchers have explored leveraging additional metadata, such as socio-demographic features (Goyal et al., 2022; Gordon et al., 2022), annotator IDs (Mokhberian et al., 2024), and partial annotation histories (Milkowski et al., 2021; Sorensen et al., 2025), to characterize individual annotation patterns and refine learning procedures.

The alignment of large language models (LLMs) with human annotation has also gained increasing attention under the context of embracing human disagreement, particularly in evaluating their ability to capture diverse perspectives and which groups' perspective that LLMs reflect (Hu and Collier, 2024; Beck et al., 2024; Salemi et al., 2024; Muscato et al., 2024). In the generation domain, MORPHEUS (Tang et al., 2024) introduces a three-stage framework to model roles from dialogue history. It compresses persona information into a latent codebook, enabling generalization to unseen roles through joint training. Lu et al. (2023) disentangle multi-faceted attributes in the latent space and use a conditional variational auto-encoder to align responses with user traits.

3 Annotation Formats for Perspective Representation

This task explores two different annotation formats (binary classification versus Likert-scale or continuous values) for representing human perspectives and investigates their influence on modeling effectiveness. The study aims to provide guidance for

future dataset construction by identifying annotation formats that best support model learning and more accurately capture the nuance of human perspectives.

3.1 Motivation and Research Hypothesis

Previous research (Plank, 2022; Mostafazadeh Davani et al., 2022) has primarily focused on label variation using discrete labels. Many studies, particularly in domains such as hate speech and offensive language detection, rely on binary annotations (Mostafazadeh Davani et al., 2022; Akhtar et al., 2020). In some cases, ordinal Likert-scale ratings are converted into binary labels in modeling procedures (Orlikowski et al., 2023).

Ovesdotter Alm (2011) argues that acceptability is a more meaningful concept than rigid "right" or "wrong" labels. Human annotators exhibit varying degrees of uncertainty for specific items, and some tasks inherently involve continuous variation, such as the level of emotional arousal (Lee et al., 2022). Simple binary classes can obscure important nuances in annotation data. It may risk oversimplifying the granularity of human perspectives, ultimately impacting model reliability and the interpretability of annotator uncertainty.

We hypothesize that continuous values or Likert scales provide a more effective source for capturing and modeling annotation variation. From the perspective of machine learning, incorporating finer-grained annotations may help align better with human judgment and enhance model performance by smoothing the decision boundary compared to rigid binary labels.

3.2 Methodology

This study undertakes interdisciplinary approach to investigate the impact of the annotation format across multiple domains, including tasks such as hate speech detection, offensive language detection and sentiment analysis[1]. By examining diverse datasets and modeling techniques, we aim to assess whether adopting finer-grained annotation scales improves the representation and learning of annotators' perspectives in a cross-domain context.

Data Construction: Two types of datasets will be used for this purpose. First, for datasets with Likert scales or continuous values, we will train models using the original values and also targets that are transformed into binary labels[2] for comparison. Second, for datasets originally with discrete labels, such as natural language inference, where three labels (entailment, contradiction, and neutrality) exist, we will annotate with an additional scale representing human uncertainty of the label selection to capture the complexity inherent in human judgment.

Modeling framework: To test the hypothesis (numerical values better represent human perspectives than binary labels, and models based on values show better effectiveness in machine learning), we will implement the three modeling architectures (Figure 2) from Mostafazadeh Davani et al. (2022) to compare the results of two types of targets (binary encoding vs. continuous values):

- Individual Annotator Modeling: Each annotator's annotations will be modeled separately using distinct neural networks to capture individual perspectives.
- Multi-target Methods: A shared neural network will be trained with all annotators' annotations represented as target vectors, allowing the model to learn patterns across annotators.
- Multi-Task Learning: A partially shared neural network will be employed, with shared layers capturing common understanding and annotator-specific layers or heads capturing individualized annotation tendencies.

Evaluation and Result Analysis: Model performance will be evaluated using both traditional metrics based on aggregated labels, label distributions and specialized evaluation on individualized prediction accuracy to assess the advantages of finer-grained annotations compared to binary labels. Since direct comparison between binary classification and regression outputs is inherently challenging, we propose two complementary evaluation strategies to facilitate a meaningful comparison:

- Binary Label Conversion: Continuous regression outputs will be converted into binary labels using a predefined threshold (consistent with the threshold used during training for label derivation). We will then compute standard classification metrics such as F1 score

[1]These tasks are known that human annotation variation exists and with relatively richer datasets annotated by multiple individuals, seen Wang et al. (2023); Akhtar et al. (2020); Waseem (2016) and Gruber et al. (2024).

[2]Different threshold values can be set for partition to assess the robustness.

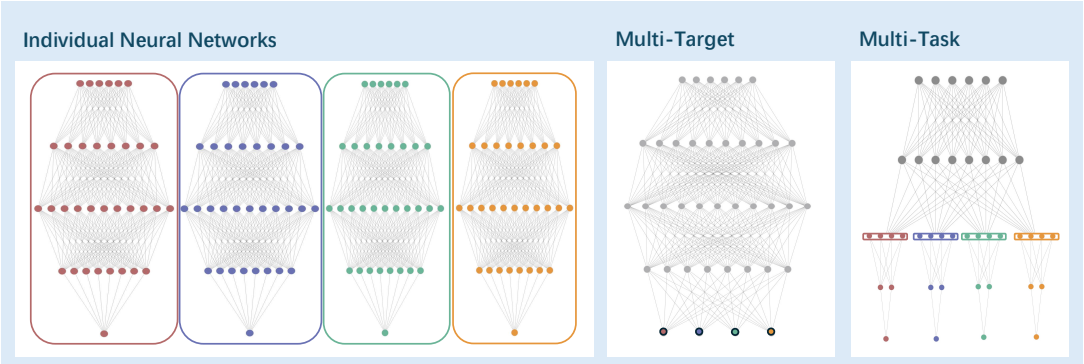

Figure 2: Neural Network Architectures for Perspective Annotation Modeling

and accuracy to evaluate the alignment between the binarized predictions and the target.

- Ranked Correlation Comparison: While classifier outputs do not offer the same level of granularity as regression values, the predicted probabilities or logits can serve as proxies for prediction confidence or intensity (e.g., degree of toxicity). These values enable a ranking-based comparison with the ground truth labels. We will compute the Spearman rank correlation (r) between the model predictions and the true target values, allowing us to compare the correlation strength across both classifiers and regressors.

4 Perspective Annotation Modeling with Demographic Features

This task investigates the extent to which socio-demographic features, such as age, gender, education level, political affiliation, and domain expertise contribute to explaining and modeling variation in human annotation.

4.1 Motivation and Research Questions

While prior research has explored this question in some NLP tasks, findings remain inconclusive with various methods and datasets. In toxicity classification, for example, Orlikowski et al. (2023) reports that incorporating group-level socio-demographic features does not significantly improve predictive performance in toxicity classification tasks, when compared to randomly assigned groups. In contrast, Gordon et al. (2022) discovered a correlation between annotator perspectives and their socio-demographic backgrounds, suggesting these features may meaningfully inform model learning of toxicity.

These conflicting results raise a question: in which application domains and with what modeling methods do socio-demographic features act effectively for modeling? Can we model the probability conditioned on socio-demographic features `Prob(annotation_pred|demographic_feature)` with a better accuracy than assuming an undifferentiated perspective `Prob(annotation_pred)` with neural networks?

We aim to explore whether socio-demographic traits enhance the performance of predicting annotations, particularly in domains that have received limited attention in previous research. Prior research primarily focuses on subjective domains such as hate speech (Sachdeva et al., 2022; Kocoń et al., 2021) or toxicity classification (Goyal et al., 2022). In linguistic annotations, more objective tasks such as natural language inference (Huang and Yang, 2023; Jiang and de Marneffe, 2022) and part-of-speech tag (POS) (Plank et al., 2014) are detected with inherent human label variations.

Extending beyond tasks that received much attention in previous research, we apply this perspective modeling framework to financial or economic domains to investigate the interpretation variation of business trends and sentiment of economic statements[3] (Malo et al., 2014; Liu et al., 2023).

Specifically, we address the following research questions: First, to what extent do socio-demographic attributes and domain expertise account for variation in annotator judgments in business-related tasks? Second, which specific attributes, if any, serve as reliable predictors of annotation variation? And third, which modeling

[3]Related datasets such as Malo et al. (2014) and Liu et al. (2023) are available with a single annotator's decision. Datasets with meta information, particularly with various socio-demographic backgrounds, should be constructed for the purpose of the current study.

methods show advantages in modeling patterns of various socio-demographic groups?

4.2 Methodology

In this task, we will improve the modeling methods in prior research to model socio-demographic features and annotation variation more efficiently. The following modeling methods are proposed for exploration:

- Socio-Demographic Embedding Learning: Embedding layers will be incorporated into neural networks to encode socio-demographic attributes, enabling the model to capture correlations and patterns of annotator attributes such as gender, nationality, and political orientation. This embedding-based model will be compared against a baseline where these attributes are randomly shuffled to assess their genuine contribution to model performance.
- Demographics-Enriched Prompts in Large Language Models (LLMs): We will experiment with prompt-based approaches to incorporate socio-demographic features into LLM predictions. Specifically, we will present demographic features in prompts with either structured key-value formats or natural language descriptions for a comparison study.
- Lightweight Fine-Tuning of LLMs: To further enhance performance, this study will adopt parameter-efficient fine-tuning techniques such as prefix tuning (Li and Liang, 2021), the methods enable personalization without extensive retraining, making them suitable for incorporating socio-demographic signals.

To assess the effectiveness of the proposed methods for modeling human perspectives, we design comparative experiments to assess the effect of socio-demographic features. Specifically, we consider the following three experimental conditions: (1) Single annotation modeling, which only makes use of the aggregated annotations obtained from multiple annotators. (2) Annotation distribution modeling that leverages the distribution of annotations without additional annotator attributes. Methods in Section 3 or approaches such as soft-loss function (Fornaciari et al., 2021; Uma et al., 2021) can serve for this purpose. (3) Socio-demographic enriched learning with three proposed methods in this section, in which predictions are conditioned on socio-demographic features. This comparison will shed light on whether demographic factors serve as useful input features for the perspective modeling of financial trends perception.

4.3 Evaluation

In the evaluation stage, we consider multiple metrics under different conditions. These include (1) Accuracy and F1 score computed from aggregated labels; (2) Measures that capture the distributional alignment of prediction and annotation, metrics including cross-entropy loss, Kullback-Leibler (KL) divergence, and Jensen-Shannon divergence. While, this study mainly focuses on (3) Model performance within specific socio-demographic groups to evaluate its effectiveness across diverse populations. To examine the influence of particular socio-demographic features on perspective attribution, we will apply statistical tests, specifically, the Student's t-test for binary features and ANOVA for categorical features, to investigate correlations between these attributes and annotation behaviors or perspectives.

5 Personalized Text Generation

Building on the perspective exploration of annotation variation, namely **label and value prediction** in the previous tasks, this section extends the research to **personalized text generation**. The goal is to generate language that aligns with individual users' backgrounds, preferences, and communication styles. This includes conditioning generation on persona-related factors such as socio-demographic attributes, historical dialogue context, and language preferences. Personalized generation aims to adapt to user needs and enhance user engagement and satisfaction.

5.1 Motivation

Generative models have demonstrated impressive capabilities of text generation across a wide range of tasks, such as summarization (Wang and Cardie, 2013), question answering (Duan et al., 2017), or dialogue generation (Li et al., 2017). While models may excel at producing coherent texts in a more general setting, they lack the ability to adapt output text to the various profiles of individual users (Zhang et al., 2024). Personalized generation aims to address this problem by integrating user-specific data, such as stated preferences, topic familiarity,

language proficiency or cultural background, to dynamically shape the generated content. This focus on personalization unlocks potential across applications like adaptive education, health support, and personalized suggestions, such as a diet plan or career recommendations.

5.2 Methodology

To achieve the goal of personalized generation, we proposed a two-stage framework: (1) Persona Retrieval and Representation; and (2) Generation with Alignment to Individual Preferences.

In the first stage, persona information can be composed of both **explicit** and **implicit** sources. Explicit features include annotator metadata such as age, gender, education level, and profession, which were collected during the dataset construction phase. Implicit cues, on the other hand, are derived from users' historical text, such as writing style, expressed interests or behaviors. These require a preliminary persona prediction or persona representation. Two strategies will be pursued for persona representation: (1) Structured persona representation, where retrieved information is formatted as key-value pairs and provided as additional context in the input prompts. (2) Latent persona embedding, building on approaches like MORPHEUS (Tang et al., 2024) and MIRACLE (Lu et al., 2023), which encode user attributes into latent vectors. These embeddings can then serve as conditioning signals during the generation phase, enabling fine-grained personalization.

In the second stage, we focus on aligning the language model's generation behavior with the identified user preferences and persona attributes. Two methodologies will be explored:

- Prompt-Based Personalization: Persona attributes will be incorporated into structured or natural language prompts to gauge the generation task with an explicit user role. This approach leverages the in-context learning capabilities of large language models (LLMs) and offers a transparent, controllable mechanism for personalized input.

- Latent Representation Learning and LLM Fine-tuning: To enable integration of personalization signals into neural networks, we will investigate lightweight fine-tuning techniques such as prefix tuning (Li and Liang, 2021), LoRA (Low-Rank Adaptation, Hu et al., 2022). These methods allow LLMs to condition on user-specific embeddings with minimal training and data requirements. Beyond model tuning, this stage may also include reinforcement learning with user feedback (RLHF) or preference modeling, where iterative refinement is guided by explicit or implicit user evaluations.

5.3 Evaluation

Evaluating personalized generation poses additional challenges besides the conventional evaluation of text generation quality. Multiple evaluation strategies will be adopted to assess generation performance: (1) Standard Generation Metrics: Including BLEU, ROUGE and METEOR to assess content quality, coherence, and relevance. While these metrics may not capture personalized generation, they are useful for verifying baseline generation quality. (2) Persona-Based Metrics: We will evaluate the alignment between generated outputs and persona information by measuring the overlap or differences between generated texts and persona sentences in datasets like PersonaChat (Jandaghi et al., 2023). To assess whether generated texts reflect target attributes, we will use classification or clustering-based evaluations, measuring whether the generated texts reflect certain persona attributes. (3) Human Evaluation: For a subset of outputs, human annotators will be used to rate the relevance, fluency, and personalization of responses with respect to their persona profiles.

6 Conclusion

This proposal advances perspective-aware modeling in natural language processing by addressing three key components: annotation format design, annotation variation modeling by leveraging socio-demographic features, and personalized text generation. First, it investigates how finer-grained annotation formats, such as Likert scales, better capture the nuances of human perspectives compared to binary labels. Second, it examines the extent to which socio-demographic features influence annotation variation, particularly in relatively underexplored domains of business and economics. Finally, methods for personalized generation that align output with user-specific attributes are proposed. These tasks aim to enhance the inclusivity and fairness of NLP systems by modeling the diversity of human perspectives.

Limitations

This proposal does not aim to comprehensively resolve all challenges associated with human annotation variation and annotator perspectives, particularly given its cross-domain property. In addition, the availability of suitable datasets for certain tasks, especially those that include detailed annotator background information required for certain modeling and generation tasks, poses challenges to this research. To address this, the study will involve the construction of new datasets or the design of additional annotation tasks tailored to perspective research.

Ethical Considerations

Research involving socio-demographic attributes and personal perspectives inherently carries ethical risks, particularly concerning the privacy and potential misuse of annotators' personal information. This study will take careful measures to protect the identities and privacy of all participants. All collected and analyzed data will be fully anonymized and handled in accordance with privacy-preserving protocols.

Special attention will be given to the ethical challenges of persona inference and demographic modeling. Minority and underrepresented viewpoints, which are essential to the study's objectives, will be treated with care and used solely for academic purposes to prevent any harm or stigmatization. Moreover, in the analysis and presentation of findings, efforts will be made to use neutral, respectful language and to avoid reinforcing stereotypes or generalizations associated with specific demographic groups.

References

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.

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

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.

Daniel Braun. 2024. I beg to differ: how disagreement is handled in the annotation of legal machine learning data sets. *Artificial intelligence and law*, 32(3):839–862.

Daniel Braun and Florian Matthes. 2024. Agb-de: A corpus for the automated legal assessment of clauses in german consumer contracts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10389–10405.

Alexander Braylan and Matthew Lease. 2020. Modeling and aggregation of complex annotations via annotation distances. In *Proceedings of The Web Conference 2020*, pages 1807–1818.

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

Silvia Casola, SODA Lo, Valerio Basile, Simona Frenda, Alessandra Cignarella, Viviana Patti, Cristina Bosco, and 1 others. 2023. Confidence-based ensembling of perspective-aware models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3496–3507. Houda Bouamor, Juan Pino, Kalika Bali.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark. Association for Computational Linguistics.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing ambiguity in crowdsourcing frame disambiguation. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6, pages 12–20.

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

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, Massimo Poesio, and 1 others. 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.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, pages 1–28.

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

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.

Nitesh Goyal, Ian D Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–28.

Cornelia Gruber, Katharina Hechinger, Matthias Assenmacher, Göran Kauermann, and Barbara Plank. 2024. More labels or cases? assessing label variation in natural language inference. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 22–32.

Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. 2018. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

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

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

Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2023. Faithful persona-based conversational dataset generation with large language models. *Preprint*, arXiv:2312.10007.

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.

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.

Hendrik Vincent Koops, W Bas De Haas, John Ashley Burgoyne, Jeroen Bransen, Anna Kent-Muller, and Anja Volk. 2019. Annotator subjectivity in harmony annotations of popular music. *Journal of New Music Research*, 48(3):232–252.

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.

John P Lalor, Hao Wu, and Hong Yu. 2017. Soft label memorization-generalization for natural language inference. *arXiv preprint arXiv:1702.08563*.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese emobank: Building valence-arousal resources for dimensional sentiment analysis. *Transactions on Asian and Low-Resource Language Information Processing*, 21(4):1–18.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169, Copenhagen, Denmark. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Jun Liu, Kai Wu, and Ming Zhou. 2023. News tone, investor sentiment, and liquidity premium. *International Review of Economics & Finance*, 84:167–181.

Zhenyi Lu, Wei Wei, Xiaoye Qu, Xian-Ling Mao, Dangyang Chen, and Jixiong Chen. 2023. Miracle: Towards personalized dialogue generation with latent-space multiple personal attribute control. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5933–5957, Singapore. Association for Computational Linguistics.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Piotr Milkowski, Marcin Gruza, Kamil Kanclerz, Przemyslaw Kazienko, Damian Grimling, and Jan Kocon. 2021. Personal bias in prediction of emotions elicited by textual opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 248–259, Online. Association for Computational Linguistics.

José Antonio Miñarro-Giménez, Catalina Martínez-Costa, Daniel Karlsson, Stefan Schulz, and Kirstine Rosenbeck Gøeg. 2018. Qualitative analysis of manual annotations of clinical text with snomed ct. *Plos one*, 13(12):e0209547.

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.

Benedetta Muscato, Chandana Sree Mala, Marta Marchiori Manerba, Gizem Gezici, Fosca Giannotti, and 1 others. 2024. An overview of recent approaches to enable diversity in large language models through aligning with human perspectives. In *In Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024*, pages 49–55. European Language Resources Association (ELRA).

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.

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.

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

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9617–9626.

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.

Ciyang Qing, Ulle Endriss, Raquel Fernández, and Justin Kruger. 2014. Empirical analysis of aggregation methods for collective annotation.

Dennis Reidsma and Jean Carletta. 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: an assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024*, pages 84–94.

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.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. 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.

Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? improving data quality

and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622.

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

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, and 1 others. Position: A roadmap to pluralistic alignment. In *Forty-first International Conference on Machine Learning*.

Yihong Tang, Bo Wang, Dongming Zhao, Jinxiaojia Jinxiaojia, Zhangjijun Zhangjijun, Ruifang He, and Yuexian Hou. 2024. Morpheus: Modeling role from personalized dialogue history by exploring and utilizing latent space. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7664–7676.

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. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Lu Wang and Claire Cardie. 2013. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405.

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.

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, Austin, Texas. 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.

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.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, and 1 others. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.