An Overview of Recent Approaches to Enable Diversity in Large Language Models through Aligning with Human Perspectives

Benedetta Muscato[1,2] **Chandana Sree Mala**[1,2] **Marta Marchiori Manerba**[2]
Gizem Gezici[1] **Fosca Giannotti**[1]
[1] Scuola Normale Superiore, [2] University of Pisa
[1]{name.surname}@sns.it, [2]{name.surname}@phd.unipi.it

Abstract

The varied backgrounds and experiences of human annotators inject different opinions and potential biases into the data, inevitably leading to disagreements. Yet, traditional aggregation methods fail to capture individual judgments since they rely on the notion of a single ground truth. Our aim is to review prior contributions to pinpoint the shortcomings that might cause stereotypical content generation. As a preliminary study, our purpose is to investigate state-of-the-art approaches, primarily focusing on the following two research directions. First, we investigate how adding subjectivity aspects to LLMs might guarantee diversity. We then look into the alignment between humans and LLMs and discuss how to measure it. Considering existing gaps, our review explores possible methods to mitigate the perpetuation of biases targeting specific communities. However, we recognize the potential risk of disseminating sensitive information due to the utilization of socio-demographic data in the training process. These considerations underscore the inclusion of diverse perspectives while taking into account the critical importance of implementing robust safeguards to protect individuals' privacy and prevent the inadvertent propagation of sensitive information.

Keywords: Text Generation, Perspectivism, Human Annotation, Bias, Diversity, Minority Groups

1. Introduction

Large Language Models (LLMs) have revolutionized NLP field by making it possible to generate human-like content. Nowadays, LLMs are competent in a wide range of downstream tasks. Human involvement, particularly concerning the data input, is responsible for the significant variance in the model results. Therefore, it is crucial to look at the training process of these models to comprehend how and why they generate biased information as well as the underlying resources they rely on. For instance, it is well-known that human annotators may introduce biases in annotations from their personal opinions or beliefs due to their distinct backgrounds in the context of supervised learning settings, which require labeled data (Romberg, 2022; Soni et al., 2024).

Perspectivism, a new current within the NLP community, advocates for the usage of datasets that gather diverse human judgments on subjective tasks such as stance identification, hate speech detection, and argumentation mining (Röttger et al., 2021). This approach embraces the annotator's disagreement, expressed through differences in annotations, which may result from ambiguity, uncertainty, genuine disagreement, or the lack of a single right answer (Plank, 2022). Moreover, perspectivism overcomes the constraints of traditional aggregation techniques, such as majority voting, which oversimplify real-world intricacies by assuming a single ground truth (Basile et al., 2021; Kanclerz et al., 2022; Mokhberian et al., 2023).

Basile (2020) and Uma et al. (2021) explore the improvement of models while trained on disaggregated datasets with multiple annotations via the development of more accurate and inclusive measures for model decisions. Likewise, Marchal et al. (2022) investigate new evaluations for data with multiple labels to enable new models to learn from fewer but valuable sources.

According to Sap et al. (2021), disagreement is common in subjective tasks and can vary depending on the identity and beliefs of the annotators. In supervised learning tasks as well as in the context of generative AI, especially LLMs, which seek to reflect human language diversity, the unreliability of a unique ground truth becomes a critical factor.

In the context of this paper, our primary goal is to demonstrate how crucial it is to give LLMs the ability to customize their outputs for distinct socio-demographic groups. First, we ask if LLMs can guarantee diversity in the perspectives they generate and why incorporating human annotations representing various viewpoints is essential. Second, by aligning LLMs with humans and using current techniques to evaluate this alignment, we investigate the possibility of fostering diversity. By tackling these issues, we hope to prevent the perpetuation of prejudices against particular communities and promote the creation of more inclusive LLMs that take into account a variety of viewpoints, including those of minority groups. Although individual studies have been carried out on these topics, to the best of our knowledge, this is the first attempt to provide an overview of the subject by adopting this particular angle.

The paper is structured as follows. In Section 2

NLPerspectives 2024@LREC-COLING 2024, pages 49–55

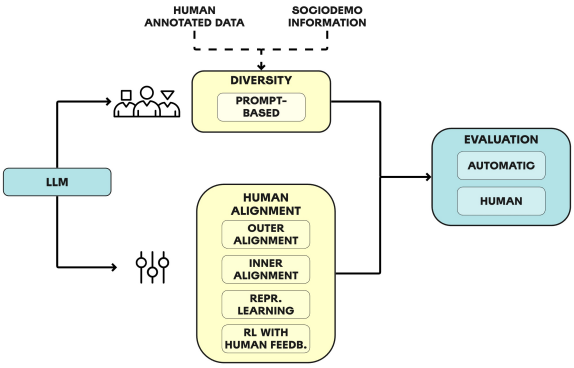

Figure 1: Outline of the topics of the paper.

we briefly present necessary background knowledge on LLMs' perspectives. The notion of diversity is examined in Section 3, while the theme of alignment is discussed in Section 4. The evaluation techniques are finally reported in Section 5. In Fig. 1, we display a diagram summarizing the topics tackled in the review.

2. Background

Recent researches concentrate on the viewpoints that LLMs embed. Kovač et al. (2023) claim that people tend to erroneously anthropomorphize LLMs by assigning certain values, personality traits, knowledge, and abilities to them. Since the context has a significant impact on LLMs' values and personality traits, the study suggests viewing LLMs as a superposition of perspectives. Because of their context-based role-playing, this may imply that LLMs are unreliable in generating diverse viewpoints that are consistent with particular human behaviors (Shanahan et al., 2023).

Some argue that LLMs are *neutral* in certain contexts, while others talk of Personalized Language Models, which can mimic people by imitating their past linguistic patterns (King and Cook, 2020). Especially in situations with limited data resources, Soni et al. (2024) recommend combining both individual and group-based features to capture an individual's identity. They acknowledge the notion that unique characteristics and group membership influence an individual's identity.

Based on the aforementioned studies, LLMs are capable of encompassing diverse viewpoints. Nevertheless, due to their significant contextual dependency, LLMs might be prone to instability over time, despite their best efforts to capture and demonstrate these differentiations, e.g., using a diversified vocabulary and personal values. We must consider how input data can shape models. Given the substantial impact on the model's outputs, it is critical to guarantee the veracity of the data and that they represent a wide range of perspectives, i.e., diversity should not be compromised by data aggregation.

3. Diversity in LLMs

Traditional aggregation approaches have a tendency to neglect the subjectivity and complexity of many tasks for the sake of seeking a single ground truth. Opposing viewpoints may naturally arise in the context of studies that require annotation of controversial topics like politics and religion, due to the subjective nature of the task. For instance, Gezici et al. (2021) illustrate the effect of annotator disagreement by querying search engines, resulting in low inter-rater agreement among crowd-workers, on controversial topics such as abortion, gay marriage, and medical marijuana.

Employing traditional aggregation methods to condense labels into a singular ground-truth label poses challenges, particularly when training black-box models. This issue becomes even more pronounced when models' learning processes feature limited transparency. One possible approach would be to gather human annotations and incorporate socio-demographic details such as gender, age, and levels of education. Even if studies have demonstrated that socio-demographic information improves LLM performance (Wan et al., 2023), one must take into account concerns about the collecting of private and sensitive information. In order to make LLMs more fair and inclusive, prior research has shown that there are valid reasons to explore the possibility of incorporating diversity to these models. Joshi et al. (2020) highlights the apparent bias of the NLP field in favor of Western perspectives, which may be viewed as a significant gap that requires attention. This argument is supported by the fact that LLMs frequently display a biased viewpoint, exhibiting a tendency towards the left and neglecting particular socio-demographic groups (Santurkar et al., 2023).

3.1. Diversity Ensuring Strategies

Criteria-based Prompting Hayati et al. (2023) introduce *maximum diversity extraction* from LLMs, an approach proposed to promote differentiation. Their objective is to investigate LLM's ability to generate diverse perspectives and justifications for subjective tasks. In other words, the researchers analyze the differences between the opinions produced automatically by LLMs and those of humans. Authors first train LLMs using human data, specifically opinions on a given statement where humans can either agree or disagree. Subsequently, the LLM is prompted to generate a variety of stances, both in agreement and disagreement with the statement, while providing reasons for each stance (Table 1).

Prompt
Given a text, how would a person of gender 'Female', race 'White', age '25 - 34', education level 'Master's degree' and political affiliation 'Liberal' rate the degree of toxicity in the text. Possible values are 'not toxic', 'slightly toxic', 'moderately toxic', 'very toxic' or 'extremely toxic'. Text: 'Well when you have a welfare state that propagates an underclass of unskilled parasites' Toxicity:
Statement: It's okay to have privacy Tell me opinions about the statement as many as possible from different people with, "Agree" or "Disagree," one-word or one-phrase criteria that is important for their opinions, and explain how they have different opinions Output: {1:{"Stance" :"Agree", "Criteria": ["personal boundaries", "autonomy"], "Reason": "Having privacy allows individuals to establish personal boundaries and maintain." their autonomy."}, 2: {"Stance": "Disagree", "Criteria": ["transparency", "trust"], "Reason": "Lack of privacy can promote transparency and build trust in relationships." ... 10: {"Stance": "...", "Criteria": [...], "Reason": " ..."}} ... Statement: You're expected to do what you are told Tell me opinions about the statement as many as possible from different people with, "Agree" or "Disagree," one-word or one-phrase criteria that is important for their opinions, and explain how they have different opinions Output:

Table 1: Examples of prompting formulation from Beck et al. (2023) and Hayati et al. (2023), respectively.

Then, the LLM extracts certain *criteria-words*, which are essentially framing keywords used to explain the model's generation process. Following that, the LLM is prompted iteratively in one-shot and few-shot learning settings, with the inputs of an initial statement and several opinions expressing agreement or disagreement concerning the given statement, first with criteria words and then without. Lastly, the opinions generated by humans and LLMs are compared and it has been revealed that human opinions are slightly more diverse than those of LLMs.

The aforementioned methodology seems promising in terms of prompting LLMs with diverse statements and asking them for the generation of new opinions using keywords that may facilitate the generation of various perspectives. The present analysis, however, does not specify whether the perspectives generated by machines and humans are representative of specific people or groups; instead, it only compares perspective generation of humans and machines. Consequently, rather than fostering diversity amongst diverse perspectives, the approach may seem to neutralize them.

Socio-demographic Prompting Beck et al. (2023) claim that varied backgrounds are associated with a higher level of disagreement, highlighting the need for the model to consider a variety of socio-demographic information to generate predictions that are socially aware. Initially, the sociodemographic details of each individual's profile — such as gender, race, level of education, and political affiliation — are provided. Subsequently, the LLM is prompted with and without sociodemographic information to obtain different perspectives. In their research, Beck et al. (2023) assess various types of LLMs with socio-demographic profiles across several datasets for NLP classification tasks including sentiment analysis, hate speech detection and toxicity detection. For instance, the toxicity detection task has been designed as follows. The LLM is prompted to ask how a person with specific characteristics (e.g. female, brown, aged 25-35 with a master's degree, liberal) would rate the toxicity level of the given text. The prompt also contains the possible labels (answers) of the given text in the context of toxicity detection. After the prompting, predictions from different profiles have been collected and further aggregated via majority voting. The goal is to compare the predictions made with and without sociodemographic information.

It has previously been argued that sociodemographic prompting may bias prompt-based algorithms to focus on certain human group annotations while ignoring others that are underrepresented in the data. Nonetheless, sociodemographic prompting has also been criticized for potentially introducing stereotypical biases, which can perpetuate negative generalizations about particular social groups (Blodgett et al., 2020; Cheng et al., 2023; Deshpande et al., 2023). Still, in some cases the strategy seems to be effective, showing improvement in zero-shot performances. However, it did not surpass the effectiveness of standard prompting when directly modeling the original annotator's sociodemographics. The effectiveness of the models varied based on factors such as size, input length, and prompt formulation. About aligning with a person's profile, this study appears to neglect preserving the subjectivity of each profile. Initially, the researchers incorporate personal data into the prompt formulation, but then, after collecting the output, they aggregate each piece of information, thereby nullifying diversity. This results in the final

goal being reduced to only comparing predictions with socio-demographic data and without.

4. LLMs Alignment

An ideal NLP model should consider a broad spectrum of perspectives, avoiding bias towards a singular viewpoint. Ouyang et al. (2022a) define *Alignment Learning* as the process of aligning the behaviors of models with human values like safety and truthfulness, while accurately adhering to the intentions of users. Especially with LLMs, producing text that is in line with human opinions could be crucial for generating and spreading more representative texts in society.

Despite their notable performance, these models are prone to certain limitations such as misunderstanding human instructions, generating potentially biased content, or factually incorrect (hallucinated) information. Acknowledging these shortcomings, the research community's focus has shifted towards aligning LLMs with human perspectives, aiming to enable models to meet user desiderata effectively.

4.1. Approaches to Align LLMs with Human Perspectives

Shen et al. (2023) identify *inner alignment* and *outer alignment* as key research agendas in AI alignment. *Inner alignment* ensures that systems are actually trained to achieve the goals set by their designers. For an in-depth overview of current inner alignment strategies, we refer to the work of Shen et al. (2023). *Outer alignment* involves selecting appropriate loss or reward functions to ensure that AI systems' training objectives align with human values.

According to Shen et al. (2023), approaches like fine-tuning and prompting, reward modeling, human-in-the-loop approaches, and adversarial training are often considered and employed in combination to address the outer alignment of LLMs with human perspectives. Outer alignment methods with Reinforcement Learning from Human Feedback (RLHF) are currently the most commonly used methods (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022b). Instead of an agent receiving feedback from a pre-defined reward function or an environment, the reward is inferred from human preferences and then used for tuning LLMs: the model, therefore, learns from direct feedback provided by users or experts. Several challenges persist in the application of RLHF. Firstly, RLHF may be susceptible to instability during fine-tuning and presents challenges in implementation (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022b). Secondly, it is hard to guarantee that the model acquires suitable behaviors through this feedback. Lastly, there is a need to develop algorithms proficient in seamlessly integrating human feedback into the learning process. While human feedback is invaluable for creating high-performing models, there are instances where complex tasks present challenges to gather this feedback, potentially leading to biases.

In line with prior research on outer alignment to steer LLMs with human perspectives, Dong et al. (2023) presents a novel framework named Reward RAnked FineTuning (RAFT), aiming to align generative models efficiently. In RAFT, generative models undergo fine-tuning using Reinforcement Learning (RL), which uses human preferences as a reward signal to fine-tune the models. Similarly, Glaese et al. (2022) employ reinforcement learning with human feedback to train their models, integrating two new components aimed at aiding human raters in evaluating agent behavior. Liu et al. (2023) propose a novel approach, denoted as Representation Alignment from Human Feedback (RAHF), which proves to be effective and computationally efficient. Extensive experiments demonstrate the efficacy of RAHF is not only in capturing, but also in manipulating representations to align with a broad spectrum of human preferences or values. RAHF's versatility in accommodating diverse human preferences shows its potential for advancing LLMs performance in adherence to human values.

5. Evaluation

Automatic Evaluation Metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are commonly adopted to assess the performance of LLMs across several datasets, especially in machine translation tasks. As LLMs' capabilities grow, their powerful generative ability can serve not only as *test takers* but also as potential *examiners* to evaluate other LLMs.

Santurkar et al. (2023) evaluate the LLMs' alignment with humans w.r.t. *representativeness* and *steerability* dimensions. The *representativeness* has been examined by comparing the default opinion distribution of LLMs with that of the US population as well as with specific demographics. *Steerability* tests models' ability to adapt to a particular demographic group represented by the data. Authors expose how, generally, LLMs trained solely on internet data, tend to align predominantly with Moderate, Protestant, and Catholic demographics, largely because of available training data. The finding underscores the propensity of LLMs to oversimplify different perspectives exposed to specific values and cultures, ignoring minority ones.

In the experiments by Beck et al. (2023), results have been evaluated through using both soft and hard-labels, the latter involving majority voting on predictions obtained via sociodemographic prompt-

ing. Notably, socio-demographic prompting has a more positive impact on soft-label evaluation, bringing predictions closer to the original annotations. However, it has been demonstrated that existing quantitative evaluation metrics do not align well with human opinions, indicating the necessity for a more nuanced assessment (Xu et al., 2023; Zheng et al., 2023; Dettmers et al., 2023).

Human Evaluation In the research conducted by Hayati et al. (2023), the effectiveness of the criteria-based prompting approach was evaluated through human assessment with the participation of crowd workers. Notably, criteria-based prompting garnered preference from humans in more than half of the total statements. The evaluation then has been extended to measure the human capacity to generate diverse opinions for given statements. Participants were instructed to express opinions of agreement or disagreement as extensively as possible on specific statements. Results revealed that individuals tended to provide fewer opinions on statements with more controversial and subjective sentiments. Although human evaluation is expensive, it often results in high-quality data and therefore should be prioritized for high-stake decision-making.

6. Conclusion

This review paper aims to highlight the need to include diverse perspectives that cover a wide range of social groups, especially minority ones. It appears that LLMs can serve as a guide to produce various perspectives while also being aligned with human opinions. One key element to enable is to embrace disagreement and diversity among annotators. Therefore, diverse datasets, including disaggregated ones, should be incorporated into the NLP pipeline (Plank et al., 2014; Dumitrache et al., 2019; Poesio et al., 2019). Proposed solutions, based on the idea of integrating human opinions and relevant personal information into prompts, like socio-demographic prompting and criteria-based prompting, aim to guide models toward responses from specific human groups, but their effectiveness depends on factors such as model size, prompt formulation, input length, and the specific task at hand.

This preliminary review serves as groundwork for future investigations to achieve inclusivity and practical alignment with human perspectives. An initial study could involve guiding LLMs via fine-tuning to generate various perspectives that account for various social groups rather than just providing socio-demographic information during the prompting phase. This strategy may result in the utilization of specialized perspective-aware models that are trained on pairs of personal data and human opinions that are grouped to represent each social group. Furthermore, leveraging human feedback to train the model with reinforcement learning may improve the degree to which LLMs align with human preferences.

Through this literature overview, we emphasize the need to develop LLMs incorporating multiple perspectives and viewpoints, ultimately encouraging participatory design and community involvement in building more equitable models. Concurrently, it is crucial to account for potential risks associated with disclosing sensitive information: socio-demographic data may be exploited to target content towards individuals without their explicit consent.

Acknowledgements

This work has been supported by the European Union under ERC-2018-ADG GA 834756 (XAI), by HumanE-AI-Net GA 952026, and by the Partnership Extended PE00000013 - “FAIR - Future Artificial Intelligence Research” - Spoke 1 “Human-centered AI”.

7. References

Valerio Basile. 2020. It’s the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *2020 AIxIA Discussion Papers Workshop, AIxIA 2020 DP*, volume 2776, pages 31–40. CEUR-WS.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a perspectivist turn in ground truthing for predictive computing. *arXiv preprint arXiv:2109.04270*.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2023. How (not) to use sociodemographic information for subjective nlp tasks. *arXiv preprint arXiv:2309.07034*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189*.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan.

2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. *arXiv preprint arXiv:1904.06101*.

Gizem Gezici, Aldo Lipani, Yucel Saygin, and Emine Yilmaz. 2021. Evaluation metrics for measuring bias in search engine results. *Information Retrieval Journal*, 24:85–113.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2023. How far can we extract diverse perspectives from large language models? criteria-based diversity prompting! *arXiv preprint arXiv:2311.09799*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, Piotr Miłkowski, Jan Kocoń, and Przemyslaw Kazienko. 2022. What if ground truth is subjective? personalized deep neural hate speech detection. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 37–45.

Milton King and Paul Cook. 2020. Evaluating approaches to personalizing language models. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2461–2469.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2023. Aligning large language models with human preferences through representation engineering. *CoRR*, abs/2312.15997.

Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3659–3668.

Negar Mokhberian, Myrl G Marmarelis, Frederic R Hopp, Valerio Basile, Fred Morstatter, and Kristina Lerman. 2023. Capturing perspectives of crowdsourced annotators in subjective learning tasks. *arXiv preprint arXiv:2311.09743*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022a. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022b. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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.

Barbara Plank. 2022. The'problem'of human label variation: On ground truth in data, modeling and evaluation. *arXiv preprint arXiv:2211.02570*.

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.

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.

Julia Romberg. 2022. Is your perspective also my perspective? enriching prediction with subjectivity. In *Proceedings of the 9th Workshop on Argument Mining*, pages 115–125.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B. Pierrehumbert. 2021. Two contrasting data annotation paradigms for subjective NLP tasks. *CoRR*, abs/2112.07475.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*.

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

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *CoRR*, abs/2309.15025.

Nikita Soni, Niranjan Balasubramanian, H Andrew Schwartz, and Dirk Hovy. 2024. Comparing human-centered language modeling: Is it better to model groups, individual traits, or both? *arXiv preprint arXiv:2401.12492*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. *arXiv preprint arXiv:2301.05036*.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593.