The Ghost Annotator: a Framework to Explore Human Label Variation in Content Moderation through Conformal Prediction

Anonymous ACL submission

Abstract

Current research predominantly focuses on model performance while overlooking uncertainty, particularly as LLMs increasingly generate annotated data. We introduce a framework combining conformal prediction with collaborative filtering to detect LLM biases. Using Non-Conformity Scores (NCS), we introduce the *Ghost Prediction* metric and *Ghost Annotator* concept to quantify and profile cases where models diverge from all human annotations. Applying Cosine similarity measures, we identify systematic biases along sociodemographic axes. Evaluating four LLMs across four content moderation datasets we revealed that smaller LLMs tend to be more confident yet less aligned with human annotations compared to larger models, and across all models, uncertainty increases as annotator disagreement rises, mirroring collective human behavior. Finally the Ghost Annotator framework unveils strong alignment between LLMs and annotators of a specific gender on particular datasets.

1 Introduction

Human Label Variation (HLV) (Plank, 2022) recently emerged as a research paradigm aimed to enhance the fairness and inclusivity of language technologies and resources. In overcoming traditional approaches based on label aggregation, HLV claims for a new generation of datasets and models (Uma et al., 2021; Cabitza et al., 2023) that are able to represent different perspectives especially on highly subjective phenomena (Frenda et al., 2025). This shift has important theoretical and practical implications, as biased technologies can harm vulnerable populations in downstream tasks such as automatic content moderation (Kocoń et al., 2021a; Anand et al., 2024).

The Natural Language Processing (NLP) community addresses issue of HLV from a wide range of perspectives: from the development of disaggregated data with annotators' metadata (Sachdeva et al., 2022; Davani et al., 2024), to the assessment and development of technologies that are able to capture different world-views (Wich et al., 2021; Van Der Meer et al., 2024) and 'listen' to the voices of minoritized groups (Vitsakis et al., 2024). Several challenges are still open, though: *i.* The generalization of findings on HLV is hindered by the lack of alignment between datasets and their annotation schemes (Fortuna and Nunes, 2018; Vidgen and Derczynski, 2020); *ii.* most research focuses on model performances, overlooking the issue of uncertainty, which is becoming central with the growing adoption of Large Language Models (LLM) to generate annotated datasets (Tan et al., 2024).

Our work tackles these challenges by presenting the Ghost Annotator: a framework to assess the presence of LLMs bias against specific categories of annotators based on uncertainty estimation. Our approach relies on Conformal Prediction (Chen et al., 2023), a methodology for models' uncertainty estimation, to profile groups of annotators and identify which identities the model align the most with.

Through the design of the Ghost Annotator we answer the following questions

[RQ1] Is there a coherence between models' uncertainty and HLV expressed in disaggregated corpora?

[RQ2] Do models align with specific categories of annotators?

Our results indicate that smaller LLMs exhibit higher confidence in their predictions while diverging more substantially from human annotations than larger models. Despite these differences, all models display confidence patterns that reflect collective annotator behavior: as disagreement among annotators increases for a given message, model uncertainty correspondingly increases, confirming findings from previous works (Schmeisser-Nieto

et al., 2024; Anand et al., 2024). Finally, our framework enables the identification of a strong alignment between LLMs and annotators of a specific gender on particular datasets, as well as a systematic divergence across different model families in the gender-based perspectives they adopt on content moderation–related phenomena.[1]

2 Related Work

The annotators' individual characteristics affect the text perception. Mieleszczenko-Kowszewicz et al. (2023) examined how the psychological and emotional traits of 40 annotators across different tasks and texts, determine the perception of text also over time. The human instability and diversity make in general the reproduction of their annotation hard. However, to lower the annotation time and costs, the use of **pre-trained models for creating dataset**, simulating humans activities and evaluating models' outputs is increasing (Tan et al., 2024; Aher et al., 2023; Li et al., 2024)[2]. This raises the need to evaluate their reliability in replacing humans (Calderon et al., 2025; Gligorić et al., 2025), and to guarantee a degree of diversity in their annotations. Besides the common approaches based on active learning approach used to optimize the annotation budget, some techniques that account for HLV were proposed recently. Wang and Plank (2023) employed an active learning strategy to learn labels from various annotators (identified by their id); Baumler et al. (2023) exploited the mismatch between model and annotator uncertainty to select examples that need more annotations relying on human disagreement; and van der Meer et al. (2024) suggested strategies to choose which human annotator should label specific instances to ensure representativeness in annotated data. However, Gruber et al. (2025) discuss the connection between optimization techniques and HLV, arguing that these techniques do not consider the distinction between HLV and annotation error, and that LLMs can provide label distributions (not only one label as human annotators) making them more attractive as annotators. However, LLMs-as-annotators tend to perform better on English datasets, are biased toward annotating texts as offensive and abusive, produce label distribution not aligned with human opinion distribution (Pavlovic and Poesio, 2024a), and even if prompted with diverse persona, struggle to generate responses as diverse as humans (Sarumi et al., 2025).

Among scholars who studied the **correlation between model prediction and the distinct human responses**, Lan et al. (2025) noticed that models struggle to capture multiple human responses in Visual Question Answering, especially when humans are uncertain about their response, and that calibration techniques based on human distribution are more effective than strategies that calibrate the model towards accuracy; while Schmeisser-Nieto et al. (2024) and (Anand et al., 2024) demonstrated how models exhibit low confidence when annotators have more disagreement with each other. Disagreement can be caused by different factors (Sandri et al., 2023; Wan et al., 2025) and when it is systematic, it is likely to be a symptom of the existence of different perspectives (Frenda et al., 2025). Especially in tasks like hate speech detection, beliefs, identities and demographics are correlated with the level of toxicity and offensive language perceived in a message (Sap et al., 2022a; Mostafazadeh Davani et al., 2024), and if the HLV is not captured by datasets and models, the result is a model unfair behavior (e.g., discrimination of minorities, reinforcement of stereotypes, or eclipsing of segments of population). To investigate the presence of biases in pre-trained models, various scholars explored the use of questionnaires, evaluation frameworks and word association tests with the purpose of unveiling their political or value preference and moral attitude (Wright et al., 2024; Jiang et al., 2025; Rao et al., 2025; Abramski et al., 2024; Dai et al., 2025). All these studies reveal how unfortunately LLMs are not suitable for a global audience.

Inspired by the work of Urbinati et al. (2025), we estimate the uncertainty of models to detect their societal biases through **conformal prediction**. The novelty of our work is a new framework that examines models correlation with HLV, and helps to position their representation, in terms of *Ghost Annotator*, across diverse sociodemographic axes. Recently introduced in NLP (Chen et al., 2023), previous studies exploited conformal prediction to trigger moderator's review in automatic hateful content moderation (Villate-Castillo et al., 2025), to estimate models uncertainty in text generation (Wang et al., 2025), machine translation (Zerva and Mar-

[1]The code of our experiment is available at the following url: https://anonymous.4open.science/r/ghost-annotator-825C/README.md

[2]LLM-as-a-judge is used also in available evaluation frameworks that score the bias of LLMs: https://deepeval.com/

tins, 2024), and text classification (Sheng et al., 2025), and to clean mislabeled data based on a small curated calibration set (Zhan et al., 2023). With our work we provide a fair framework, based on a statistical guaranteed technique (Campos et al., 2024), to evaluate and use conscientiously pre-trained models in the creation and augmentation of training dataset ensuring diverse annotations.

3 Experimental Setting

In this section we present the experimental setting that drives our research. In Section 3.1 we present Conformal Prediction, which is used to estimate models uncertainty against human annotations. In Section 3.2 we describe the Ghost Prediction, an alternative to accuracy-based metric that is used to quantify models divergence from disaggregated human annotators. In Section 3.3 we describe the Ghost Annotator, a framework to profile models and human annotators inspired by Collaborative Filtering and built upon Conformal Prediction and Ghost Predictions. Sections 3.4 and 3.5 respectively present the datasets and models that we adopted in our experiment.

3.1 Conformal Prediction

Conformal Prediction (Angelopoulos et al., 2023; Fontana et al., 2023) is a framework for the estimation of models confidence returning, for each prediction, a distribution of the probabilities of all the possible labels. Depending on this distribution, it is possible to associate the prediction to a Non Conformity Score (NCS), which functions (e.g., Brier score) as a proxy to quantify the uncertainty of the prediction: the higher is the NCS, the more uncertain is the model. The core idea behind Conformal Prediction is that it is possible to calibrate a model by computing its average NCS on a limited set of data (the calibration set) and then use this score to assess the uncertainty of model's predictions on unseen data.

In this work, we adopts Conformal Prediction to estimate model uncertainty against individual annotators in order to identify general patterns of misalignment between models an humans. Specifically, we consider the average NCS as a proxy of this misalignment: the higher is the NCS, the higher is the divergence between model's predictions and human preferences in the annotation. This approach is extremely flexible because it can be adopted to capture individual preferences or group dynamics by partially aggregating annotators.

3.2 Ghost Prediction

Commonly the model evaluation in classification tasks relies on the accuracy performance based on the comparison between model predictions and the *ground truth* obtained aggregating human labels or their distribution (Leonardelli et al., 2025). Recently, some methods of evaluation that take into account HLV were proposed. These consider the comparison of model's predictions with annotators' labels grouped by similar profiles (Akhtar et al., 2021; Gordon et al., 2022; Frenda et al., 2023), and with individual annotators' labels (Mostafazadeh Davani et al., 2022; Mokhberian et al., 2024; Orlikowski et al., 2025; Lo et al., 2025b). Moreover, all these works mainly rely on the computation of accuracy-based metrics (e.g., F1 score, MAE). Inspired by works on human bias investigation (Kocoń et al., 2021b; Mieleszczenko-Kowszewicz et al., 2023) and differently from previous works on LLMs bias measurement (see Section 2), we introduce the *Ghost Prediction* metric. Overcoming the evaluation of model outputs in terms of performance, we look at how frequently the model outputs a label that is not selected by humans, exploiting all the available labels per item.

3.3 Annotators Representation through Collaborative Filtering

Collaborative filtering (CF) is probably the most popular technique in the area of recommender systems (Ricci et al., 2022; Schafer et al., 2007), i.e., software tools which generate personalized suggestions (*recommendations*) promoting items that are most likely to match the needs, preferences or interests of a certain user (Burke et al., 2011) (the *target*), with the aim of mitigating the so-called *information overload* problem (Maes, 1994). The original version of CF, also known as *user-based* CF, draws on the idea that users who agreed on their evaluations for some items in the past are likely to agree on others too: hence, this approach generates recommendations based on items liked by other users with similar tastes, namely, with a similar rating history (Goldberg et al., 1992).

In this work we combine Conformal Prediction with CF to provide annotator representations based on models uncertainty. Our methodology relies on the following assumptions:

1. we consider an LLM as an user that interacts

with an annotation provided by a human annotator;

2. the output of this interaction is a NCS that measures the uncertainty of the model in relation to the annotation provided by the human annotator;

3. each annotator is represented by all the interactions that the model had with their annotations, namely a distribution of all the NCSs associated to their annotations;

4. the model is represented by all the NCSs associated to all its Ghost Prediction (Section 3.2), namely all the occurrences in which the model predict a label that diverges from all the labels expressed by human annotators.

In order to adequately compare annotators that labeled different amount of messages, we represent them as a 3-dimensional vector derived from the quartiles of the NCS distributions.

The model is represented as a 3-dimensional vector, as well. Since the model's representation is based on the NCSs of its ghost predictions, we define this representation as **Ghost Annotator**: a virtual annotator that minimizes the NCS by providing an annotation that aligns with the predicted label.

By applying Cosine similarity between the Ghost Annotator and all the representations of human annotators, we are able to identify human perspectives to which the model better align with.

We adopt this methodology to systematically explore the alignment of models with specific categories of annotators in perceiving relevant phenomena for content moderation (e.g., Hate Speech, offensiveness.

3.4 Datasets

We chose four datasets annotated for topics related to content moderation, in order to assess the generalization of our method across different phenomena. We followed two guiding principles for data selection to ensure comparability between corpora: *i.* we only selected datasets with a scalar annotation scheme to ensure a coherent scheme across them; *ii.* we did not include datasets provided by the same research group to avoid research bias (Hovy and Prabhumoye, 2021) and maximize their reciprocal independence. The benchmark includes the following datasets:

Attitudes (Sap et al., 2022b): a corpus of 627 tweets annotated for Hate Speech (HS) detection on a scale from 1 to 5.
CADE (Lo et al., 2025a): a corpus of $2,094$ YouTube comments ranked on the basis of their unacceptability on a scale from 1 to 4.
Disentangling (Davani et al., 2024): a corpus of $4,554$ messages from Wikipedia Talk pages[3] and Civil Comments[4] annotated for offensiveness on a scale from 0 to 4.
MHS (Sachdeva et al., 2022): a corpus of $39,461$ tweets annotated according to a multidimensional annotation scheme on a scale from 0 to 4. For this study we focused on the axis of violence

Three types of descriptive statistics have been extracted from each dataset to identify different and common features between them.

Distribution of majority types. Inspired by existing work of Leonardelli et al. (2021), we described each message according to the type of majority formed by annotators: unanimity ($x = 1$), qualified majority ($0.66 < x < 1$), absolute majority ($0.5 < x < 0.66$), and relative majority ($x <= 0.5$). As it can be observed in Table 1, the distribution of majority types significantly differ between datasets suggesting divergent annotation behaviors across datasets.

Label fitting and average number of annotators. This statistics describes the average percentage of scalar values that have been selected by at least one annotator. Excluding Disentangling, whose average number of 32.2 annotators *per* message causes a very high label fitting, differences also emerge between datasets with a comparable average number of annotators.

Annotators Isolation. This statistic reports the average percentage of annotators to label a message in contrast with the majority. Coherently with the high number of annotations *per* message, Disentangling has the highest annotation isolation but differences also arise between the other corpora.

3.5 Models

In this experiment, a selection of pre-trained language models was employed to tackle a text annotation task, aimed at classifying social media posts based on the presence of harmful content (the prompts and the experimental setup are reported in

[3] https://en.wikipedia.org/wiki/Help:Talk_pages
[4] civilcomments.com

Dataset	Avg. Ann.	Rel. Maj.	Abs. Maj.	Qual. Maj.	Unan.	Label Fitting (avg)	Isolation (avg)
Attitudes (Hate Speech) (Sap et al., 2022b)	5.523	0.632	0.038	0.311	0.019	0.578	0.350
CADE (Acceptability) (Lo et al., 2025a)	5.700	0.343	0.206	0.337	0.114	0.505	0.325
Disentangling (Offensiveness) (Davani et al., 2024)	32.324	0.599	0.217	0.179	0.005	0.916	0.472
MHS (Violence) (Sachdeva et al., 2022)	5.856	0.354	0.003	0.280	0.363	0.351	0.261

Table 1: Description of datasets according to the following axes. **Majorities**: percentage of relative majority ($<= 0.50$), absolute majority ($0.50 < x < 0.66$), qualified majority ($0.66 < x < 1$). **Label fitting**: all the labels chosen by at least one annotator / all the possible labels; **Isolation**: % of times in which an annotator diverges from majority.

Appendix A and B). The focus is identifying different categories of harmful content such as violence, hate speech, acceptability, and offensiveness.

The chosen models were tasked with generating probabilities for specific labels and calculating the NCS. Two families of LLMs, Qwen and Llama, were selected for benchmarking performance across different model scales. Therefore, we employ two smaller models (i.e., Qwen/Qwen2.5-1.5B-Instruct, Meta-llama/Llama-3.2-1B-Instruct), one from the Qwen family and one from the Llama family, along with their respective medium-sized counterparts (i.e., Qwen/Qwen2.5-7B-Instruct, Meta-llama/Llama-3.1-8B-Instruct). In particular, these models were selected for their ability to understand instructions and generate responses tailored to classification tasks.

4 Results

In this section we present the results of our experiments. Section 4.1 presents results about the comparison between models' uncertainty and HLV (RQ1); Section 4.2 describes the impact of datasets and models in the alignment of LLMs with annotators' gender (RQ2).

4.1 [RQ1] Is there a coherence between HLV and models' uncertainty?

Our first experiment is aimed at exploring the coherence between LLMs uncertainty and collective behaviors in the context of dataset annotation. We jointly study the average NCS of LLMs across datasets (Section 3.1) and their tendency to output ghost predictions (Section 3.2). We observe whether there is a pattern between different majority types emerging between human annotators and models uncertainty.

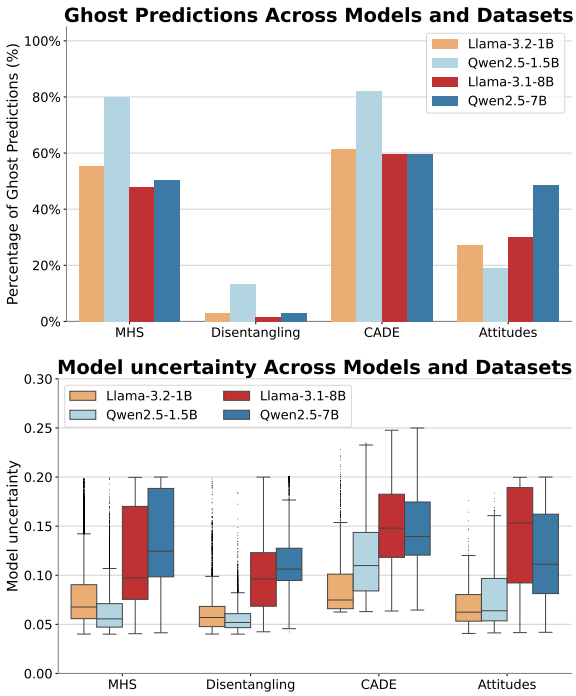

Figure 1: (Top) Percentage of the Ghost prediction across models and datasets. (Bottom) The box plot illustrates the distribution of model uncertainty for each dataset, with separate colors representing different models. Outliers are marked with small black dots. The datasets are labeled on the x-axis, and the NCS (indicating model uncertainty) are represented on the y-axis.

Smaller models are more confident about the 'wrong' label. The average conformity score of models across datasets (Figure 1 Bottom) shows that smaller models are more confident about their predictions, regardless their LLM family and the type of datasets. Excluding Qwen2.5-1.5B on the CADE dataset, the conformity score of smaller LLMs exhibit lower variation than their counterparts. However, the higher confidence of these models appears to be associated with a higher divergence with human annotations. As it can be observed in Figure 1 Top, the Ghost Prediction

metric, namely the frequency with which a model outputs a label not chosen by any annotator, is higher for smaller models in three cases out of four. Qwen2.5-1.5B systematically scores the highest Ghost Annotator on MHS, Disentangling, and CADE. Observing variation across datasets it is worth mentioning the significantly lower Ghost Annotator in models predictions on Disentangling's dataset, which appears to be caused by the higher number of annotations *per* message, and the higher Ghost Annotator achieved by models on MHS and CADE. This behavior may be explained by the type of phenomena that are less explored in NLP and thus it is more likely that models diverge from human annotations in their classification.

Models uncertainty is coherent with majority types. Figure 2 Top shows that the correlation between models NCS and the majority types has almost always a negative sign: the higher is the NCS in prediction, the smaller is the majority of annotators who label a message with the same label. This pattern, which shows to a certain extent a coherence between LLMs and human uncertainty, emerges with different magnitude, depending on the dataset and the model. Larger LLMs shows a moderate negative correlation between NCS and majority types in predictions on Disentangling Corpus (-0.44 and -0.45); Llama-3.1-8B on CADE (-0.39). Smaller models exhibit a moderate negative correlation on CADE (Llama3.2-1B, -0.31) and Attitudes (Qwen2.5-1.5B, -0.51). MHS is the only dataset in which NCS correlation with majority types is weak for all LLMs.

The correlation between models NCS and annotator isolation (Figure 2 Bottom) is specular with pattern emerging from the analysis of majority types: the higher frequency of annotators who vote against majority, the higher is the models uncertainty. Notably, different datasets are characterized by a positive correlation with a higher magnitude. E.g., larger LLMs show a moderate correlation ($+0.30$) between NCS and annotator isolation on MHS.

4.2 [RQ2] Do models align with specific categories of annotators?

Our second experiment adopts the Ghost Annotator (Section 3.3) to identify whether LLMs align with the perspectives of annotators characterized by specific socio-demographic traits. Since gender is the only socio-demographic trait consistently

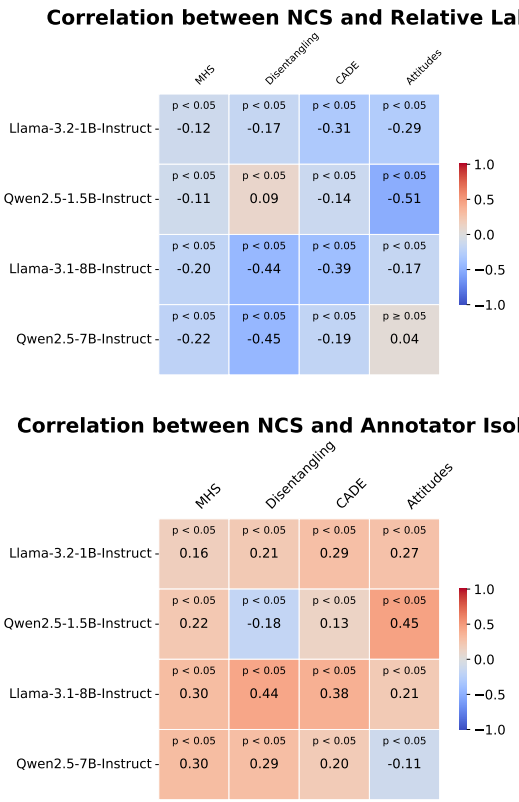

Figure 2: (Top) The Heatmap displays the Pearson correlation between the NCS and majority types. (Bottom) The figure shows the Pearson correlation between NCS and Annotator Isolation for the analyzed models and datasets.

shared across datasets we limited our analysis to this feature.

Given a model and a dataset, we performed the following steps:

1. we generated a representation of the model by computing the Ghost Annotator (Section 3.3);
2. we selected the 10 women annotators and the 10 men annotators whose representation based on NCS is more similar to the Ghost Annotator[5];
3. we identified the annotators group that is more similar to the Ghost Annotator interpreting this similarity as the proxy of a higher alignment of the model with a specific gender;
4. keeping the Ghost Annotator derived from the initial corpus, we compared this representation with human annotators from other corpora

[5]We replicated the experiment for top-20 and all annotators (Appendix C)

This approach allows us to identify patterns that characterize the interaction between models, human annotators, and datasets: we do not only identify the eventual alignment of a model with a specific gender but **we also observe whether this alignment is generalizable outside the context of a specific corpus**. We repeated the experiment for each model for each dataset to analyze differences between models and the impact of specific datasets.

Results of this experiments are shown in Figure 3. Each sub-figure represents the Ghost Annotator alignment of a model with men and women: the rows correspond to the dataset used to profile the Ghost Annotator, the columns represent the dataset used to profile the annotators. A value near to 1 means that the model is nearer to women; near to -1 the opposite; near to 0 the lack of alignment.

Models representation is not stable across datasets. A first notable result emerging from Figure 3 is that the Ghost Annotator consistently aligns with men annotators in CADE and women annotators in MHS, regardless of the corpus used to profile the Ghost Annotator. This alignment between models and gender is always statistically significant, ranging from 0.31 to 0.78 in MHS and from -0.24 to -0.87 in CADE. By contrast, the profiles derived from the Disentangling and Attitudes corpora are less stable, though still comparably pronounced. Specifically, the alignment of the Ghost Annotator with human annotators profiled from the Disentangling corpus ranges from -0.39 to 0.33, while for Attitudes it ranges from -0.24 to 0.50. A possible interpretation of the stability of models representation in CADE and MHS might be the tendency of annotators with different genders to be more polarized in interpreting phenomena like violence (MHS) and acceptability (CADE).

Significant differences emerge between LLMs. A second relevant pattern emerging from the experiment is the centrality of LLMs biases in determining the alignment of models with specific gender. Observing the similarity between Ghost and human annotators profiled on Attitudes (Figure 3.3, top right square in each heatmap) it is possible to observe a strong divergence between Llama and Qwen families. Both the smaller and the larger Llama are more aligned with men (-0.17 and -0.24); smaller and larger Qwen with women (0.38 and 0.50). In other cases the divergence between models appears to depend by their size. Smaller models are both consistently aligned with men profiled on MHS and compared with human annotators profiled on Disentangling (second square of the last row, starting drom left): Llama Ghost Annotator scores -0.32; Qwen -0.38. Higher models are slightly more alignment with women (Llama 0.07; Qwen 0.12). This variability might be the proxy of existing biases in LLMs on specific tasks that influence their alignment with specific genders.

5 Conclusion

This work proposes the **Ghost Annotator**, a framework designed to uncover and assess biases in LLMs against specific groups of annotators through uncertainty estimation. While prior research emphasized accuracy-based performance of LLMs, neglecting uncertainty—despite the growing use of LLMs for data annotation—our approach integrates conformal prediction and collaborative filtering to detect sociodemographic biases. Specifically, leveraging Non-Conformity Scores, we introduce the Ghost Prediction metric and the Ghost Annotator concept to capture when and how the model outputs diverge from all human annotations. We evaluated four models (two large and two small from the Qwen and Llama families) across four datasets reporting disaggregated scalar annotations on diverse dimensions of abusive language: violence, hate speech, acceptability and offensiveness. Finally, employing the Euclidean similarity between the Ghost Annotator and all the representations of human annotators, we identified human perspectives to which the model better align with.

Our findings show that smaller LLMs are generally more confident in their outputs but deviate more from human annotations than larger models. Nonetheless, in line with previous works' findings (Schmeisser-Nieto et al., 2024; Anand et al., 2024), all models exhibit confidence trends that mirror annotator consensus, with higher uncertainty arising when annotators disagree more. Additionally, the proposed framework uncovers strong gender-specific alignment between certain LLMs and annotators on specific datasets (in particular, male annotators in CADE and female annotators in MHS), along with consistent differences across model families about which (female or male) perspectives they adopt to judge content moderation especially for the hate speech phenomenon (male perspective in Llama models and female perspective in Qwen models).

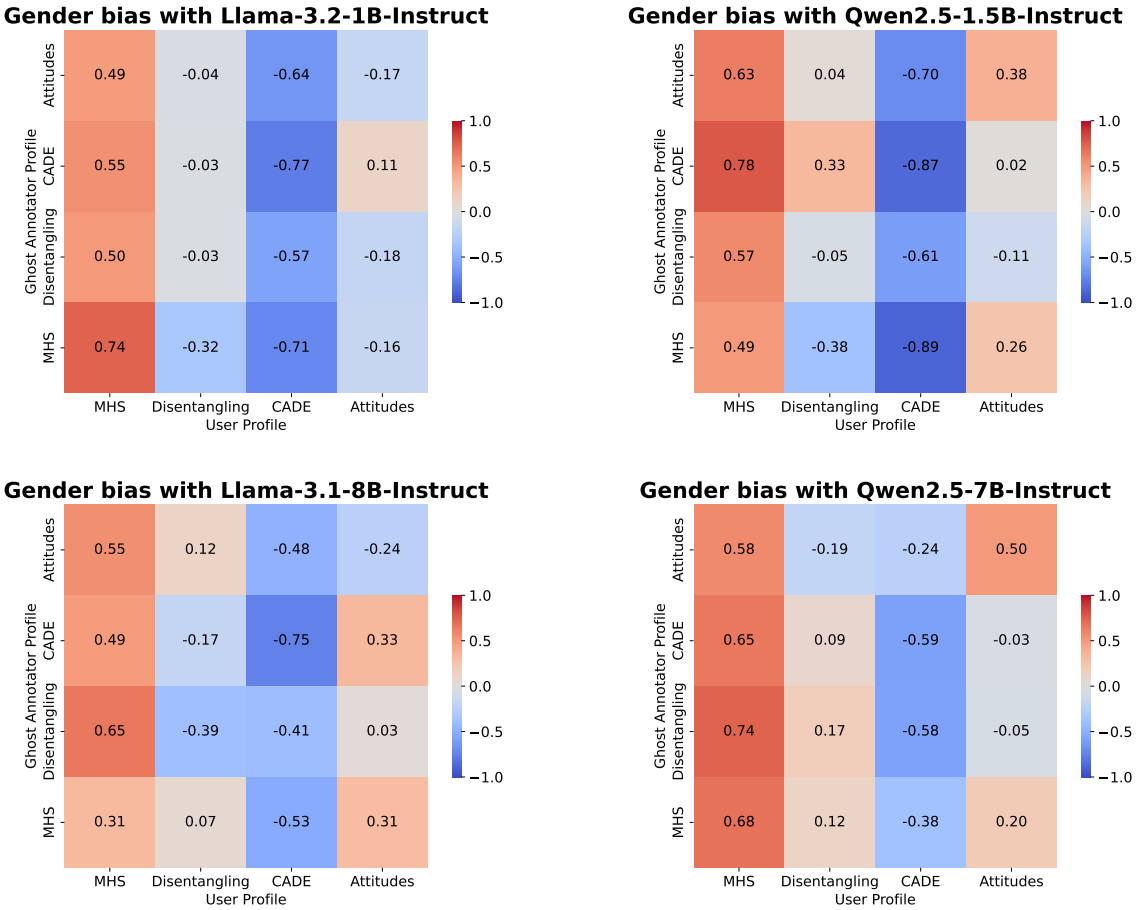

Figure 3: The figure shows the profile of the ghost annotator extracted from all models for each dataset. Each of the four sub-figures represents the similarity index between the ghost annotator, calculated for that model, and the four datasets. In each sub-figure, the rows correspond to the dataset used to profile the ghost annotator based on its NCS quartiles, while the columns represent the dataset used to profile the annotators based on their NCS quartiles. A value of 0 indicates identical distance between male annotators and the ghost, as well as between female annotators and the ghost. A value of -1 indicates that the ghost annotator is closer to the male profile than the female one, while a value of 1 indicates that the ghost annotator profile is closer to the female one. Only the 10 male and 10 female annotators most similar to the ghost annotator are considered.

Limitations

In our work, we used the gender as a binary dimension as a case study to prove the bias detection ability of our framework. However, we are aware of the limitation of the use of binary gender sociodemographic information, and for this reason we consider important to prove the generalization of our approach with multi-categorical dimension. Indeed, we mapped the diverse gender information provided by the authors of these datasets in a binary variable to obtain comparable categories for all the datasets.

Another limitation of our analysis is about the model families and size we evaluated. For the selection of models we took into account their open availability and their possible extensive use because of small and medium size (i.e., requiring lower computational power). However, we are aware that services and applications for daily assistant activities are fed mainly by close models, and in the future we consider to employ the proposed framework to evaluate the imperfections of real-world applications.

Ethical Considerations

Our research focuses on capturing sociodemographic biases in models already used by users worldwide. We are conscious that it is risky to consider limited societal biases and adopt a binary categorization for gender. However, the proposed framework is employable to multiple categories and societal dimensions (e.g., ethnicity, origin, disabilities, educational status and so on). We hope our framework can be used to analyze the safety of the models before their release, and that this investigation can encourage attention to societal issues in the creation of AI.

References

Katherine Abramski, Clara Lavorati, Giulio Rossetti, Massimo Stella, and 1 others. 2024. Llm-generated word association norms. *Frontiers in Artificial Intelligence and Applications*, 386:3–12.

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.

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*.

Abhishek Anand, Negar Mokhberian, Prathyusha Kumar, Anweasha Saha, Zihao He, Ashwin Rao, Fred Morstatter, and Kristina Lerman. 2024. Don't blame the data, blame the model: Understanding noise and bias when learning from subjective annotations. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 102–113, St Julians, Malta. Association for Computational Linguistics.

Anastasios N Angelopoulos, Stephen Bates, and 1 others. 2023. Conformal prediction: A gentle introduction. *Foundations and trends® in machine learning*, 16(4):494–591.

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.

Robin D. Burke, Alexander Felfernig, and Mehmet H. Göker. 2011. Recommender systems: An overview. *AI Mag.*, 32(3):13–18.

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.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria. Association for Computational Linguistics.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário AT Figueiredo, and André FT Martins. 2024. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516.

Zecong Chen, Yuhan Xie, and Mark Fishel. 2023. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*.

Xunlian Dai, Li Zhou, Benyou Wang, and Haizhou Li. 2025. From word to world: Evaluate and mitigate culture bias in LLMs via word association test. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24521–24537, Suzhou, China. Association for Computational Linguistics.

Aida Davani, Mark Díaz, Dylan Baker, and Vinodkumar Prabhakaran. 2024. Disentangling perceptions of offensiveness: Cultural and moral correlates. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2007–2021.

Matteo Fontana, Gianluca Zeni, and Simone Vantini. 2023. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23.

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

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: 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.

Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel Candes, and Dan Jurafsky. 2025. Can unconfident llm annotations be used for confident conclusions? 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 3514–3533.

David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. 1992. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70.

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.

Cornelia Gruber, Helen Alber, Bernd Bischl, Göran Kauermann, Barbara Plank, and Matthias Aßenmacher. 2025. Revisiting active learning under (human) label variation. In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 75–86, Suzhou, China. Association for Computational Linguistics.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and linguistics compass*, 15(8):e12432.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2025. Can language models reason about individualistic human values and preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6757–6794, Vienna, Austria. Association for Computational Linguistics.

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

Jan Kocoń, Marcin Gruza, Julita Bielaniewicz, Damian Grimling, Kamil Kanclerz, Piotr Miłkowski, and Przemysław Kazienko. 2021b. Learning personal human biases and representations for subjective tasks in natural language processing. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1168–1173.

Jian Lan, Diego Frassinelli, and Barbara Plank. 2025. Mind the uncertainty in human disagreement: Evaluating discrepancies between model predictions and human responses in vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):4446–4454.

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: The 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, Sara Tonelli, and 1 others. 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. Association for Computational Linguistics.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Soda Marem Lo, Oscar Araque, Rajesh Sharma, and Marco Antonio Stranisci. 2025a. That is unacceptable: the moral foundations of canceling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6625–6639, Vienna, Austria. Association for Computational Linguistics.

Soda Marem Lo, Silvia Casola, Erhan Sezerer, Valerio Basile, Franco Sansonetti, Antonio Uva, and Davide

Bernardi. 2025b. 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.

Pattie Maes. 1994. Agents that reduce work and information overload. *Commun. ACM*, 37(7):30–40.

Wiktoria Mieleszczenko-Kowszewicz, Kamil Kanclerz, Julita Bielaniewicz, Marcin Oleksy, Marcin Gruza, Stanislaw Wozniak, Ewa Dziecіol, Przemyslaw Kazienko, and Jan Kocon. 2023. Capturing human perspectives in nlp: Questionnaires, annotations, and biases. In *NLPerspectives@ ECAI*.

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 Diaz, Dylan K Baker, and Vinodkumar Prabhakaran. 2024. D3CODE: Disentangling disagreements in data across cultures on offensiveness detection and evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18511–18526, Miami, Florida, USA. 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.

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.

Maja Pavlovic and Massimo Poesio. 2024a. 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.

Maja Pavlovic and Massimo Poesio. 2024b. Understanding the effect of temperature on alignment with human opinions. *Proceedings of Algorithmic Fairness through the lens of Metrics and Evaluation Workshop*.

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.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

Francesco Ricci, Lior Rokach, and Bracha Shapira. 2022. Recommender systems: Techniques, applications, and challenges. In Francesco Ricci, Lior Rokach, and Bracha Shapira, editors, *Recommender Systems Handbook*, pages 1–35. Springer US.

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

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

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

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

Olufunke O. Sarumi, Charles Welch, Daniel Braun, and Jörg Schlötterer. 2025. The impact of annotator personas on LLM behavior across the perspectivism spectrum. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 121–136, Southern

Denmark University, Odense, Denmark. Association for Computational Linguistics.

J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. *Collaborative Filtering Recommender Systems*, pages 291–324. Springer Berlin Heidelberg, Berlin, Heidelberg.

Wolfgang S. Schmeisser-Nieto, Pol Pastells, Simona Frenda, and Mariona Taule. 2024. Human vs. machine perceptions on immigration stereotypes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8453–8463, Torino, Italia. ELRA and ICCL.

Huanxin Sheng, Xinyi Liu, Hangfeng He, Jieyu Zhao, and Jian Kang. 2025. Analyzing uncertainty of LLM-as-a-judge: Interval evaluations with conformal prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11286–11328, Suzhou, China. Association for Computational Linguistics.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

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.

Alessandra Urbinati, Mirko Lai, Simona Frenda, and Marco Stranisci. 2025. Are you sure? measuring models bias in content moderation through uncertainty. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18061–18076, Suzhou, China. Association for Computational Linguistics.

Michiel Van Der Meer, Neele Falk, Pradeep 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.

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.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Guillermo Villate-Castillo, Javier Del Ser, and Borja Sanz. 2025. A collaborative content moderation framework for toxicity detection based on multitask neural networks and conformal estimates of annotation disagreement. *Neurocomputing*, 647:130542.

Nikolas Vitsakis, Amit Parekh, and Ioannis Konstas. 2024. Voices in a crowd: Searching for clusters of unique perspectives. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12517–12539, Miami, Florida, USA. Association for Computational Linguistics.

Ruyuan Wan, Haonan Wang, Ting-Hao Kenneth Huang, and Jie Gao. 2025. From noise to nuance: Enriching subjective data annotation through qualitative analysis. In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pages 240–254, Suzhou, China. Association for Computational Linguistics.

Sean Wang, Yicheng Jiang, Yuxin Tang, Lu Cheng, and Hanjie Chen. 2025. Copu: Conformal prediction for uncertainty quantification in natural language generation. *arXiv preprint arXiv:2502.12601*.

Xinpeng Wang and Barbara Plank. 2023. ACTOR: Active learning with annotator-specific classification heads to embrace human label variation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2046–2052, Singapore. Association for Computational Linguistics.

Maximilian Wich, Christian Widmer, Gerhard Hagerer, and Georg Groh. 2021. Investigating annotator bias in abusive language datasets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1515–1525, Held Online. INCOMA Ltd.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. LLM tropes: Revealing fine-grained values and opinions in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17085–17112, Miami, Florida, USA. Association for Computational Linguistics.

Chrysoula Zerva and André FT Martins. 2024. Conformalizing machine translation evaluation. *Transactions of the Association for Computational Linguistics*, 12:1460–1478.

Xianghao Zhan, Qinmei Xu, Yuanning Zheng, Guangming Lu, and Olivier Gevaert. 2023. Reliability-based cleaning of noisy training labels with inductive conformal prediction in multi-modal biomedical data mining. *arXiv preprint arXiv:2309.07332*.

A Prompt and Prediction Extraction

The goal of the prompt is to classify social media posts according to specific content-related cate-

gories, with a clear scoring system[6].

Each dataset used in the experiment is associated with a distinct prompt, where the user is instructed to annotate a social media post with one of the predefined values from the following options:

- **Violence**: 0 (no violence) to 4 (extremely violent)
- **Hate Speech**: 1 (no hate speech) to 5 (extremely hate speech)
- **Unacceptability**: 1 (acceptable) to 4 (extremely unacceptable)
- **Offensiveness**: 0 (no offensiveness) to 4 (extremely offensive)

The prompt structure follows a similar format across all datasets, which ensures consistency in the model's task understanding and response generation. An example prompt for violence classification is as follows:

Task: You are a participant in an annotation task for the recognition of violence.

Instruction: Read the following social media post and annotate it with one value from the following options. Options: 0, 1, 2, 3, 4. 0 == no violence, 4 == extremely violent.

Output format: The answer should follow this template 'answer': option. Answer only in JSON. No extra text.

Question: How much does the following social media post contain violence?

The trans women reading this tweet right now is beautiful

To extract the predictions, we used the transformers library from Hugging Face. Initially, the text input, which includes both the prompt and the social media post, was tokenized using the model's corresponding tokenizer. This tokenized input was then passed through the model to generate predictions. The model outputs raw logits, which are uncalibrated scores representing the model's confidence for each possible token. These logits were then passed through a log-softmax function in order to normalize the logits and to provide a probability distribution where the sum of all token probabilities equals one.

[6]Differently from previous works (Pavlovic and Poesio, 2024b; Sarumi et al., 2025), we do not tune the models to exhibit behavior similar to humans.

The probabilities for the target labels (e.g., '0', '1', '2', '3', '4') were gathered across the generated tokens. These probabilities were averaged over multiple steps of token generation to provide a more robust prediction.

B Hardware and Experimental Setup

Each experimental run was allocated a single compute node with the following specifications: 4 CPU cores, 25GB of RAM, and one NVIDIA H200 GPU. The experiment ran for 40 hours. Models were always initialized with their default setup of hyperparameters.

C Sanity Check

This Appendix reproduces the similarity of the ghost annotator with men and women in two different settings: top-20 similar women and men (Figure 4); considering all women and men (Figure 5)

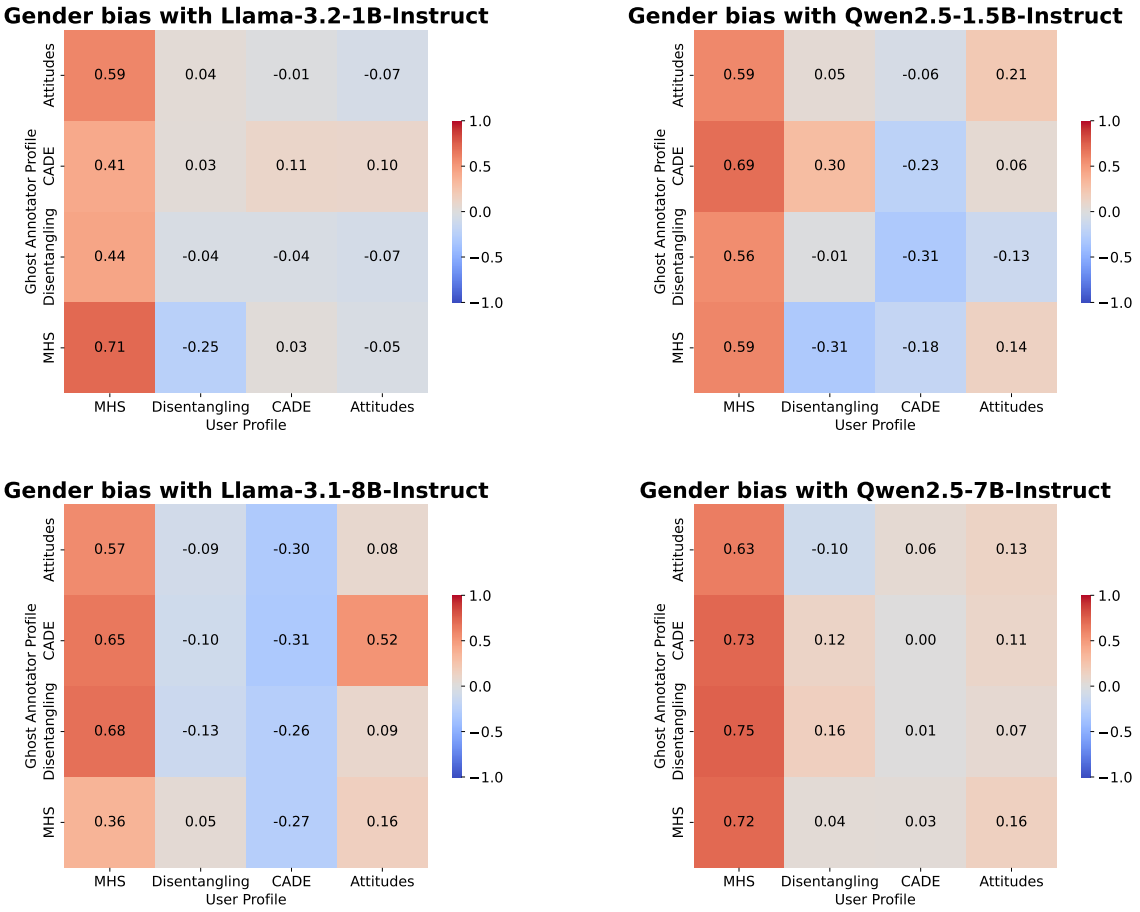

Figure 4: The figure shows the profile of the ghost annotator extracted from all models for each dataset. Each of the four sub-figures represents the similarity index between the ghost annotator, calculated for that model, and the four datasets. In each sub-figure, the rows correspond to the dataset used to profile the ghost annotator based on its NCS quartiles, while the columns represent the dataset used to profile the annotators based on their NCS quartiles. A value of 0 indicates identical distance between male annotators and the ghost, as well as between female annotators and the ghost. A value of -1 indicates that the ghost annotator is closer to the male profile than the female one, while a value of 1 indicates that the ghost annotator profile is closer to the female one. Only the 20 male and 20 female annotators most similar to the ghost annotator are considered.

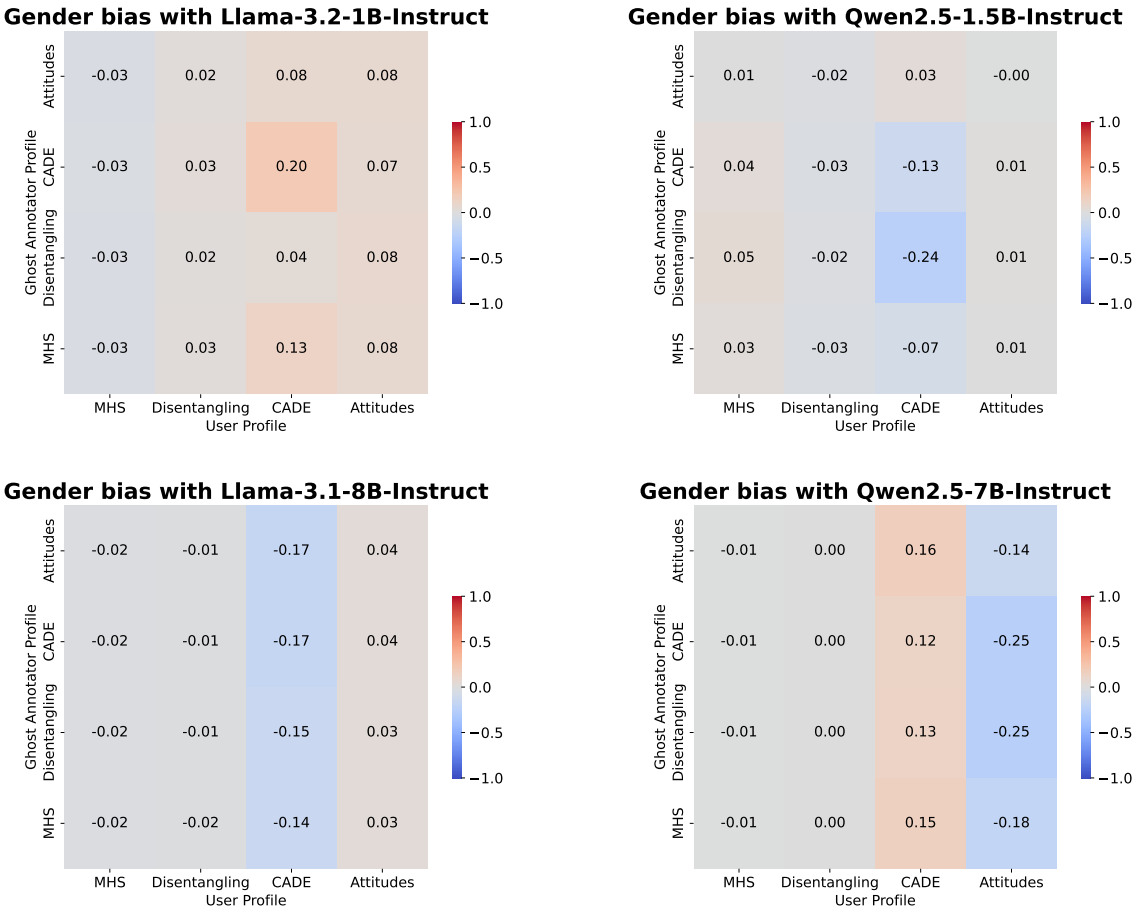

Figure 5: The figure shows the profile of the ghost annotator extracted from all models for each dataset. Each of the four sub.figures represents the similarity index between the ghost annotator, calculated for that model, and the four datasets. In each sub-figure, the rows correspond to the dataset used to profile the ghost annotator based on its NCS quartiles, while the columns represent the dataset used to profile the annotators based on their NCS quartiles. A value of 0 indicates identical distance between male annotators and the ghost, as well as between female annotators and the ghost. A value of -1 indicates that the ghost annotator is closer to the male profile than the female one, while a value of 1 indicates that the ghost annotator profile is closer to the female one. All annotator are considered here.