The Impact of Annotator Personas on LLM Behavior Across the Perspectivism Spectrum

Olufunke O. Sarumi[1], Charles Welch[2], Daniel Braun[1], Jörg Schlötterer[1,3]
[1]University of Marburg, [2]McMaster University, [3]University of Mannheim
{sarumio,daniel.braun,joerg.schloetterer}@uni-marburg.de[1], cwelch@mcmaster.ca[2]

Abstract

In this work, we explore the capability of Large Language Models (LLMs) to annotate hate speech and abusiveness while considering predefined annotator personas within the strong-to-weak data perspectivism spectra. We evaluated LLM-generated annotations against existing annotator modeling techniques for perspective modeling. Our findings show that LLMs selectively use demographic attributes from the personas. We identified prototypical annotators, with persona features that show varying degrees of alignment with the original human annotators. Within the data perspectivism paradigm, annotator modeling techniques that do not explicitly rely on annotator information performed better under weak data perspectivism compared to both strong data perspectivism and human annotations, suggesting LLM-generated views tend towards aggregation despite subjective prompting. However, for more personalized datasets tailored to strong perspectivism, the performance of LLM annotator modeling approached, but did not exceed, human annotators.

1 Introduction

Perspectivism in Natural Language Processing (NLP) aims to preserve the spectrum of opinions held by annotators in corpora (Cabitza et al., 2023). Dataset annotation for this purpose often uses a descriptive paradigm (Rottger et al., 2022), involving minimal instructions and multiple annotators providing labels for every corpus sentence to capture diverse viewpoints. The number of annotators involved can range significantly, from a minimum of 2 to 2500 or more (Plepi et al., 2022; Frenda et al., 2024).

Most traditional approaches aggregate labels to obtain a single majority label (Davani et al., 2022; Aroyo and Welty, 2015), which is commonly used for training models. However, the perspectivist approach argues that critical information is lost when labels are aggregated. More importantly, the opinions of the minority, which may represent a significant population, are undermined, leading to underrepresentation and overshadowing of nuances inherent in the dataset. This is crucial because people's views and opinions are indeed shaped by different socio-cultural, demographic, economic, and experiential backgrounds (Akhtar et al., 2021; Almanea and Poesio, 2022; Demszky et al., 2020; Kennedy et al., 2022). These factors impact how individuals perceive, interpret, and respond to various topics, making it unrealistic to assume everyone shares similar views on the same subject. Recognizing and reflecting opinion differences in our models is therefore important for developing socially aware NLP systems, treating disagreements not as errors but as distinct perspectives. To address this, models have been developed that can learn from such disaggregated labels (Leonardelli et al., 2023; Sullivan et al., 2023; Vitsakis et al., 2023; García-Díaz et al., 2023; Cui, 2023; Xu et al., 2024).

Furthermore, while some disagreements stem from different perspectives, other factors also cause disagreement in data annotations, including temporal factors, annotator inconsistencies, uncertainty, ambiguities, lack of task understanding, or a perfunctory approach to annotation (Fleisig et al., 2024). When modeling perspectives obtained from subjective tasks, these perspectives are often mixed with noise and errors, raising the question of whether true perspectives or merely annotator inconsistencies have been modeled. Some literatures have quantified these uncertainties to a minimal extent (Klemen and Robnik-Šikonja, 2022; Davani et al., 2022).

In this work, we aimed to investigate how existing annotator modeling techniques would behave when trained on deterministic LLM-generated annotations, in contrast to earlier work that explored modeling individual human annotators' perspectives using disaggregated labels. We generated

new annotations for the HS-Brexit and ConvAbuse datasets using Llama2-13B, guided by persona-based prompting derived from annotator information provided by the original authors.

In generating these annotations, we implemented two perspectivism approaches: *strong* and *weak* data perspectivism. Weak perspectivism, also known as reduced perspective, involves considering multiple labels which are ultimately aggregated into one, representing a group opinion. Strong perspectivism, by contrast, utilizes and retains all distinct labels from training through evaluation (Cabitza et al., 2023; Frenda et al., 2024).

Our findings show that LLMs struggle to generate responses as diverse as humans, even with diverse personas. They still partially align with human annotations but tend to pick up only selected persona features. Furthermore, we identified latent annotation prototypes shared by multiple human annotators. These alignment patterns vary across datasets and perspectivism strategies: for instance, HS-Brexit with contrasting demographic attributes shows stronger alignment with human annotations under weak perspectivism, whereas ConvAbuse demonstrates closer alignment with human annotations when strong data perspectivism is used, involving highly personalized and overlapping persona features.

2 Related Work

The first part of this section addresses how Large Language Models (LLMs) have been used to generate different perspectives and their ability to adopt an assigned persona. It also highlights the lack of connection between perspectivism, based on defined personas and annotations in subjective tasks. The second part focuses on the use of LLMs as annotators, examining their ability to generate discrete multiple labels, identifying the lack of persona-based labeling, and replicating human annotation behavior to enable alignment with human annotations.

2.1 LLMs in Perspectivism and Adopting Personas

LLMs have been explored for their ability to simulate diverse human perspectives. Subjective tasks often involve annotators with different backgrounds, leading to divergent opinions which often reflect demographic variation, different and substantial opinions, these make label aggregation inadequate (Rottger et al., 2022). Some works argue that LLMs naturally contain persona traits, as they are trained in large corpora, often culled from social networks that contain crowd-sourced data rich with diverse viewpoints (Hu and Collier, 2024; Vitsakis et al., 2023). For example, Hayati et al. (2024) showed that it is possible to generate multiple perspectives from LLMs and quantify the maximum number of perspectives derivable from an LLM. However, the influence of persona prompting remains debated and the influence of specific persona traits remains underexplored (Beck et al., 2024; Sun et al., 2025). Hu and Collier (2024) suggests that personas have minimal effect on LLM outputs, whereas a psycholinguistic research found that LLMs can generate human-like outputs, even surpassing humans in turing experiments, yet exhibit unnaturally high accuracy that is not possible within human populations (Aher et al., 2023). Furthermore, Wang et al. (2024) found that LLMs risk homogenizing or misrepresenting marginalized identity groups, particularly when asked to simulate them. These challenges highlight the difficulty in separating the LLM's inherent persona from externally applied persona prompts. Despite this, prompting LLMs with well-defined personas, particularly those grounded in demographic traits from existing datasets, offers a practical way to examine how perspective alignment occurs between machines and humans. However, small variations in prompt configurations can lead to large differences in output, complicating reproducibility and fairness evaluations.

2.2 LLM Annotations and Label Generation

Beyond simulating perspectives, LLMs are being explored as direct substitutes for human annotators (Ivey et al., 2024; Bavaresco et al., 2024), especially in settings where collecting human annotations is expensive or slow (Huang et al., 2023; Gligorić et al., 2024). Recent studies have examined the ability of LLMs to generate discrete labels for classification tasks, often using crowd-sourced datasets as benchmarks (Pavlovic and Poesio, 2024a; Gilardi et al., 2023). Gilardi et al. (2023) found that LLMs outperformed crowd-sourced workers in certain annotation tasks, while Pavlovic and Poesio (2024b) demonstrated that adjusting temperature values can control LLM behavior to better simulate annotation disagreement or consistency. These findings suggest that LLMs can be tuned to exhibit behavior similar to individual or aggregated hu-

man annotators. LLMs have also been deployed in replicating prior annotation experiments. For example, Pavlovic and Poesio (2024a) replicated a Learning With Disagreement task (Leonardelli et al., 2023) using GPT-3 but did not incorporate the demographic background of annotators, limiting their insight into perspective-specific agreement. While many experiments rely on LLMs generating explanations or engaging in dialogue-based tasks, fewer works have explored their ability to produce discrete, disaggregated annotations comparable to crowdsourced annotators. Likewise, existing annotator modeling techniques are yet to be fully evaluated on annotations generated by LLMs. The impact of LLM annotations and predefined personas on existing annotator modeling approaches remains unexplored and is a key area we address in our study.

3 Dataset

We used two datasets from the SemEval-2023 task on learning with disagreements (Leonardelli et al., 2023) and used Llama2-13B to generate annotations for weak and strong data perspectivism variants resulting in six (6) datasets. Strong perspectivism used prompts tailored to individual persona descriptions, while weak perspectivism used group descriptions to simulate aggregated viewpoints; however, the persona descriptions in each variant were limited to the demographic information and features provided in the original work. All datasets use binary labels for classification. Original dataset statistics are presented in Table 1.

HS-Brexit The Hate Speech Brexit (HS-Brexit) dataset (Akhtar et al., 2021) comprises 1,120 tweets concerning Brexit and immigration, annotated for hate speech, aggressiveness, and offensiveness. This dataset features annotations from two distinct groups of three individuals: a target group of Muslims and first- or second-generation immigrants to the UK (also classified as migrants in the original study) and a control group of researchers with a Western background making six annotators in all.

ConvAbuse The Conversational Abuse (ConvAbuse) dataset, as described by Cercas Curry et al. (2021), comprises roughly 4,000 English dialogues between users and two conversational agents. These user conversations were labeled by a minimum of three gender studies experts, using a hierarchical annotation system that included categories for presence, severity, and directness of abuse. We binarized the annotations into two classes, 0 and 1. The ConvAbuse dataset is characterised by eight (8) annotators, each providing a significant number of annotations. Also, not all the 8 annotators labeled every instance contrary to the HS Brexit, but each annotator has annotations.

4 Methodology

Firstly, we explore the ability of Llama2-13B to generate discrete binary annotations on the datasets, using defined personas. Secondly, we modeled these personas with existing annotator modeling techniques.

4.1 Annotation Generation

For the strong perspectivism variant of the datasets, we prompt Llama2-13B with each text in the original corpus. We extended the dataset with the generated annotations for each corresponding persona, maintaining the original structure of the dataset from the SemEval-2023 task. The *strong* variant uses specific individual descriptions for each persona as seen in Figure 1. In the original ConvAbuse dataset, not all annotators annotated all instances, but in the LLM version, all eight annotators were represented in all instances. We generate annotations at temperatures: 0, 0.1, 0.2, 0.5 and 0.8, for each perspectives. We used the demographic description presented in the original work as guide for our persona features. In *weak* perspectivism, we followed the same approach. Figures 3 and 4 show persona descriptions and Table 2 shows a sample of the prompt used. The prompt and personas are fully described in the Appendices A and B, respectively. Also in Table 3, we show a summary of the data statistics and the variance observed in the inter-annotator agreement K-α as temperature increases.

4.2 Annotator Modeling

We trained existing annotator models (Oluyemi et al., 2024; Davani et al., 2022) using the LLM-generated labels, following a classification pipeline originally used with the human-annotated corpus. We replicated these annotator modeling techniques—User Token, Composite Embedding, Composite+User Token Embedding, and Multi-task to model perspectives by modeling annotators, and we also added a text-only implementation without annotator information with SBERT. These techniques used annotator IDs and text, with 6 annotations per instance for the HSBrexit and 8 annotations per

	#A	#I	N	A/I	K-α
HS-Brexit	6	1,120	1,120.00± 0.00	6.00± 0.00	0.35
ConvAbuse	8	4,050	1,521.00± 206.91	3.00± 0.88	0.65

Table 1: Original Dataset Statistics by Human Annotators. #A: number of annotators, #I: number of total instances, N: number of annotations per annotator, A/I: annotations per instance, K-α: Krippendorff's alpha agreement.

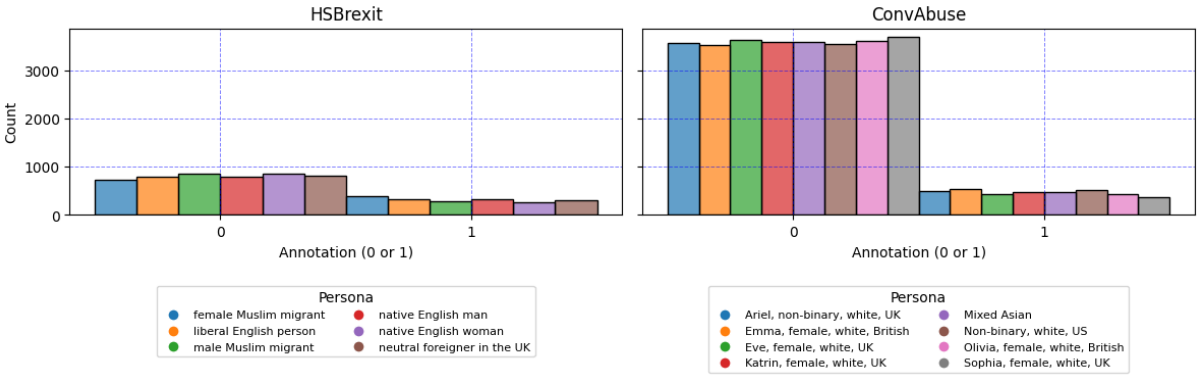


Figure 1: Distribution of Annotations for ConvAbuse and HS-Brexit (Strong Perspective) by Persona

annotate this statement regarding Brexit. as a group of (persona) in the UK or Europe, you must assign a label to the statement Use '0' if the statement does not express hate towards Brexit, and '1' if it does express hate towards Brexit. Provide only the number as your response without any additional text or explanation for example "####Annotator:"####Annotator:0" or "####Annotator:1"

Table 2: An example prompt for weak perspectivism in HS-Brexit

instance for the ConvAbuse to capture the *persona* perspectives using the labels obtained from the generations at all temperatures but used the best scores (generally between temperatures 0 and 0.1) in our results and analysis. The annotator ID represents each unique annotator(persona), encoded as embeddings. Each technique uses a different method to generate encodings used to uniquely model the personas. Finally, we compared the performance of these annotator modeling techniques on LLM-generated annotations and human annotations.

5 Results

Table 4 presents the F1 scores for the annotator modeling techniques evaluated on both the original and the synthetic datasets. Our analyses show some trends in the performance of these models. In existing results that used human annotations, overall performance was observed on the ConvAbuse dataset. The inter-annotator agreement measured by Krippendorff's alpha was high for ConvAbuse and comparatively lower for the HSBrexit dataset. Interestingly, the Llama2-annotated versions showed significantly higher agreement levels than the original human annotations across all temperature settings, including at a high randomness level (Temperature = 0.8) as seen in Tables 1 and 3. Prior research established that the effectiveness of annotator modeling techniques is largely dependent on the degree of agreement and the number of annotations per annotator (Oluyemi et al., 2024). Specifically, the User-Token modeling approach performs best for datasets with low agreement, while the Composite Embedding + User Token method is optimal for datasets with high agreement. Both methods rely on an explicit naming system, using annotator IDs to individually predict the label outputs for each annotator. However, our results indicate that models without explicit annotator information outperformed others on the Llama2 persona-based datasets. For instance, SBERT, with no annotator information and Composite Embedding- an approach that did not use explicit naming convention (annotator ID) for modeling, both outperformed the best-performing models on HSBrexit and achieved comparable results on ConvAbuse. This suggests that the optimal annotator modeling techniques for human annotations may not be directly transferable or equally effective for data annotated through LLM personas.

	#A	#I	N	A/I	K-α (Strong)	K-α (Weak)
HS-Brexit	6	1,120	1,120.00± 0.00	6.00± 0.00	0.58 – 0.81 (T=0.8 – 0)	0.55 – 0.75 (T=0.8 – 0)
ConvAbuse	8	4,050	4,050.00± 0.00	8.00± 0.00	0.60 – 0.91 (T=0.8 – 0)	0.62 – 0.93 (T=0.8 – 0)

Table 3: LLAMA2 Dataset Statistics. #A: number of annotators, #I: number of total instances, N: number of annotations per annotator, A/I: annotations per instance, K-α: Krippendorff's alpha agreement (T=temperature range). The K-α values are presented as a range from temperature 0.8 to 0, that is agreement decreases as temperature increases.

Method	SBERT	User Token	Composite Embedding	Composite Embedding + User Token	Multi-Tasking
Human-annotations					
HS-Brexit	68.6	77.6	67.6	77.3	71.7
ConvAbuse	85.9	88.5	85.8	88.6	82.3
LLAMA2-13B strong perspectivism					
HS-Brexit	**72.2**	69.4	**71.8**	71.2	65.1
ConvAbuse	85.7	84.4	84.6	84.4	81.1
LLAMA2-13B weak perspectivism					
HS-Brexit	**73.2**	72.2	**72.4**	71.7	62.0
ConvAbuse	85.2	83.7	83.7	81.8	79.8

Table 4: Model performance based on individual annotator and persona F1 scores. Results for human annotations was adapted from Oluyemi et al. (2024). We reported the best LLM results for temperatures 0 and 0.1.

5.1 Strong vs Weak Data Perspectivism in Annotator Modeling

As presented in Tables 6 and 7 of Appendix C, we adapted the two versions of data perspectivism described by (Cabitza et al., 2023) and evaluated the annotator modeling techniques on the datasets. The strong perspectivist approach, which used fine-grained persona profiles, generally produced higher performance that was more aligned with the results from human modeling for the ConvAbuse dataset at temperature 0.1. The weak perspectivism approach, characterized by contrasting group descriptions, showed improved performance over the human version in the HS-Brexit dataset across both strong and weak variants, with a greater improvement observed in the weak, group-based variant. However, this performance increase was exclusively observed in the Composite Embedding and SBERT models without explicit annotator information.

5.2 Annotation Quality and Uncertainty

We analyzed the quality of annotations generated by Llama2-13B across a spectrum of temperature parameters. Even at high randomness with temperature set to 0.8, inter-annotator agreement remained high cf. Table 3. The distribution of labels diverged significantly from that of the human annotators. To illustrate this, we compared the label distributions using Probability Density Functions (PDFs). The human annotations showed a sharp peak near class 0, indicating a highly consistent assignment of non-abusive class, despite disagreement, in the HS-Brexit dataset as seen in Figure 2. In contrast, the PDF for the strong perspectivist variant of the LLM showed a slightly right-skewed peak between 0.1 and 0.2, suggesting that the LLM assigned marginally higher soft labels than human annotators. The weak perspectivist PDF was flatter and more dispersed, with a small density spike near a probability of 0.2, reflecting greater uncertainty and inconsistency in labeling. The PDFs for the ConvAbuse dataset is presented in the Appendix D.

5.3 Prototypical *Persona* Annotators and Human Alignment

Ablation 1: Table 5 shows that annotator models trained on LLM annotations perform worse when tested on human labels, indicating a lack of alignment. The decline likely comes from the lack of corresponding match between LLM personas and the unknown individual human annotators.

Ablation 2: Figures 3 and 4 present an alignment analysis between LLM personas and human annotators. We compute cosine similarity between their annotation vectors. Using sample sizes of 5, 10, 50, and 100, stronger alignment was observed at sizes 50 and 100. In the ConvAbuse *strong* vari-

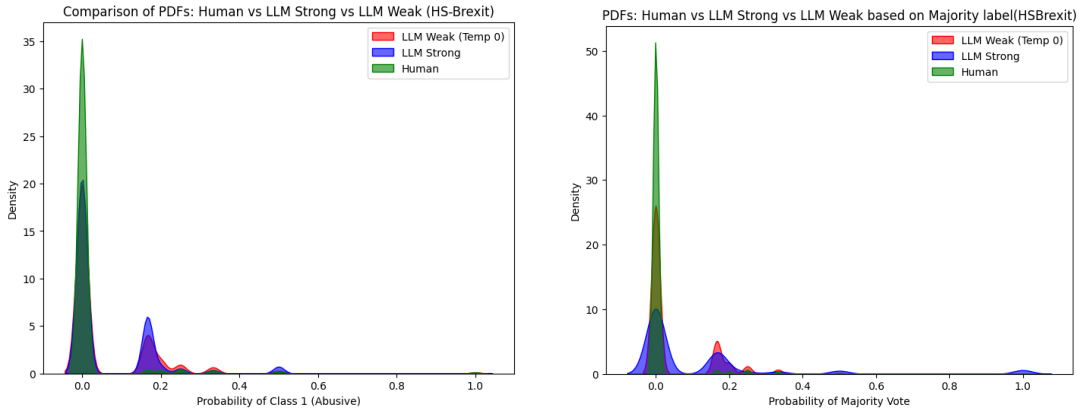
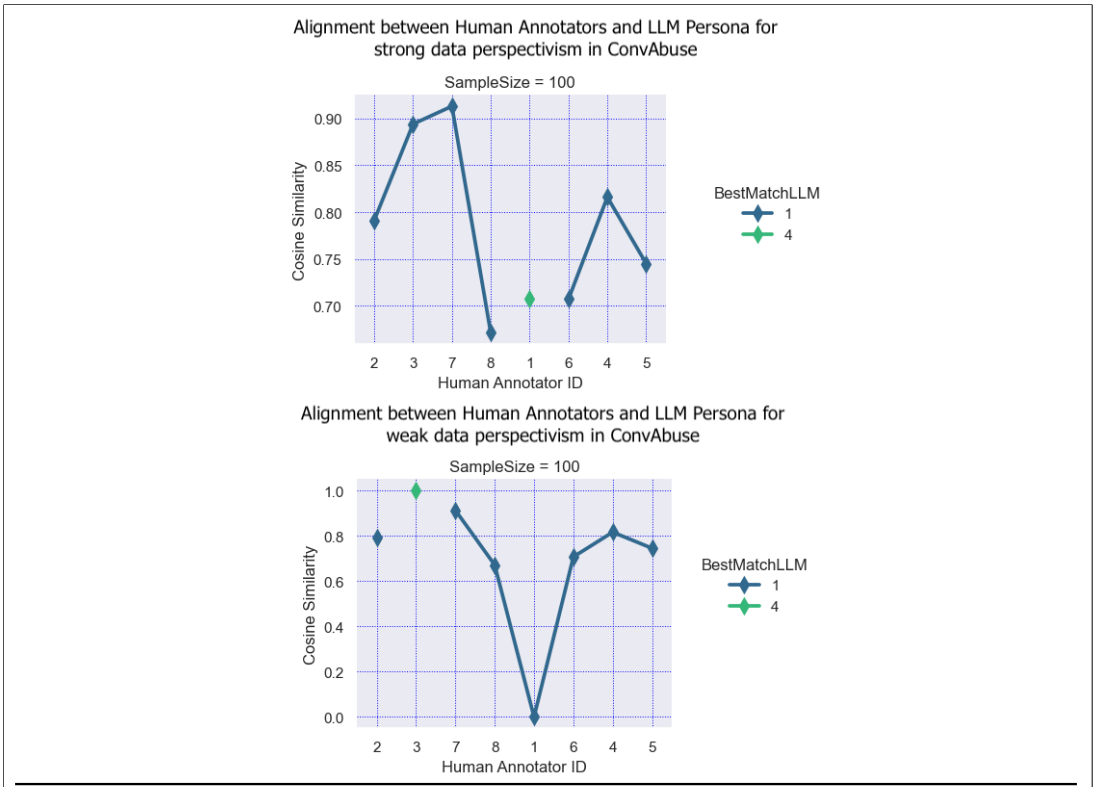


Figure 2: Figure showing the Probability Density Function illustrating Uncertainty in LLM annotations Vs Human in HSBrexit



Human/LLM ID	Persona (strong)	Persona (weak)
1	Olivia, a female and white British person	white British female people
2	Emma, a female and white British person	white British with non-binary gender orientation
3	Ariel, a white person from the United Kingdom with a non-binary gender orientation	non-binary gender people from the United States
4	Sophia, a female and white person from the United Kingdom	white female people from the United Kingdom
5	Katrin, a female and white person from the United Kingdom	white female from United States
6	Eve, a female and white person from the United Kingdom	mixed Asian with a non-binary gender orientation
7	mixed Asian person	mixed Asian female
8	a white person from the United States with a non-binary gender orientation	mixed Asian female

Figure 3: Figure showing Prototypical LLM annotators and Alignment with Human Annotators in ConvAbuse

ant, ANN(2–8) showed varying degrees of alignment with LLM Persona 1 (Olivia, female, white, British), while ANN(1) aligns more closely with LLM Persona 4 (Sophia, female, white, from the UK). Other LLM personas (2, 3, 5–8) exhibit no correspondence with any human annotator. We further trained annotator models on annotations from LLM Personas 1 and 4, and evaluated them against human-labeled data. These models showed improved performance, approaching human-level results for both Composite Embedding and SBERT, as shown in Table 5.

Model	SBERT	User Token	Composite Embedding	Composite Embedding + User Token
HL	85.9	88.5	85.8	88.6
LLM	85.7	84.4	84.6	84.4
LLM-H	83.1	83.4	84.5	84.2
LLM(1,4)-H	85.4	82.6	85.1	85.9

Table 5: Model performance based on different training and testing label splits: HL (models trained and tested on Human Labels), LLM (models trained and tested on LLM Labels), LLM-H (models trained on LLM Labels, tested on Human Labels), and LLM(1,4)-H (models trained on the most aligned LLM personas 1 and 4 to human labels, tested on Human Labels).

In the HS-Brexit dataset, alignment is less consistent. In Figure 4, we see Persona 1, Male Muslim migrant, belonging to the *target* group mapped to annotators 4 and 5 of the human annotators belonging to the *control* group in the strong variant. Human annotators 1–3 belong to the Muslim or migrant group, while annotators 4–6 belong to the group with Western background, denoted as *locals*. Also, Persona 3 of the migrant group representing "neutral foreigner" shows positive alignment in the *weak* variant to the migrant group in human when "Muslim" was removed. These findings suggest that Llama2 includes prototypical personas capable of partially representing multiple human annotators. However, other defined personas fail to map to any observed human annotation patterns (cf. Appendix E).

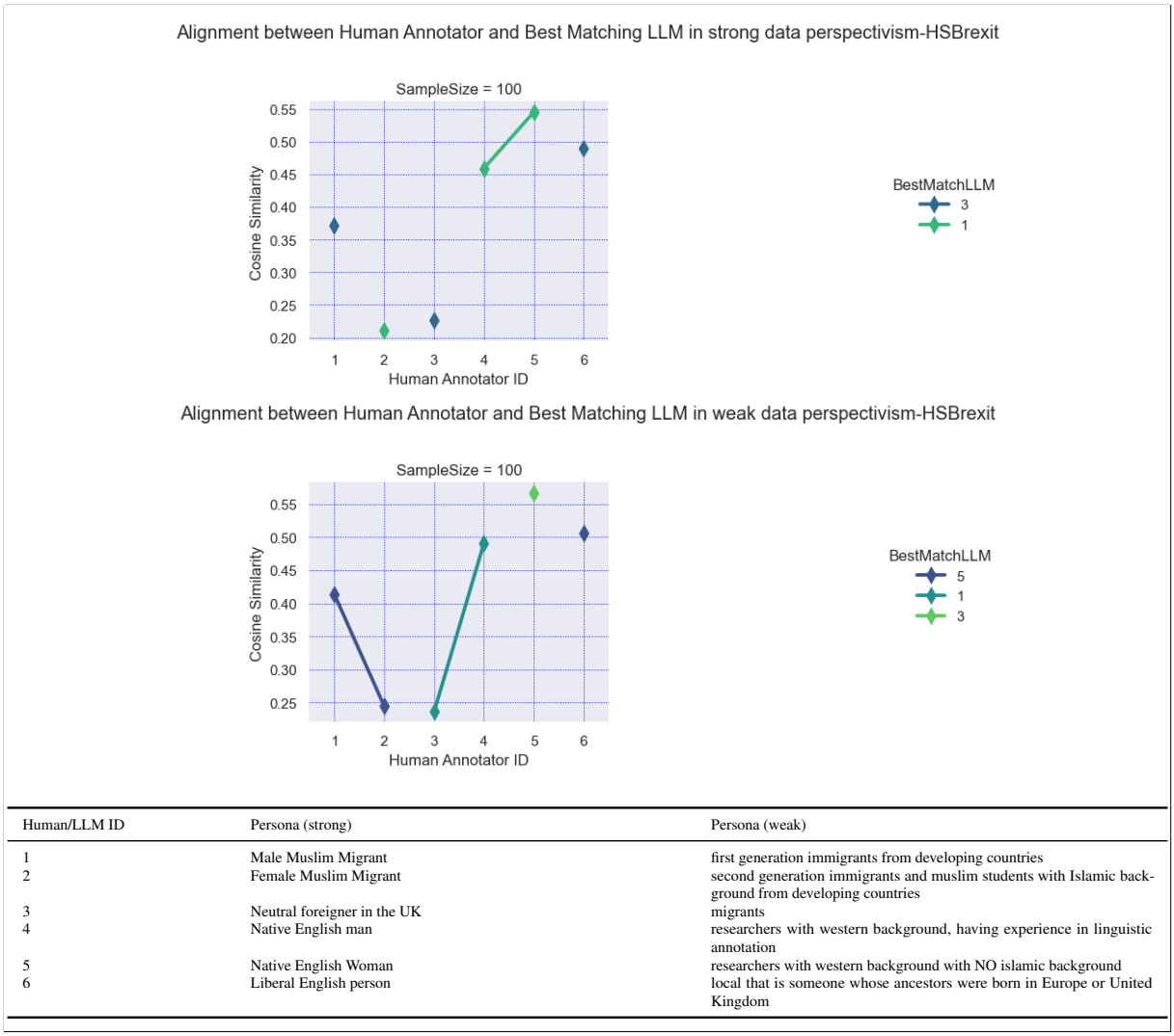
6 Discussion and Conclusion

This work investigates Llama2's capacity to generate disaggregated labels for hate speech and offensiveness datasets using predefined personas, under two perspectivism frameworks: strong (individual) and weak (group) data perspectivism. We examine the quality and alignment of LLM-generated annotations with human-annotated datasets and evaluate downstream performance across existing annotator modeling techniques.

Llama2 annotations consistently exhibited higher inter-annotator agreement (Krippendorff's alpha ranging 0.55–0.91) than human annotations across both ConvAbuse and HS-Brexit datasets, though agreement decreased at higher temperatures. PDF analysis further indicated that LLM annotations tend to converge around features inherent in the model's underlying corpus, suggesting a divergence from human perspectives. As seen in Figure 2, the PDF using the soft label distribution of the abusive class shows human annotations aligning towards the non-abusive class, strong perspectivism aligning more towards the abusive class, and weak perspectivism showing a relatively flat and dispersed distribution depicting high uncertainty.

In terms of performance of annotator modeling methods, LLM annotations shifted model efficacy. While prior work confirmed that annotator models trained on human-annotated datasets with high agreement (e.g., ConvAbuse) performed best with the Composite Embedding + User Token model, and those with low agreement (e.g., HS-Brexit) favored the User Token model, our findings with LLM-generated annotations demonstrate that simpler models, specifically SBERT and Composite Embedding models without explicit annotator information, showed improved results. This shift implies that LLM-generated annotations align more with generalized perspectives and are less suited to highly personalized approaches. Comparing the two perspectivism approaches, strong data perspectivism on ConvAbuse, characterized by overlapping and more personalized features, improved the performance of annotator modeling techniques over its weak counterpart. Conversely, weak perspectivism on HS-Brexit, with its contrasting demographic features in groups, yielded improved performance specifically with SBERT and Composite Embedding models, suggesting that contrasting demographic diversity tends to influence the choice of perspectivism approach and annotator modeling performance in LLMs.

Our ablation studies revealed LLM personas do not directly correspond to human annotators. However, as seen in Figure 3, we identified generalized "prototypical persona features" working as representatives of groups of humans (e.g., ANN 2-8 mapping to LLM Persona 1, ANN1 to LLM Persona 4). Swapping the labels of corresponding annotators in the original dataset with these prototypical annotator labels, and evaluating with the human test set, slightly improved results, as seen in Table 5, presenting a novel approach for modeling perspectivism in LLMs. These findings suggest that while



Human/LLM ID	Persona (strong)	Persona (weak)
1	Male Muslim Migrant	first generation immigrants from developing countries
2	Female Muslim Migrant	second generation immigrants and muslim students with Islamic background from developing countries
3	Neutral foreigner in the UK	migrants
4	Native English man	researchers with western background, having experience in linguistic annotation
5	Native English Woman	researchers with western background with NO islamic background
6	Liberal English person	local that is someone whose ancestors were born in Europe or United Kingdom

Figure 4: Figure showing Prototypical LLM annotators and Alignment with Human Annotators in HSBrexit

LLMs offer insights into subjective domains, their capacity to fully embody external personas remains limited to their underlying corpus, supporting an aggregated view rather than personalization. Future work should focus on standardization and generate more diversified personas, systematically varying features, and expanding evaluation to other LLMs to fully investigate these prototypical attributes and their potential in capturing a wider scope of perspectives.

7 Limitations

This study is based on two datasets and focuses exclusively on binary classification tasks for hate and offensive speech detection. One potential limitation is that the data used to train Llama2-13B may have been filtered, reducing its sensitivity to detecting abusive content, potentially influencing the observed results. Our analysis is also limited to this model, and we did not investigate how newer variants of Llama or other LLMs, like GPT 4o, might influence the results. The personas used for generating annotations were limited to the demographic features explicitly provided in the original datasets, with slight modifications to fit the perspectivist spectrum. Furthermore, we did not quantify the extent to which the model's attention was distributed between the persona and the input sentences. Understanding this balance could provide deeper insight into how strongly LLMs personalize their annotations.

Another limitation of this study arises from the design of the annotation prompt for the HS-Brexit dataset variant, which focused on 'hate speech towards Brexit'. However, the prompt was structured to provide general contextual information about Brexit and simulate the prior knowledge of human annotators. A follow-up experiment analysing the

model's attention mechanism revealed that the instance of "Brexit" appearing first in the prompt received a significantly higher attention score of 0.0654 than the "Brexit" label target which received an attention score of 0.0050. Furthermore, when 'immigrants' was targeted instead, it received an attention score of 0.0117, which was higher than that given to 'Brexit' as a target. This suggests that the models have learned to recognise plausible targets for hate speech, which warrants further investigation. However, this paper's specific focus is to investigate the impact of Annotator Personas on LLM behaviour across the perspectivism spectrum. It therefore does not include a deep analysis of the model's sensitivity to target plausibility. Nevertheless, we present this as a compelling avenue for future research, while maintaining that our core findings regarding persona-driven perspectivism remain valid within the described experimental setup. Our codes are publicly available[1] to support future work.

Acknowledgments

This research was supported by funding from Hessian.AI. Any opinions, findings, conclusions, or recommendations in this material are those of the authors and do not necessarily reflect the views of Hessian.AI.

References

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 *International Conference on Machine Learning, ICML 2023*, 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. *Preprint*, arXiv:2106.15896.

Dina Almanea and Massimo Poesio. 2022. Armis - the arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291. European Language Resources Association.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *Preprint*, arXiv:2406.18403.

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. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Thirty-Seventh AAAI Conference on Artificial Intelligence*, pages 6860–6868. AAAI Press.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. Convabuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational ai. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403. Association for Computational Linguistics.

Xia Cui. 2023. xiacui at semeval-2023 task 11: Learning a model in mixed-annotator datasets using annotator ranking scores as training weights. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1076–1084. 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.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054. Association for Computational Linguistics.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292. Association for Computational Linguistics.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide

[1] https://doi.org/10.5281/zenodo.16744588

Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*.

José Antonio García-Díaz, Ronghao Pan, Gema Alcaráz-Mármol, María José Marín-Pérez, and Rafael Valencia-García. 2023. Umuteam at semeval-2023 task 11: Ensemble learning applied to binary supervised classifiers with disagreements. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1061–1066. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel J. Candès, and Dan Jurafsky. 2024. Can unconfident llm annotations be used for confident conclusions? *Preprint*, arXiv:2408.15204.

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

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in llm simulations. *Preprint*, arXiv:2402.10811.

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

Jonathan Ivey, Shivani Kumar, Jiayu Liu, Hua Shen, Sushrita Rakshit, Rohan Raju, Haotian Zhang, Aparna Ananthasubramaniam, Junghwan Kim, Bowen Yi, Dustin Wright, Abraham Israeli, Anders Giovanni Møller, Lechen Zhang, and David Jurgens. 2024. Real or robotic? assessing whether llms accurately simulate qualities of human responses in dialogue. *Preprint*, arXiv:2409.08330.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Cardenas, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and Morteza Dehghani. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108.

Matej Klemen and Marko Robnik-Šikonja. 2022. Ulfri at semeval-2022 task 4: Leveraging uncertainty and additional knowledge for patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 525–532. Association for Computational Linguistics.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. Semeval-2023 task 11: Learning with disagreements (lewidi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318. Association for Computational Linguistics.

Sarumi Oluyemi, Béla Neuendorf, Joan Plepi, Lucie Flek, Jörg Schlötterer, and Charles Welch. 2024. Corpus considerations for annotator modeling and scaling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1029–1040. 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. ELRA and ICCL.

Maja Pavlovic and Massimo Poesio. 2024b. Understanding the effect of temperature on alignment with human opinions. *Preprint*, arXiv:2411.10080.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. 2022. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402. Association for Computational Linguistics.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective nlp tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190. Association for Computational Linguistics.

Michael Sullivan, Mohammed Yasin, and Cassandra L. Jacobs. 2023. University at buffalo at semeval-2023 task 11: Masda–modelling annotator sensibilities through disaggregation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 978–985. Association for Computational Linguistics.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with llms. *Preprint*, arXiv:2311.09730.

Nikolas Vitsakis, Amit Parekh, Tanvi Dinkar, Gavin Abercrombie, Ioannis Konstas, and Verena Rieser.

2023. ilab at semeval-2023 task 11 le-wi-di: Modelling disagreement or modelling perspectives? In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1660–1669. Association for Computational Linguistics.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2024. Large language models should not replace human participants because they can misportray and flatten identity groups. *Preprint*, arXiv:2402.01908.

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. Association for Computational Linguistics.

Appendix

A Prompt Design

Prompt for Strong HS-Brexit *annotate this statement regarding Brexit. As a (persona) in the UK, you must assign a label to the statement. Use 'O' if the statement does not express hate towards Brexit, and '1' if it does express hate towards Brexit. Provide only the number as your response without any additional text or explanation for example "###Annotator:0" or "###Annotator:1"*

Prompt for Weak HS-Brexit *annotate this statement regarding Brexit. for a group of (persona) in the UK or Europe, you must assign a label to the statement. Use 'O' if the statement does not express hate towards Brexit, and '1' if it does express hate towards Brexit. Provide only the number as your response without any additional text or explanation for example "### Annotator:0" or "###Annotator:1"*

Prompt for Weak ConvAbuse *annotate these conversations between users and bots. As a group of persona within the age range of 19-21, a level 1 English speaker and an Undergraduate student in Gender Studies and Sociology, you must assign a label to the statement. Use 'O' if the statement is not abusive, and '1' if it does express abusiveness by the user towards the bot or agent. Provide only the number as your response without any additional text or explanation. for example "### Annotator:0" or "### Annotator:]"*

Prompt for Strong ConvAbuse *annotate these conversations between users and bots. As a persona within the age range of 19-21, a level 1 English speaker and an Undergraduate student in Gender Studies and Sociology, you must assign a label to the statement. Use 'O' if the statement is not abusive, and '1' if it does express abusiveness by the user towards the bot or agent. Provide only the number as your response without any additional text or explanation. for example "### Annotator:0" or "### Annotator:]"*

B Persona Descriptions

HS-Brexit Persona for Strong Perspectives

- Male Muslim Migrant
- Female Muslim Migrant
- Neutral foreigner in the UK
- Native English man
- Native English Woman
- Liberal English person

HS-Brexit Persona for Weak Perspectives

- researchers with Western background having experience in linguistic annotation
- first or second generation muslim immigrant students from developing countries

ConvAbuse Persona for Weak Perspectives

- white British female people
- white British with non-binary gender orientation
- non-binary gender people from the United States
- white female people from the United Kingdom
- white female from United States
- mixed Asian with a non-binary gender orientation
- mixed Asian female
- white people from the United States with a non-binary gender orientation

ConvAbuse Persona for Strong Perspectives

- Olivia, a female and white british person
- Emma, a female and white british person
- Ariel, a white person from the United Kingdom with a non-binary gender orientation
- Sophia, a female and white person from the United Kingdom
- Katrin, a female and white person from the United Kingdom
- Eve, a female and white person from the United Kingdom
- a mixed Asian person
- a white person from the United States with a non-binary gender orientation

C Model performance for Strong and Weak Data Perspectivism

Model	α	User-Token	Composite	Composite+ User-Token	Multitasking	SBERT
Strong Perspectivism						
Human	0.65	88.5	85.8	88.6	82.3	85.9
0	0.91	84.1	83.0	84.4	46.9	83.1
0.1	0.87	84.4	**84.6**	84.3	81.1	**85.7**
0.2	0.81	80.5	81.5	80.0	46.8	81.5
0.5	0.68	69.9	70.7	71.1	45.1	69.4
0.8	0.60	63.5	65.1	64.4	62.6	64.6
Weak Perspectivism						
0	0.93	83.7	83.7	81.8	69.0	85.2
0.1	0.88	80.1	79.3	78.3	79.8	82.0
0.2	0.82	81.2	81.5	81.2	70.3	82.1
0.5	0.67	71.7	69.7	71.4	64.7	69.5
0.8	0.62	61.7	61.4	62.3	58.1	61.4

Table 6: Performance of Annotator modeling methods for Strong and Weak data Perspectivism **(ConvAbuse dataset)** across various temperatures.

Model	α	User-Token	Composite	Composite+ User-Token	Multitasking	SBERT
Strong Perspectivism						
Human	0.35	77.6	67.6	77.3	71.7	68.6
0	0.81	69.3	71.3	71.2	65.1	72.2
0.1	0.73	69.4	71.8	71.0	61.8	69.2
0.2	0.67	66.3	63.8	61.9	61.4	67.2
0.5	0.62	61.5	61.3	61.4	49.5	62.2
0.8	0.58	52.4	56.1	54.2	51.2	56.6
Weak Perspectivism						
0	0.75	72.2	**72.4**	71.7	60.3	**73.2**
0.1	0.69	66.6	65.8	65.5	62.0	69.1
0.2	0.62	62.2	63.8	69.9	59.2	66.8
0.5	0.54	58.0	58.4	57.9	39.2	56.1
0.8	0.55	55.2	57.8	56.7	55.4	56.6

Table 7: Performance of Annotator modeling methods for Strong and Weak data Perspectivism **(HS-Brexit dataset)** across various temperatures.

D Probability Density Function for Uncertainty and Annotation Quality

The Figure 5 shows the probability density function of the weak data perspectivism in ConvAbuse using the majority class as a reference point.

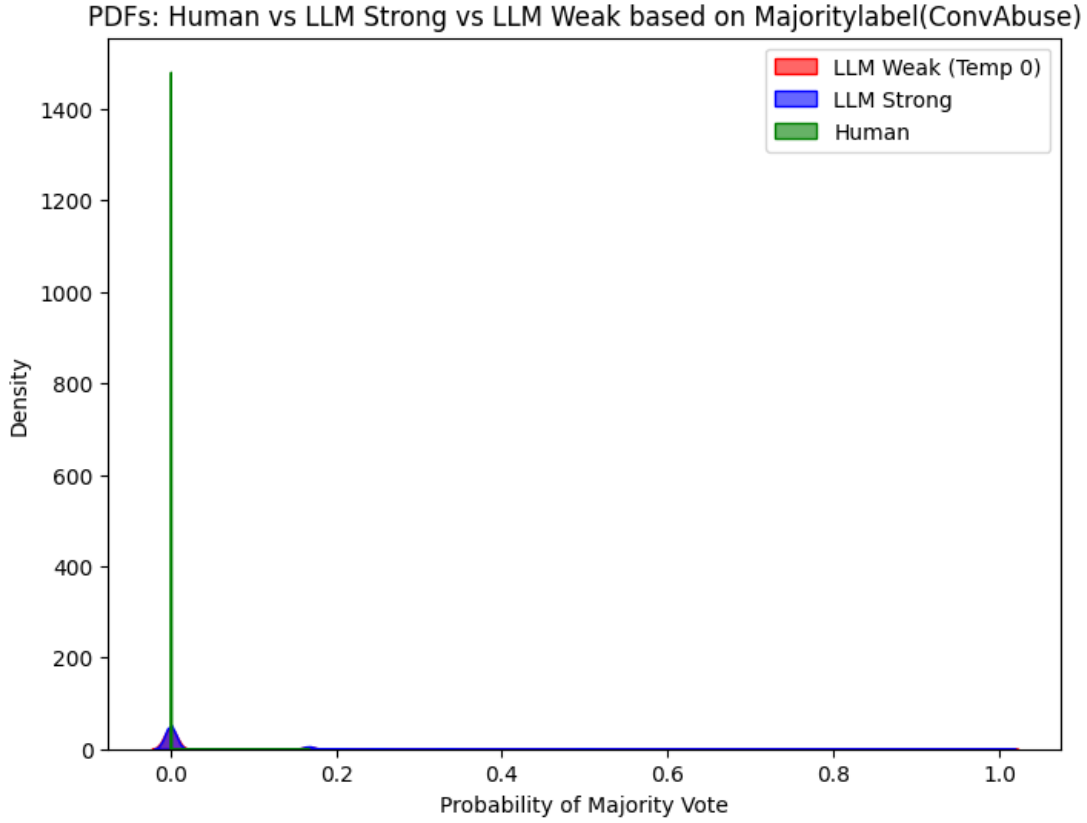


Figure 5: Probability Density Function ConvAbuse Dataset

E Human Vs Persona Alignment and Prototypes ConvAbuse Dataset

Human Annotator	SampleSize	Best Match LLM Persona	Similarity score
2	100	1	0.791
3	100	1	0.894
7	100	1	0.913
8	100	1	0.671
1	100	4	0.707
6	100	1	0.707
4	100	1	0.816
5	100	1	0.745

Table 8: Mapping of Human Annotators to Best Matching LLM Personas based on Cosine Similarity.

Prototypical Annotators and their Alignment with Human Annotators Across Varying Sample sizes

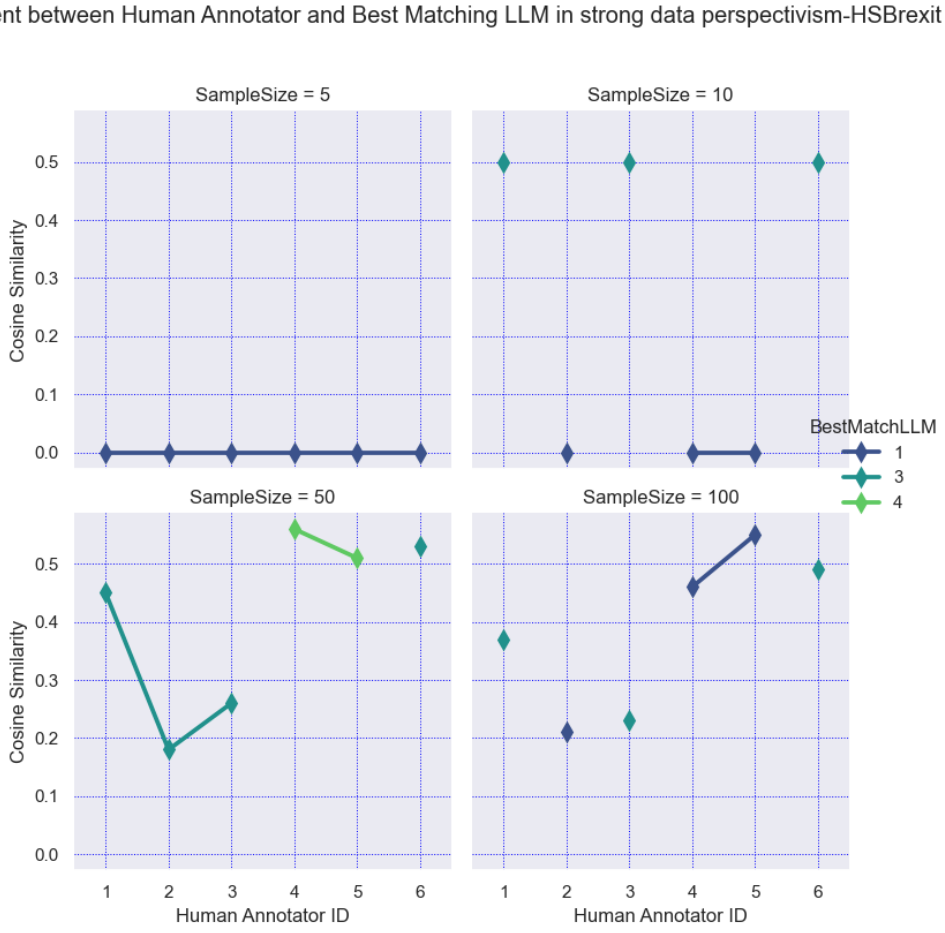


Figure 6: Showing the identified Prototypical annotators in HS-Brexit dataset and the alignment with human annotators

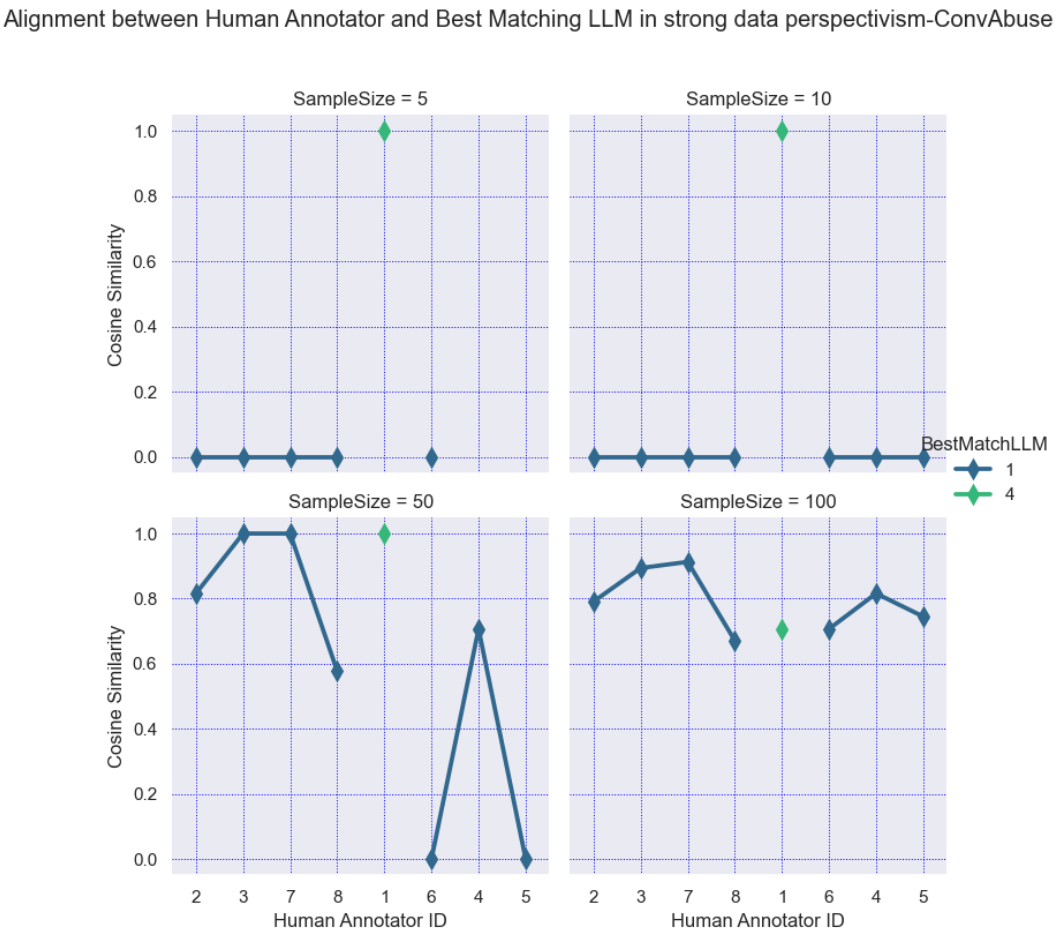


Figure 7: Showing the identified Prototypical annotators in ConvAbuse dataset and the alignment with human annotators