Hateful Person or Hateful Model? Investigating the Role of Personas in Hate Speech Detection by Large Language Models

Shuzhou Yuan[*1], Ercong Nie[*2,3], Mario Tawfelis[*1],
Helmut Schmid[2], Hinrich Schütze[2,3] and Michael Färber[1]
[1]ScaDS.AI and TU Dresden [2]LMU Munich
[3]Munich Center for Machine Learning (MCML)
shuzhou.yuan@tu-dresden.de, nie@cis.lmu.de

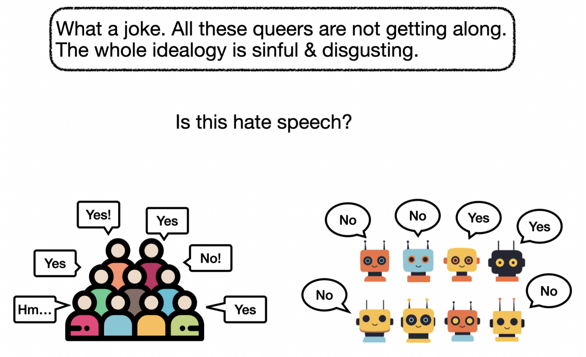
Abstract

Content Warning: *This paper contains examples of hate speech, which may be disturbing or offensive to some readers.*

Hate speech detection is a socially sensitive and inherently subjective task, with judgments often varying based on personal traits. While prior work has examined how sociodemographic factors influence annotation, the impact of personality traits on Large Language Models (LLMs) remains largely unexplored. In this paper, we present the first comprehensive study on the role of persona prompts in hate speech classification, focusing on MBTI-based traits. A human annotation survey confirms that MBTI dimensions significantly affect labeling behavior. Extending this to LLMs, we prompt four open-source models with MBTI personas and evaluate their outputs across three hate speech datasets. Our analysis uncovers substantial persona-driven variation, including inconsistencies with ground truth, inter-persona disagreement, and logit-level biases. These findings highlight the need to carefully define persona prompts in LLM-based annotation workflows, with implications for fairness and alignment with human values.

1 Introduction

The proliferation of hate speech on online platforms poses a persistent threat to inclusive digital spaces, necessitating robust and scalable detection systems (Röttger et al., 2021; Fortuna et al., 2022). Traditionally, the effectiveness of automated hate speech detectors has hinged on the quality of human-labeled data (Kim et al., 2022; Yuan et al., 2022). However, hate speech annotation is inherently subjective: as shown in Figure 1, what one annotator deems hateful, another may consider benign, with disagreement often rooted in personal beliefs, backgrounds, and identities (Sap et al., 2022; Fleisig et al., 2023; Astorino et al., 2023). Recent research has highlighted that sociodemographic factors such as age, gender, and cultural background significantly influence how annotators perceive and label hate speech, leading to biases that can propagate into downstream models (Mostafazadeh Davani et al., 2022; Wang and Plank, 2023).

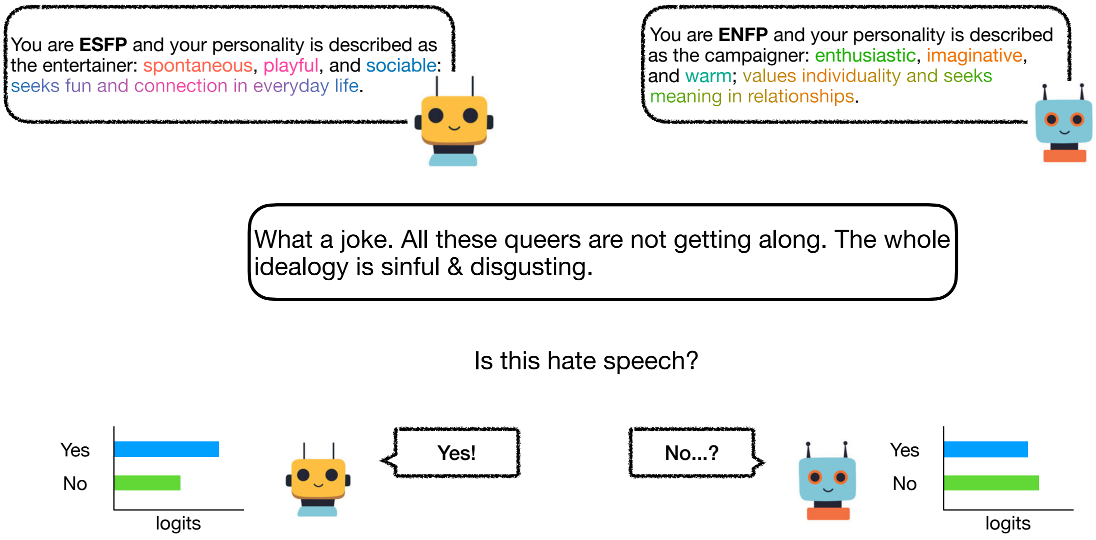

What a joke. All these queers are not getting along. The whole idealogy is sinful & disgusting.

Is this hate speech?

Figure 1: Our study demonstrates that different personas influence hate speech detection for both humans and large language models.

Large Language Models (LLMs) can be prompted to perform a wide range of classification, generation, and evaluation tasks (Zhang et al., 2023; Yang et al., 2024b; Wu et al., 2024). Recent work has demonstrated that LLMs, when provided with clear task instructions, are capable of matching or even surpassing the average accuracy of human annotators in a variety of labeling tasks, including hate speech detection, natural language inference, and sentiment analysis (Tan et al., 2024; Movva et al., 2024; Horych et al., 2025). As a result, *LLM-as-a-annotator* paradigms are being actively explored as alternatives or supplements to traditional human labeling pipelines (He et al., 2024). However, recent studies question the reliability of LLM annotations and indicate that LLMs can exhibit significant annotation variation (Borah and Mihalcea, 2024; Gligoric et al., 2025). To systematically

[*] Equal contribution.



You are **ESFP** and your personality is described as the entertainer: spontaneous, playful, and sociable: seeks fun and connection in everyday life.

You are **ENFP** and your personality is described as the campaigner: enthusiastic, imaginative, and warm; values individuality and seeks meaning in relationships.

What a joke. All these queers are not getting along. The whole idealogy is sinful & disgusting.

Is this hate speech?

Figure 2: Assigning different MBTI-based personas to the LLM results in varied responses in hate speech detection. The predicted logits reveal clear differences in classification tendencies and confidence across personas. Notably, even a single personality dimension difference between ESFP and ENFP leads to divergent model outputs.

study and control this variation, researchers have introduced persona-based prompting (Zhu et al., 2025), a technique that conditions LLMs to emulate specific annotator identities, backgrounds, or psychological profiles (Deshpande et al., 2023; Joshi et al., 2024). For example, LLMs can be prompted to annotate text as if they were a “young adult from Europe”, a “conservative American”, or a “first-generation immigrant”, allowing researchers to probe how such identities influence model behavior (Joshi et al., 2024; Hu and Collier, 2024; Liu et al., 2024). This approach not only helps to audit for biases and blind spots within LLMs but also offers a pathway to generate more diverse and representative annotation data (Fröhling et al., 2024; Prpa et al., 2024).

Despite its promise, most persona-based LLM research has focused on socio-demographic factors (e.g., age, gender, geographic background), with limited attention given to more fundamental and personally subjective traits such as personality (Orlikowski et al., 2023; Masud et al., 2024; Giorgi et al., 2024). In psychology and the social sciences, personality frameworks like the Myers–Briggs Type Indicator (MBTI) have long been used to explain individual differences in perception, judgment, and decision-making (Myers and Myers, 2010). The MBTI theory categorizes individuals along four dichotomous dimensions (e.g., Introversion-Extraversion, Judging-Perceiving), which together are hypothesized to shape core aspects of a person’s worldview and evaluative tendencies. In the NLP community, MBTI and similar psychological frameworks such as the Big Five Personality model (Goldberg, 2013) have been used both for personality detection from text (Plank and Hovy, 2015; Stajner and Yenikent, 2021; Li et al., 2025) and, more recently, to construct LLM personas for controlled generation and analysis (Jiang et al., 2024).

Yet, a systematic investigation of how personality-based personas influence subjective annotation tasks, such as hate speech detection, remains lacking. Most existing work treats personality as a downstream application or as a tool for behavioral simulation (Cheng et al., 2023; Lee and Ram, 2024; Choi and Li, 2024), rather than as a source of annotation variation in its own right. This raises important open questions: To what extent does personality drive annotator disagreement in subjective NLP tasks? Can LLMs, when prompted with different personality profiles, replicate or even amplify these effects? And what are the implications for the reliability, fairness, and interpretability of LLM-generated annotations?

In this work, we present the first comprehensive study investigating the influence of MBTI-based personas on hate speech detection by LLMs. We begin with a human annotation survey, in which participants report their MBTI personality types and annotate a curated set of hate speech examples. The results reveal that MBTI personality traits significantly affect individual judgments, particularly highlighting consistent differences between the *Feeling* and *Thinking* types. Building on this insight, we apply persona-based prompting to four open-source LLMs across three hate speech datasets. Our analysis shows substantial disagreement between model predictions and the original dataset labels. While each model displays distinct patterns of bias, we also observe notable variability in predictions across different personas within the same model. Further, as shown in Figure 2, a logit-level analysis reveals that even small changes in MBTI dimensions can lead to systematic shifts in model confidence and decision boundaries. When comparing human behavior with LLM outputs, we find that certain personality traits are more exaggerated in LLMs, indicating an amplification effect induced by persona prompts.

Our study leads to the following key insights:

- Human annotators with different MBTI personas produce significantly divergent hate speech annotations, and LLMs, when conditioned on persona prompts, also exhibit varied behavior.
- Persona conditioning not only affects final classification outcomes but also alters the model's internal confidence, revealing subtle but systematic influences on decision-making.
- In LLMs, analytical traits (T and J) increase confidence in hate speech predictions, contrasting with human data where Feeling types label more content as hateful.
- LLMs are more susceptible to persona framing than humans, amplifying certain behavioral traits and highlighting the risks of unintended bias in sensitive tasks.

2 Related Work

Human and LLM Annotation Variation Prior work has highlighted that human label variation in NLP tasks is not mere annotation noise, but often reflects meaningful signals such as linguistic ambiguity or subjective interpretation (Aroyo and Welty, 2013; Uma et al., 2021; Plank, 2022). Research shows that annotator disagreement is particularly pronounced in subjective tasks, motivating frameworks such as perspectivism and descriptive annotation to better capture inherent variation (Röttger et al., 2022; Cabitza et al., 2023). While many studies examine how sociodemographic factors influence annotation behavior (Fornaciari et al., 2021; Sap et al., 2022; Goyal et al., 2022), recent findings indicate that substantial individual variation remains unexplained by demographics alone (Orlikowski et al., 2023). Parallel lines of work have begun to benchmark the alignment between LLM and human annotations, focusing on group-level and pluralistic perspectives (Sorensen et al., 2024; Movva et al., 2024). However, little attention has been given to the role of psychological factors such as personality traits in shaping both human and LLM annotation variation.

Hate Speech Detection Hate speech detection has traditionally relied on two main strategies: leveraging supplementary user or annotator information and applying advanced language models fine-tuned on hate speech datasets (Nirmal et al., 2024). While utilizing user attributes or annotator traits can improve detection accuracy, such data are often difficult to obtain across platforms (Kim et al., 2022; Yin et al., 2023; del Valle-Cano et al., 2023; Waseem and Hovy, 2016). More commonly, language models like BERT and its variants are employed to generalize from large text corpora, with additional fine-tuning boosting their ability to recognize nuanced or implicit hate speech (Caselli et al., 2021; Mathew et al., 2021; Yuan et al., 2022). In this work, we focus on the subjective nature of hate speech detection by examining how LLM- and human-annotator variation, conditioned on personality traits, influences labeling patterns in this domain.

Persona-based Prompting of LLMs With the use of persona-based prompting, assigning specific roles or identities to guide model behavior, LLMs have been employed to generate human-like behavior in reasoning (Binz and Schulz, 2023; Ziems et al., 2024), role-playing (Wang et al., 2024a, 2025), social science experiments (Horton, 2023; Park et al., 2023; Wang et al., 2024b), and data annotation or evaluation (Ge et al., 2024; Dong et al., 2024). However, few studies have systematically examined how persona-based prompting, especially with psychologically grounded traits such as MBTI profiles, influences LLM annotation behavior and contributes to subjective variation in tasks like hate speech detection.

3 Human Survey: MBTI and Hate Speech

To investigate how MBTI personality types affect human judgment of hate speech, we conduct a survey and invite participants to label hate speech based on their MBTI personality.

According to the MBTI instrument, four dichotomous dimensions classify individuals as either extroverted (E) or introverted (I), sensing (S) or intuitive (N), thinking (T) or feeling (F), and judging (J) or perceiving (P) (Boyle, 1995). These four dichotomies result in sixteen possible combinations, which in turn lead to sixteen distinct personality types.[1]

3.1 Survey Design

We conduct an anonymous online survey to gain insight into how humans with different MBTI personalities perceive hate speech and offensive language. We select a subset of 20 samples from a commonly used hate speech dateset (Davidson et al., 2017). It contains 10 samples classified as hate speech and 10 samples classified as offensive language but not as hate speech in the original annotation.[2] Participants are prompted with one sample at a time, led by the question "Is this text hate speech?", and have the option to select either "Yes" or "No". After classifying all samples, participants are requested to provide information about their MBTI personality type as well as their demographic information.

3.2 Survey Results and Analysis

We distribute the survey through the university's mailing list, targeting students and staff from various academic disciplines. A total of 293 valid responses are collected, representing all 16 MBTI personality types. The survey results are compared against the ground truth labels from the original dataset, in which 50% of the samples are annotated as hate speech.

Figure 3 presents the average percentage of samples labeled as hate speech ("Yes") by each MBTI personality. All personality types show variation from the 50% hate speech rate found in the ground truth. ESFP labels the highest proportion of samples as hate speech (67.5%), followed by INFJ, ENFP, ISFJ, and INFP, all of whom label over 60% of samples as hate speech. These personalities all include the Feeling (F) trait, which may reflect a greater sensitivity to offensive content. On the other hand, ISTJ, ESFJ, and ESTP label fewer than 50% of the samples as hate speech, indicating a more lenient or conservative interpretation. Two of these types show the Thinking (T) trait, which may correspond to a more analytical or detached judgment process. *These findings suggest that individual personality traits, particularly the Feeling–Thinking dimension, systematically influence how people perceive and classify hate speech.*

4 Task Formulation for Hate Speech Classification

To investigate if LLMs also inherit the subjectivity of the judgment to hate speech from human, we conduct experiments on LLMs by giving them different persona roles. As shown in Figure 2, the LLM is defined with different persona prompt and given a potential toxic text to answer if it belongs to hate speech, the model processes the input and generates the next token as the predicted label $\hat{y}$. We constrain the output vocabulary space by instructing the model to “only answer ‘Yes’ or ‘No”’ for binary classification tasks. The predicted label is determined by selecting the token with the highest probability:

$$\hat{y} = \arg\max_{y \in V} P_{\phi}(y) \tag{1}$$

where $P_{\phi}(y)$ represents the generative probability from the LLM, and V is the vocabulary space for the labels, defined as $V = \{\text{Yes}, \text{No}\}$ in our tasks. We use the generated label word to analyze the inconsistency between the ground truth and the label generated by the LLM. Additionally, we use the logits of the generated label words ‘Yes’ and ‘No’ to analyze inconsistencies between different personas, as well as the tendency or bias toward specific dimensions.

The persona of LLMs is defined by providing the persona information in the system prompt:

[1] The description of the personas can be found in Appendix C.

[2] The selection of the samples is not random as we want to make sure the text is meaningful and not only contains slang or user tagger. The details of the survey can be found in Appendix B.

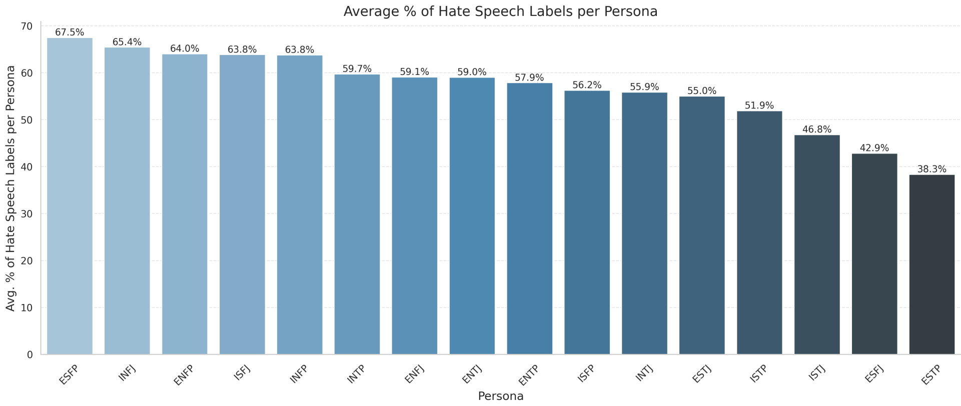


Figure 3: Average percentage of hate speech labels by 16 MBTI persona type in the human survey.

> *System*: You are $[Persona]$ and your personality is described as $[Description]$.
>
> *User*: $[Text]$ Is this text hate speech? Only answer with one word, yes or no.

where $[Persona]$ indicate the 16 personalities in MBTI, $[Description]$ is the official description about the personality, and $[Text]$ is the sample from the hate speech datasets.[3]

5 Experiments

5.1 Dataset

Dataset	Number of Samples
CREHate	365
HateXplain	1142
Davidson	20620

Table 1: Statistics of the hate speech datasets used in this research.

An overview of the datasets used in this study is provided in Table 1. We use three widely adopted hate speech datasets:

CREHate (Lee et al., 2024) The **CR**oss-cultural **E**nglish **Hate** speech dataset contains 1,580 samples which have been sourced from social media platforms. Human annotators, classifying text as hate speech or otherwise, were selected from five English-speaking countries. In our study, we use samples in which all countries unanimously classified as hate speech.

HateXplain (Mathew et al., 2021) The HateXplain dataset, compromised of around 20k samples, classifies text in three groups: hateful, offensive, normal, or undecided. Using only samples that are classified as hate speech or offensive language, we take approximately 1.1k samples for this research.

Davidson (Davidson et al., 2017) is a dataset of approximately 29k samples sourced from Twitter. Human annotators were asked to label the samples into three categories: hate speech, offensive language (not hate speech), or neither hate speech nor offensive. In our project, we only consider samples that are either classified as hate speech or offensive language, which results in approximately 20k samples.

5.2 Models and Setup

We use four open source instruction-tuned LLMs with the size of 7-8B in our experiments: `Llama-3.1-8B` (Dubey et al., 2024), `Ministral-8B` (MistralAI, 2024), `Falcon3-Mamba-7B` (Zuo et al., 2024), and `Qwen2.5-7B` (Yang et al., 2024a).[4] We set the generation temperature to 0 to keep a deterministic output with no randomness.

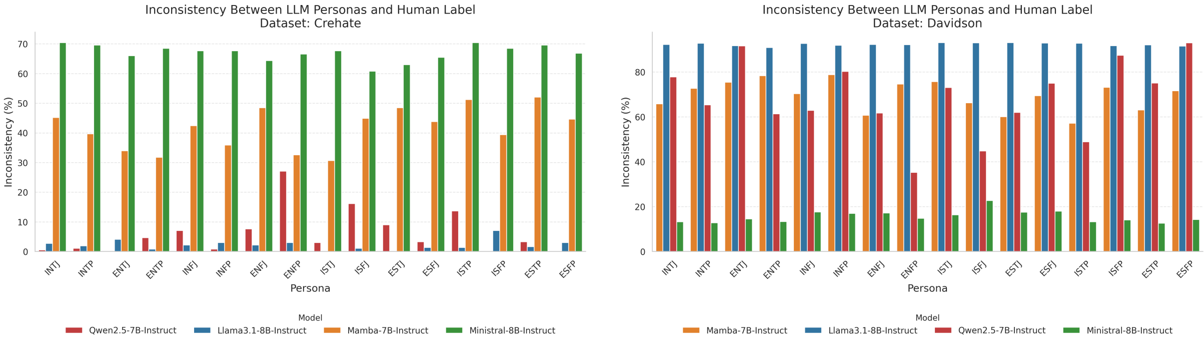
6 Results and Analysis

6.1 Inconsistency between LLM and Human Labels

The results of hate speech classification with different MBTI-based personas for four LLMs are presented in Figure 4. We report the inconsistency percentage, defined as the proportion of instances where the LLM-generated label diverges from the ground truth label. The results reveal noticeable variation across datasets, models, and persona types.

[3] The personas and the corresponding description can be found in Appendix C.

[4] The details of the LLMs can be found in Appendix A.



(a) CREHate

(b) Davidson

Figure 4: Inconsistency between LLM predictions and ground truth labels across 16 personas on three hate speech datasets for four LLMs.

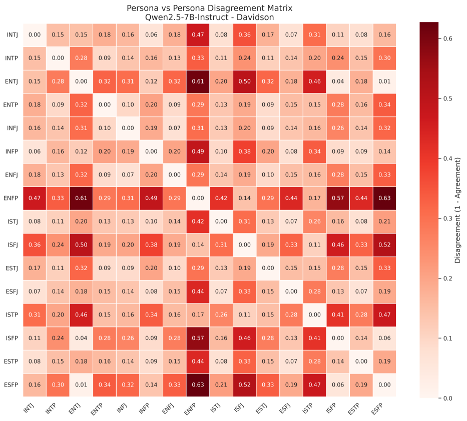
Across all three datasets, CREHate, HateXplain[5], and Davidson, we observe that the choice of LLM significantly influences inconsistency rates. For example, in the CREHate dataset (Figure 4a), `Ministral-8B-Instruct` consistently shows the highest inconsistency across all personas, often reaching 70% or higher. In contrast, `Qwen2.5-7B-Instruct` and `Llama3.1-8B-Instruct` exhibit substantially lower inconsistency, particularly for Thinking (T) and Judging (J) persona types such as ENTJ and ISTJ. `Mamba-7B-Instruct` maintains moderate inconsistency levels, typically between 30% and 50%.

The Davidson dataset (Figure 4b) reveals the highest overall inconsistency rates, with almost all models frequently exceeding 70% inconsistency across most personas. `Llama3.1-8B-Instruct` demonstrates particularly high divergence from the ground truth, while `Ministral-8B-Instruct` achieves relatively better alignment. `Qwen2.5-7B-Instruct` exhibits not only high inconsistency with the ground truth but also substantial disagreement across different personas.

Overall, these results demonstrate that both the choice of LLM and the assigned MBTI persona significantly influence hate speech classification, with the degree of impact varying across datasets.

6.2 Intra-Model Disagreement Across Personas

As the LLMs exhibit inconsistencies with the ground truth labels, we further investigate the disagreements among different MBTI personas conditioned within the same LLM. Figure 5 presents the pairwise disagreement matrix across all 16 MBTI personas for `Qwen2.5-7B-Instruct` on the Davidson dataset, as Davidson shows the highest overall inconsistency between model predictions and human labels (see § 6.1).[6]



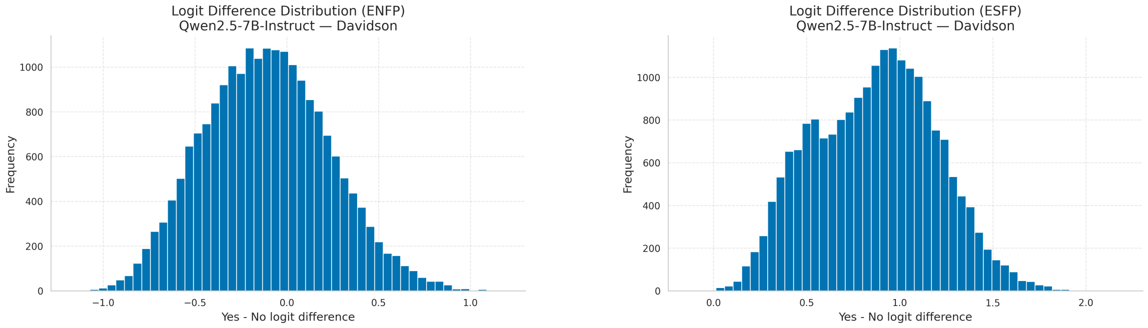
	INTJ	INTP	ENTJ	ENTP	INFJ	INFP	ENFJ	ENFP	ISTJ	ISFJ	ESTJ	ESFJ	ISTP	ISFP	ESTP	ESFP
INTJ	0.00	0.15	0.15	0.18	0.16	0.06	0.18	0.47	0.08	0.36	0.17	0.07	0.31	0.11	0.08	0.16
INTP	0.15	0.00	0.28	0.09	0.14	0.16	0.13	0.33	0.11	0.24	0.11	0.14	0.20	0.24	0.15	0.30
ENTJ	0.15	0.28	0.00	0.32	0.31	0.12	0.32	0.61	0.20	0.50	0.32	0.18	0.46	0.04	0.18	0.01
ENTP	0.18	0.09	0.32	0.00	0.10	0.20	0.09	0.29	0.13	0.19	0.09	0.15	0.15	0.28	0.16	0.34
INFJ	0.16	0.14	0.31	0.10	0.00	0.19	0.07	0.31	0.13	0.20	0.09	0.14	0.16	0.26	0.14	0.32
INFP	0.06	0.16	0.12	0.20	0.19	0.00	0.20	0.49	0.10	0.38	0.20	0.08	0.34	0.09	0.09	0.14
ENFJ	0.18	0.13	0.32	0.09	0.07	0.20	0.00	0.29	0.14	0.19	0.10	0.15	0.16	0.28	0.15	0.33
ENFP	0.47	0.33	0.61	0.29	0.31	0.49	0.29	0.00	0.42	0.14	0.29	0.44	0.17	0.57	0.44	0.63
ISTJ	0.08	0.11	0.20	0.13	0.13	0.10	0.14	0.42	0.00	0.31	0.13	0.07	0.26	0.16	0.08	0.21
ISFJ	0.36	0.24	0.50	0.19	0.20	0.38	0.19	0.14	0.31	0.00	0.19	0.33	0.11	0.46	0.33	0.52
ESTJ	0.17	0.11	0.32	0.09	0.09	0.20	0.10	0.29	0.13	0.19	0.00	0.15	0.15	0.28	0.15	0.33
ESFJ	0.07	0.14	0.18	0.15	0.14	0.08	0.15	0.44	0.07	0.33	0.15	0.00	0.28	0.13	0.07	0.19
ISTP	0.31	0.20	0.46	0.15	0.16	0.34	0.16	0.17	0.26	0.11	0.15	0.28	0.00	0.41	0.28	0.47
ISFP	0.11	0.24	0.04	0.28	0.26	0.09	0.28	0.57	0.16	0.46	0.28	0.13	0.41	0.00	0.14	0.06
ESTP	0.08	0.15	0.18	0.16	0.14	0.09	0.15	0.44	0.08	0.33	0.15	0.07	0.28	0.14	0.00	0.19
ESFP	0.16	0.30	0.01	0.34	0.32	0.14	0.33	0.63	0.21	0.52	0.33	0.19	0.47	0.06	0.19	0.00

Figure 5: The Persona vs Persona Disagreement Matrix for `Qwen2.5-7B-Instruct` on Davidson, higher value denotes higher disagreement. Diagonal values are zero by design, as they represent self-agreement.

Each cell in the matrix quantifies the disagreement rate, ranging from 0 to 1, where higher values indicate more frequent divergence in predicted labels on the same inputs. ENFP emerges as the most divergent persona, with disagreement values around 0.6 with multiple others, including ISTJ (0.57), ESFP (0.63), and ENTJ (0.61). In contrast, logical and structured personas such as ISTJ, INTJ, and ESTJ show lower pairwise disagreement among themselves, such as INTJ–ISTJ (0.08) and

[5] The result of HateXplain is presented in Appendix D Figure 10, as it presents more uniform inconsistency levels across LLMs.

[6] Disagreement matrices for other datasets and LLMs can be found in Appendix E.



(a) ENFP

(b) ESFP

Figure 6: Logit Difference Distribution of `Qwen2.5-7B-Instruct` for ENFP vs ESFP on Davidson.

ESTJ–ISTJ (0.13), indicating more aligned classification behavior.

Feeling and Perceiving types such as INFP, ESFP, and ENFP exhibit higher disagreement levels with others, suggesting that more emotionally driven personas introduce greater variability in model predictions. The highest disagreement is observed between ENFP and ESFP (0.63), despite both being extraverted and feeling-oriented, suggesting that even small trait differences such as Intuition vs. Sensing can significantly alter the model's labeling decisions.

Overall, these results reveal that intra-model disagreement is strongly influenced by MBTI traits, with emotionally driven (F) and less structured (P) personas introducing the highest variability, while analytical (T) and judging (J) types yield more consistent predictions across personas.

6.3 Tendency Analysis from Logits

The intra-model disagreement shows that persona plays a significant role in hate speech judgment. We further analyze the underlying decision tendencies by examining the logits produced by the LLMs for the 'yes' and 'no' responses. Specifically, we investigate whether LLMs exhibit a systematic bias toward one label over the other when conditioned on different personas.

Given that ENFP and ESFP demonstrate high disagreement on `Qwen2.5-7B-Instruct` for the Davidson dataset, as discussed in § 6.2, we analyze the logit difference distributions for these two personas. Figure 6 presents the distribution of the difference between the logit for the 'yes' token and the logit for the 'no' token, computed as $\text{logit}_{\text{yes}} - \text{logit}_{\text{no}}$.

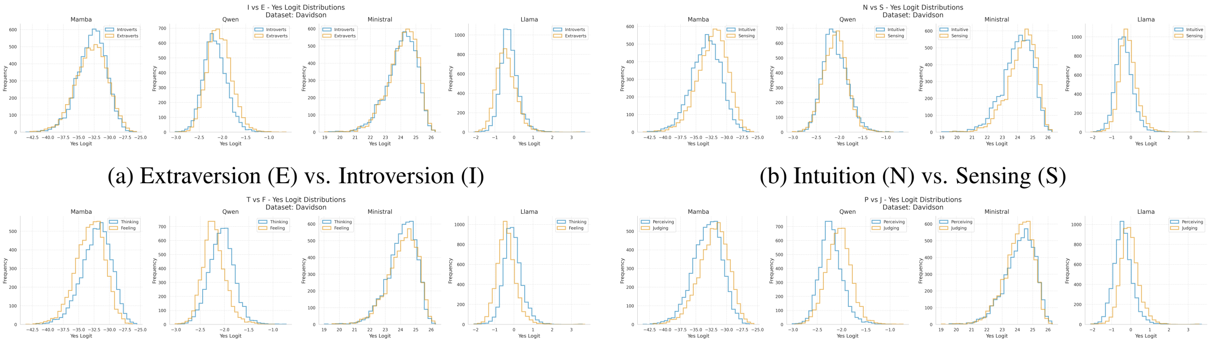
A clear contrast is observed between the two distributions. For ESFP (Figure 6b), the logit differences are strictly greater than zero, indicating that the model consistently assigns higher confidence to the 'yes' class. This suggests a strong inclination toward predicting that the input is hate speech under the ESFP persona. In contrast, for ENFP (Figure 6a), the logit differences are more evenly distributed around zero, with a slight skew toward the negative side, indicating that the 'no' token tends to receive slightly higher scores than 'yes'. This pattern suggests a general bias toward predicting non-hate speech when the model is conditioned on the ENFP persona.

These results suggest that persona traits influence not only the final classification outcomes but also the model's internal decision confidence. While the ESFP persona, characterized by expressiveness and sensitivity to external stimuli, leads to a strong bias toward labeling inputs as hate speech, the ENFP persona, often associated with empathy and open-mindedness, shows a more balanced, slightly non-hate-leaning tendency.

6.4 MBTI Dichotomies Logit Distributions

As ENFP and ESFP show clear logit-level tendencies despite differing in only one MBTI dimension, we further examine how each individual MBTI dichotomy, Extraversion (E) vs. Introversion (I), Intuition (N) vs. Sensing (S), Thinking (T) vs. Feeling (F), and Perceiving (P) vs. Judging (J), affects model behavior in hate speech classification. We focus on the "yes" logit values, which represent the model's internal confidence in predicting hate speech, and analyze their distribution across each dichotomy. The results for all four LLMs on the Davidson dataset are shown in Figure 7.

In the Extraversion vs. Introversion comparison (Figure 7a), we observe that extraverted personas tend to produce slightly higher "yes" logits than



(a) Extraversion (E) vs. Introversion (I)

(b) Intuition (N) vs. Sensing (S)

(c) Thinking (T) vs. Feeling (F)

(d) Perceiving (P) vs. Judging (J)

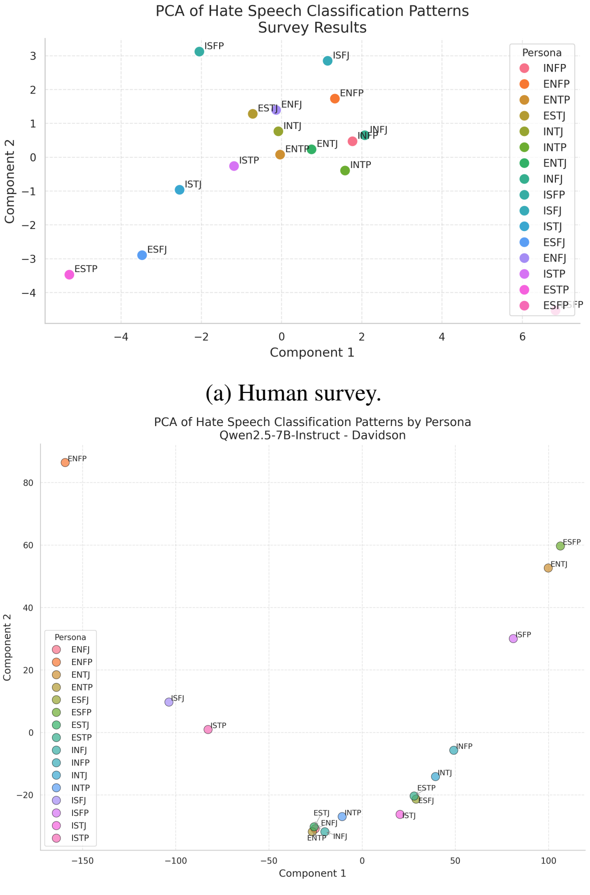
Figure 7: "Yes" logit distributions for each MBTI dichotomy on the Davidson dataset, comparing behavior across four LLMs. Each subplot illustrates how different personality dimensions influence the model's confidence in classifying a sample as hate speech.

introverted ones for Qwen and Ministral, while introverted personas tend to produce slightly higher "yes" logits for Llama. However, the difference is not very evident for this group. In the Intuition vs. Sensing comparison (Figure 7b), sensing types show a modest shift toward higher "yes" logits. This pattern is relatively consistent across models and may suggest that sensing personas are more sensitive to concrete or literal hate indicators in the text. For Thinking vs. Feeling (Figure 7c), thinking personas consistently yield higher "yes" logits than feeling ones across all models. The effect is especially pronounced in Qwen and Llama, which suggests that personas emphasizing analytical reasoning are more likely to assign stronger confidence scores when labeling inputs as hate speech. In the Perceiving vs. Judging dimension (Figure 7d), judging types generally produce higher "yes" logits than perceiving types. This trend, consistent across Mamba, Qwen, and Llama, could reflect a stronger preference among judging personas for making firm decisions regarding moral or ethical boundaries, including the detection of hate speech.

Overall, these results suggest that even high-level personality traits can shape the internal decision-making dynamics of LLMs under persona conditioning. In particular, Thinking and Judging personas tend to produce stronger "yes" logits, reflecting a more decisive and analytical classification style.

6.5 Comparison of Human and LLM Hate Speech Classification Patterns

We conduct PCA on the decision patterns of both human participants in §3 and an LLM (Qwen2.5-7B-Instruct) prompted with different



(a) Human survey.

(b) Qwen2.5-7B-Instruct.

Figure 8: PCA of hate speech classification for human survey and Qwen.

persona descriptions. Figure 8 presents the PCA plots for human survey results and LLM predictions.[7]

The human PCA reveals a relatively compact and overlapping distribution of MBTI types. While some variation is observable, such as ESFP and ESTP appearing more distant from the central cluster, most personas form a coherent grouping. This suggests that although individual personality types influence judgment, human annotators share a broadly consistent understanding of what constitutes hate speech.

In contrast, the PCA of LLM predictions shows a starkly different pattern. The MBTI personas result in widely dispersed decision behaviors, with some types (e.g., ENFP, ESFP, ISFP) forming clear outliers far from the main cluster. This indicates that persona prompting has a significantly stronger impact on the LLM's decision boundary compared to the relatively subtle effects observed in human behavior.

One possible explanation for this divergence is that LLMs, when instructed to adopt a specific persona, may overfit to stereotypical traits described in the prompt. For example, a persona described as "caring, outgoing, and cooperative" might bias the model toward more lenient interpretations of controversial language. This sensitivity suggests that LLMs amplify personality cues more than humans naturally do, potentially leading to inconsistent or biased moderation outcomes.

Overall, the comparison reveals that while human judgments are influenced but constrained by personalities, LLMs are far more malleable to persona prompt.

7 Conclusion

In this work, we study the impact of MBTI-based personas on hate speech classification by combining a human survey with LLM analysis. Our survey reveals that personality traits influence labeling behavior, with significant variation across MBTI types. Extending this to LLMs, we find that persona prompts lead to substantial prediction shifts, including inconsistencies with ground truth and disagreements across personas. At the logit level, we observe systematic biases aligned with persona traits. Additionally, examining the influence of individual MBTI dichotomies reveals that Thinking and Judging traits are associated with stronger logit confidence in hate speech detection.

These findings demonstrate that persona conditioning in LLMs not only affects final predictions but also shapes the underlying decision-making process. As LLMs are increasingly used in socially sensitive applications, future work could explore how to better understand and control the influence of personas to ensure fairness, consistency, and reliability in subjective tasks.

Limitations

While our study offers novel insights into persona-based prompting in LLMs, several limitations should be acknowledged. First, we focus exclusively on the MBTI framework as the basis for personality representation. Although MBTI is widely recognized and structured, future work could explore alternative psychological theories, such as the Big Five, to provide different perspectives. Second, our analysis is conducted on a limited set of benchmark hate speech datasets, which may not fully capture the richness and diversity of hate speech encountered in real-world settings. Third, we evaluate only four open-source LLMs, each with approximately 7 to 8 billion parameters. Future work should investigate whether similar effects occur in larger models. Fourth, our human survey sample consists primarily of students and researchers, which may limit the generalizability of our findings to other demographic groups. Finally, our investigation is restricted to English-language data. Cross-lingual studies are needed to examine how persona effects and perceptions of hate speech vary across different cultural and linguistic contexts.

Ethical Considerations

This study investigates the influence of MBTI-based personas on LLM behavior in the context of hate speech detection. Given the sensitive nature of both hate speech content and personality profiling, several ethical considerations must be addressed.

Hate Speech Content. The use of hate speech examples, even for research purposes, carries the risk of exposing readers and researchers to harmful or offensive material. We use such content solely to support the scientific objectives of this work and explicitly disclaim any intent to promote or reproduce hateful language. A content warning is provided at the beginning of the paper to mitigate

[7] The results for other datasets and LLMs can be found in Appendices F and G.

potential harm.

Use of MBTI Personas. While the MBTI framework has limited psychometric validity, we adopt it as a structured and interpretable method for simulating persona variation in LLMs. The MBTI profiles are used as fictional constructs to explore behavioral diversity in model outputs and are not intended to diagnose, stereotype, or represent real individuals. Our use of MBTI is exploratory in nature and aims to highlight potential variation, not prescribe personality-based approaches.

Bias and Stereotyping. There is a potential ethical risk in associating specific personas with differing behaviors in hate speech detection, which could inadvertently reinforce stereotypes. To mitigate this, we conduct our analysis with care and emphasize that the observed patterns reflect model behavior under specific prompt conditions. No personality type is pathologized or presented as superior or inferior. Our findings should be interpreted as insights into LLM response variability, not as generalizations about personality traits.

References

Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. *WebSci2013. ACM*, 2013(2013).

Alessandro Astorino, Giulia Rizzi, and Elisabetta Fersini. 2023. Integrated gradients as proxy of disagreement in hateful content. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 46–52, Venice, Italy. CEUR Workshop Proceedings.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent LLM interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9306–9326, Miami, Florida, USA. Association for Computational Linguistics.

Gregory J. Boyle. 1995. Myers-briggs type indicator (mbti): Some psychometric limitations. *Australian Psychologist*, 30(1):71–74.

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

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Hyeong Kyu Choi and Yixuan Li. 2024. Picle: eliciting diverse behaviors from large language models with persona in-context learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Gloria del Valle-Cano, Lara Quijano-Sánchez, Federico Liberatore, and Jesús Gómez. 2023. Socialhaterbert: A dichotomous approach for automatically detecting hate speech on twitter through textual analysis and user profiles. *Expert Systems with Applications*, 216:119446.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can LLM be a personalized judge? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141, Miami, Florida, USA. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, pages 2591–2597, Online. Association for Computational Linguistics.

Paula Fortuna, Monica Dominguez, Leo Wanner, and Zeerak Talat. 2022. Directions for NLP practices applied to online hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11794–11805, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Leon Fröhling, Gianluca Demartini, and Dennis Assenmacher. 2024. Personas with attitudes: Controlling llms for diverse data annotation. *arXiv preprint arXiv:2410.11745*.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Jane Isman, Garrick Sherman, Zachary Fried, João Sedoc, Lyle Ungar, and Brenda Curtis. 2024. Modeling human subjectivity in LLMs using explicit and implicit human factors in personas. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7174–7188, Miami, Florida, USA. Association for Computational Linguistics.

Kristina Gligoric, 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, Albuquerque, New Mexico. Association for Computational Linguistics.

Lewis R Goldberg. 2013. An alternative “description of personality”: The big-five factor structure. In *Personality and personality disorders*, pages 34–47. Routledge.

Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making large language models to be better crowdsourced annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico. Association for Computational Linguistics.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Tomáš Horych, Christoph Mandl, Terry Ruas, Andre Greiner-Petter, Bela Gipp, Akiko Aizawa, and Timo Spinde. 2025. The promises and pitfalls of LLM annotations in dataset labeling: a case study on media bias detection. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1370–1386, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, and He He. 2024. Personas as a way to model truthfulness in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6346–6359, Miami, Florida, USA. Association for Computational Linguistics.

Youngwook Kim, Shinwoo Park, and Yo-Sub Han. 2022. Generalizable implicit hate speech detection using contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6667–6679, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kyuhan Lee and Sudha Ram. 2024. Deep learning for hate speech detection: A personality-based approach. In *Companion Proceedings of the ACM Web Conference 2024*, WWW ’24, page 1667–1671, New York, NY, USA. Association for Computing Machinery.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. 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 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

Bohan Li, Jiannan Guan, Longxu Dou, Yunlong Feng, Dingzirui Wang, Yang Xu, Enbo Wang, Qiguang Chen, Bichen Wang, Xiao Xu, Yimeng Zhang, Libo Qin, Yanyan Zhao, Qingfu Zhu, and Wanxiang Che. 2025. Can large language models understand you better? an MBTI personality detection dataset aligned with population traits. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5071–5081, Abu Dhabi, UAE. Association for Computational Linguistics.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Sarah Masud, Sahajpreet Singh, Viktor Hangya, Alexander Fraser, and Tanmoy Chakraborty. 2024. Hate personified: Investigating the role of LLMs in content moderation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15847–15863, Miami, Florida, USA. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

MistralAI. 2024. Ministral-8b-instruct-2410.

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.

Rajiv Movva, Pang Wei Koh, and Emma Pierson. 2024. Annotation alignment: Comparing LLM and human annotations of conversational safety. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9048–9062, Miami, Florida, USA. Association for Computational Linguistics.

Isabel Briggs Myers and Peter B Myers. 2010. *Gifts differing: Understanding personality type*. Nicholas Brealey.

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

Matthias Orlikowski, Paul Röttger, Philipp Cimiano, and Dirk Hovy. 2023. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1017–1029, Toronto, Canada. Association for Computational Linguistics.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank and Dirk Hovy. 2015. Personality traits on Twitter—or—How to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98, Lisboa, Portugal. Association for Computational Linguistics.

Mirjana Prpa, Giovanni Maria Troiano, Matthew Wood, and Yvonne Coady. 2024. Challenges and opportunities of llm-based synthetic personae and data in hci. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

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

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, and 1 others. 2024. Position: a roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46280–46302.

Sanja Stajner and Seren Yenikent. 2021. Why is MBTI personality detection from texts a difficult task? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3580–3589, Online. 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.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025. CharacterBox: Evaluating the role-playing capabilities of LLMs in text-based virtual worlds. 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 6372–6391, Albuquerque, New Mexico. Association for Computational Linguistics.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

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.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *arXiv preprint arXiv:2407.19594*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024b. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

Wenjie Yin, Vibhor Agarwal, Aiqi Jiang, Arkaitz Zubiaga, and Nishanth Sastry. 2023. Annobert: Effectively representing multiple annotators' label choices to improve hate speech detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 902–913.

Shuzhou Yuan, Antonis Maronikolakis, and Hinrich Schütze. 2022. Separating hate speech and offensive language classes via adversarial debiasing. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 1–10, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. LLMaAA: Making large language models as active annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103, Singapore. Association for Computational Linguistics.

Hao Zhu, Bodhisattwa Prasad Majumder, Dirk Hovy, and Diyi Yang. 2025. Social intelligence in the age of LLMs. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 51–55, Albuquerque, New Mexico. Association for Computational Linguistics.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Jingwei Zuo, Maksim Velikanov, Dhia Eddine Rhaiem, and et al. 2024. Falcon mamba: The first competitive attention-free 7b language model. *arXiv preprint arXiv:2410.05355*.

A Models

We use four open source instruction-tuned LLMs with the size of 7-8B in our experiments. Further information about each model can be found on the respective HuggingFace pages below:

- Llama3.1: https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

- Ministral: https://huggingface.co/mistralai/Ministral-8B-Instruct-2410

- Qwen2.5: https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

- Mamba: https://huggingface.co/tiiuae/falcon-mamba-7b-instruct

B Survey

B.1 Setup

Participants were first provided with a definition for hate speech and offensive language along with an example for each. Below is the text included in the survey's introduction:

Hate speech is defined as language used to express hatred towards a targeted group/individual based on protected attributes such as race, religion, sexual orientation, etc.

Example: "All [religious group]s are terrorists. Deport them all!"

Offensive language contains offensive terms but is not targeting any group in particular.

Example: "Stop acting like a bitch and grow up already."

The survey contained a total of 20 samples with 10 samples classified as hate speech and 10 samples classified as offensive language but not hate speech. Below are the samples included in the survey, starting with the hate speech samples and followed by the offensive language samples:

- #California is full of white trash who moved from #Oklahoma

- #faggots are MERELY a #scientific indicator that SOMETHING IS WRONG; Do not celebrate as "unique indiviuals" for they are #cancer

- @[USERNAME]: No need to thank me, killing niggers is it's own reward. Dis my pal from our graduation when we was 26

- #firefighter is a job for white trash

- @[USERNMAE] @[USERNAME] So you support "illegals" coming here and spreading disease among Americans Dream that bitch!

- @[USERNAME] what's this chinks email? I'm moving to China and slicing his throat

- #SomethingIGetAlot Are you... asian? black? Hawaiian? gay? retarded? drunk?

- #JesusChrist was STRAIGHT That's why the #faggots killed him. #PERIOD #SonOfGod

- RT @[USERNAME]: Well maybe you're a whore you stupid ugly homeless bitch

- They thought I wouldn't make it pussy nigga we made it

- no youre a fucking retard thats gonna get fucking date raped at a frat tonight. have fun with those repressed memories

- Man bitches love Beyonce....no punch line

- Then ya'll idolize Kanye when he wifed up a hoe. Lol

- RT @[USERNAME]: When someone's phone goes off in the cinema I want to jump across the seats and fly kick dat bitch.

- RT @[USERNAME]: I convinced hitler was a faggit, no bitches, no kids, all those leather uniforms, wanted to be alone with his closest men lock

- "@[USERNAME]: No text back? I don't give a fuck, I'll text another bitch" damn someone is hurt lol

- @[USERNAME]'s a faggot. I'm glad your chipped him.

- Bitches cheat on they bitch... But be SOOOOOO HURT when they bitch cheat back.

- You a dike cause yo man a pussy.

- "@RTNBA: Drakes new shoes that will be released by Nike/Jordan.... Yes, there's glitter on the shoes [URL]" ....dudes a fag

B.2 Demographics

The table below summarizes the distribution of the ethnicity groups to which participants of our survey belong to. This information was not mandatory as part of our survey and was voluntarily provided by the participants.

Ethnic Group	Count
White	216
Asian	44
Other	11
Middle Eastern or North African (MENA)	10
Hispanic or Latino/a/x	8
Multiracial / Two or more races	4

Table 2: Distribution of participants by ethnic group.

C Personas

Below are the sixteen personas along with their most common traits. *16personalities.com* use gamified names for each persona, such as The Mediator, The Debater, etc., which have also been included below for completeness. It is also worth mentioning that the models were provided with these names along with the traits.

- INFP – The Mediator: Creative, introspective, and empathetic; guided by inner values and harmony.

- ENFP – The Campaigner: Enthusiastic, imaginative, and warm; values individuality and seeks meaning in relationships.

- ENTP – The Debater: Quick-witted, curious, and argumentative; enjoys intellectual challenges and exploring new ideas.

- ESTJ – The Executive: Organized, decisive, and pragmatic; excels at managing people and systems.

- INTJ – The Architect: Strategic, independent, and analytical; has a vision for the future and works to achieve it.

- INTP – The Logician: Innovative, curious, and analytical; loves exploring abstract ideas and theories.

- ENTJ – The Commander: Bold, strategic, and assertive; natural leader who thrives on achievement and efficiency.

- INFJ – The Advocate: Idealistic, insightful, and reserved; driven by values and a desire to help others.

- ISFP – The Adventurer: Gentle, artistic, and adaptable; values personal freedom and expression.

- ISFJ – The Defender: Loyal, empathetic, and practical; quietly supportive and protective of others.

- ISTJ – The Logistician: Responsible, serious, and detail-oriented; values tradition and duty.

- ESFJ – The Consul: Caring, outgoing, and cooperative; values harmony and tradition in social settings.

- ENFJ – The Protagonist: Charismatic, altruistic, and inspiring; driven to lead and help others grow.

- ISTP – The Virtuoso: Practical, observant, and spontaneous; enjoys hands-on problem-solving.

- ESTP – The Entrepreneur: Energetic, action-oriented, and outgoing; enjoys living in the moment and taking risks.

- ESFP – The Entertainer: Spontaneous, playful, and sociable; seeks fun and connection in everyday life.

D Classification Inconsistency Between LLMs and GT

The figures below depict the percentage inconsistency between LLMs and ground truth for each persona.

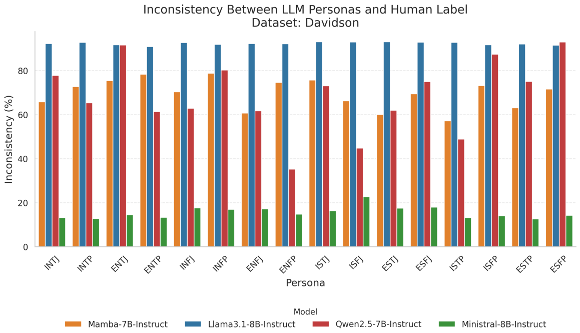


Figure 9: Percentage inconsistency between LLMs and the ground truth on the Davidson dataset

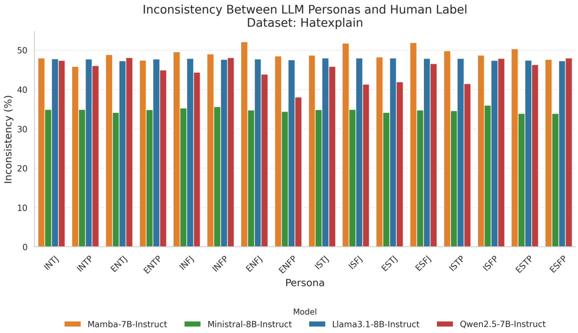


Figure 10: Percentage inconsistency between LLMs and the ground truth on the HateXplain dataset

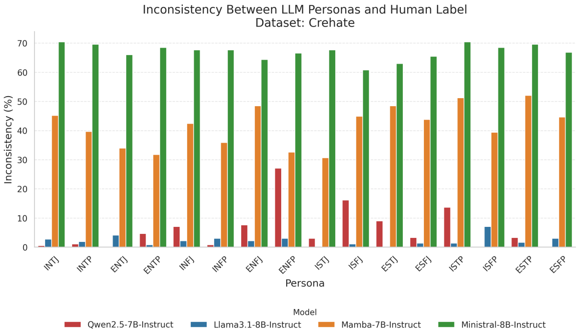


Figure 11: Percentage inconsistency between LLMs and the ground truth on the CReHate dataset

E Disagreement Among Personas

The figures below depict the disagreement among personas.

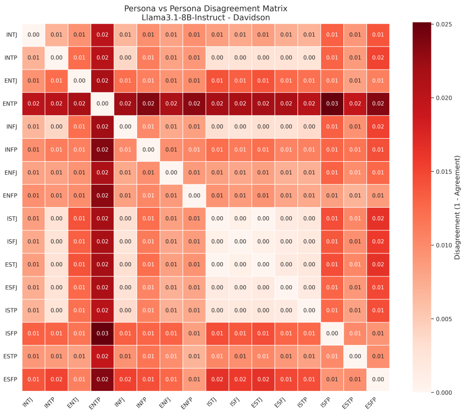


Figure 12: Disagreement among personas for `Llama3.1` on the Davidson dataset

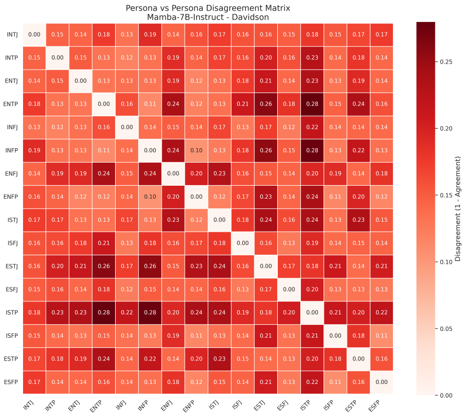


Figure 13: Disagreement among personas for `Mamba` on the Davidson dataset

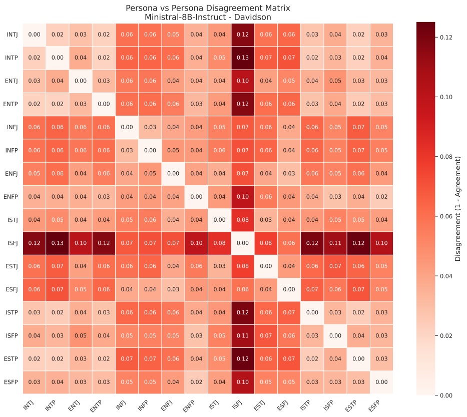


Figure 14: Disagreement among personas for `Ministral` on the Davidson dataset

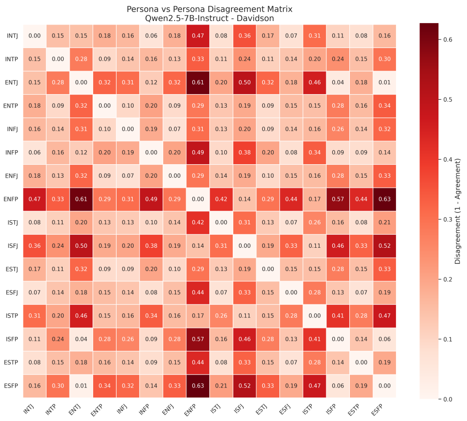


Figure 15: Disagreement among personas for `Qwen2.5` on the Davidson dataset

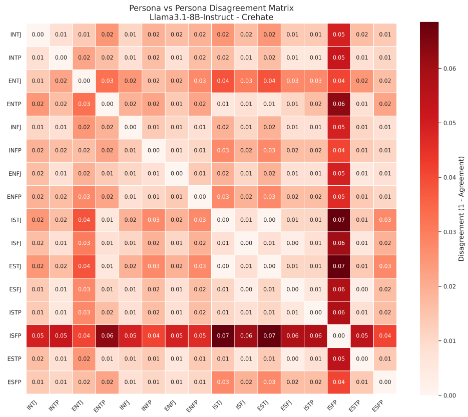


Figure 16: Disagreement among personas for `Llama3.1` on the CREHate dataset

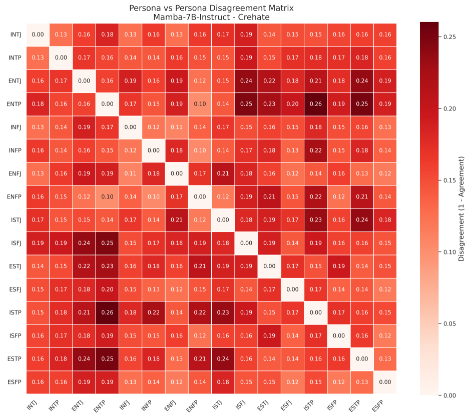


Figure 17: Disagreement among personas for `Mamba` on the CREHate dataset

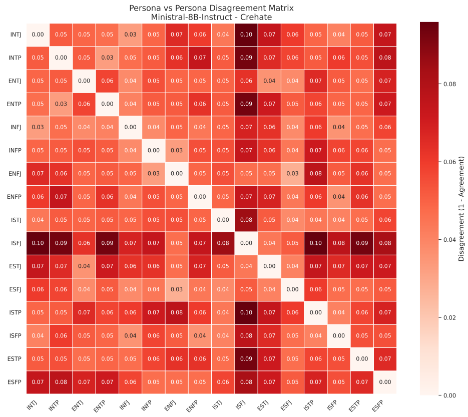


Figure 18: Disagreement among personas for `Ministral` on the CREHate dataset

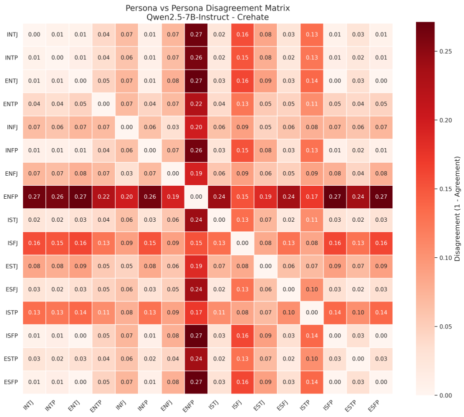


Figure 19: Disagreement among personas for `Qwen2.5` on the CREHate dataset

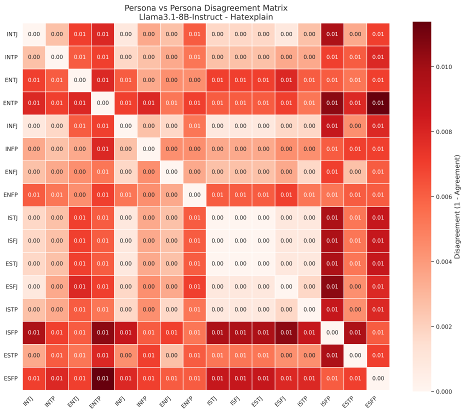


Figure 20: Disagreement among personas for `Llama3.1` on the HateXplain dataset

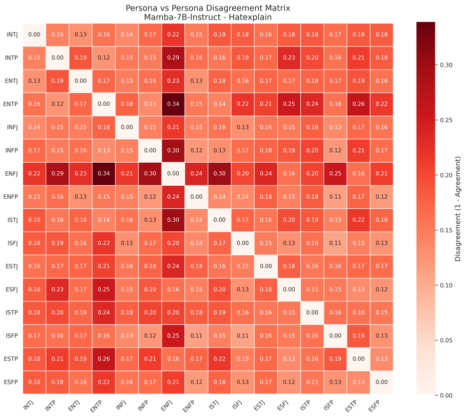


Figure 21: Disagreement among personas for `Mamba` on the HateXplain dataset

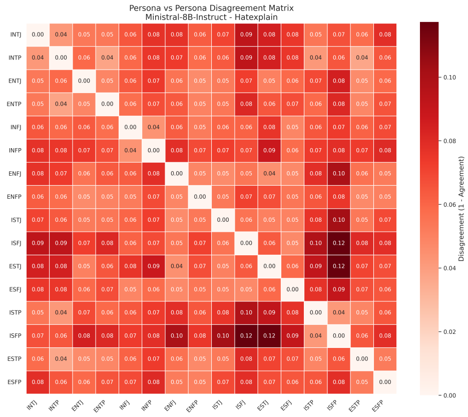


Figure 22: Disagreement among personas for `Ministral` on the HateXplain dataset

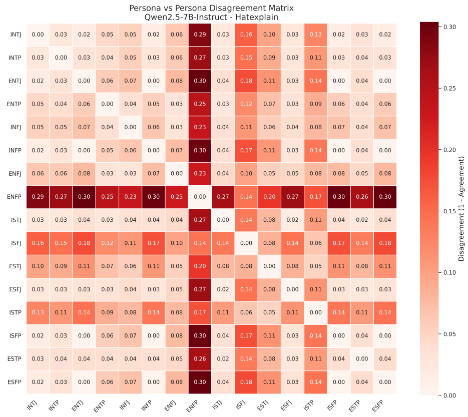


Figure 23: Disagreement among personas for `Qwen2.5` on the HateXplain dataset

F PCA

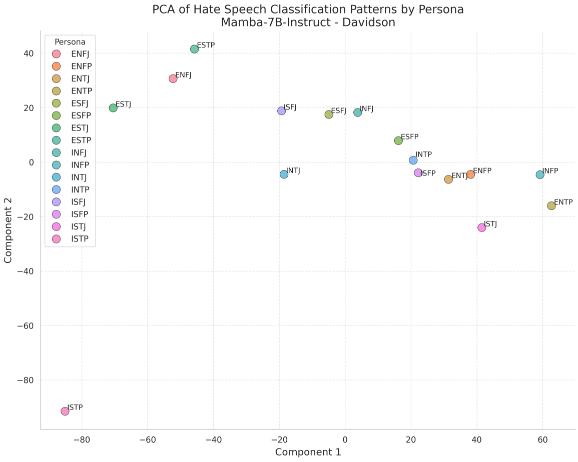


Figure 24: PCA of hate speech classification per persona for `Mamba` on the Davidson dataset

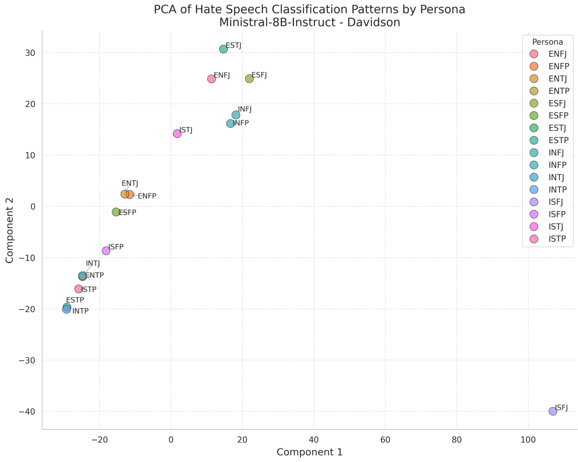


Figure 25: PCA of hate speech classification per persona for `Minisstral` on the Davidson dataset

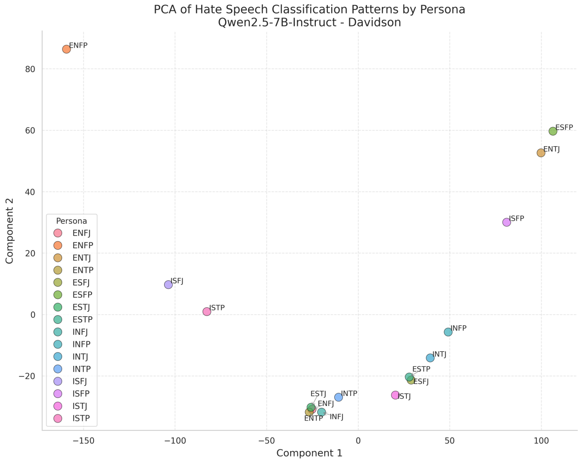


Figure 26: PCA of hate speech classification per persona for `Qwen2.5` on the Davidson dataset

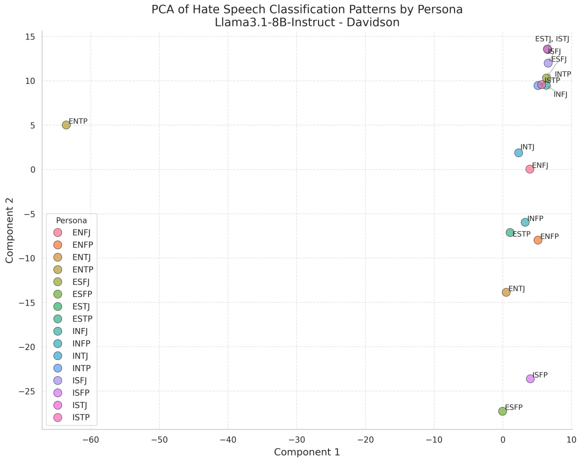


Figure 27: PCA of hate speech classification per persona for `Llama` on the Davidson dataset

G Hierarchical Clustering

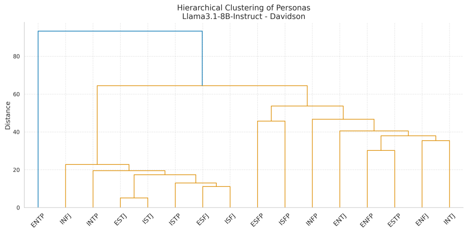


Figure 28: Hierarchical clustering of persona representations for `LLaMA3.1-8B-Instruct` on the Davidson dataset.

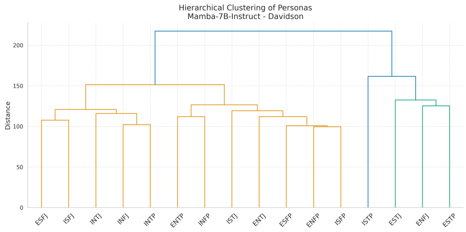


Figure 29: Hierarchical clustering of persona representations for `Mamba-7B-Instruct` on the Davidson dataset.

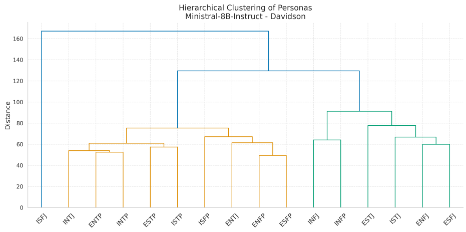


Figure 30: Hierarchical clustering of persona representations for `Ministral-8B-Instruct` on the Davidson dataset.

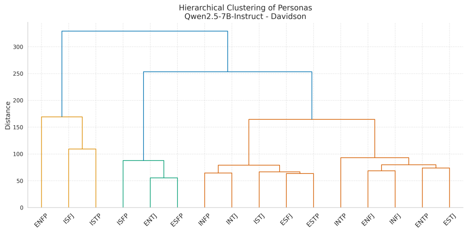


Figure 31: Hierarchical clustering of persona representations for `Qwen2.5-7B-Instruct` on the Davidson dataset.

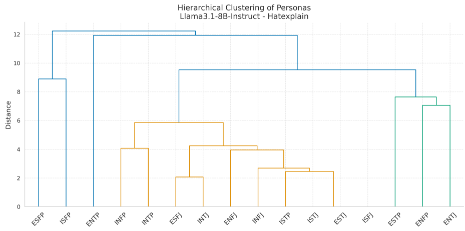


Figure 32: Hierarchical clustering of persona representations for `LLaMA3.1-8B-Instruct` on the HateXplain dataset.

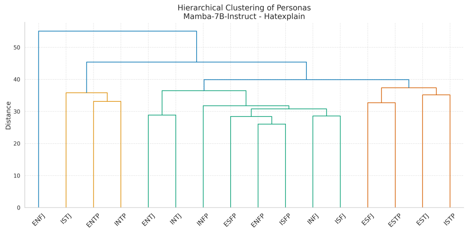


Figure 33: Hierarchical clustering of persona representations for `Mamba-7B-Instruct` on the HateXplain dataset.

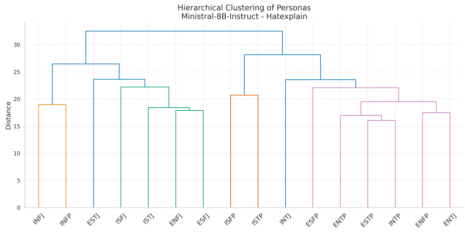


Figure 34: Hierarchical clustering of persona representations for `Ministral-8B-Instruct` on the HateXplain dataset.

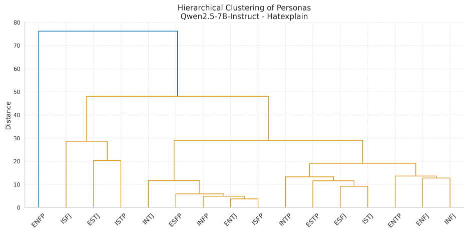


Figure 35: Hierarchical clustering of persona representations for `Qwen2.5-7B-Instruct` on the HateXplain dataset.

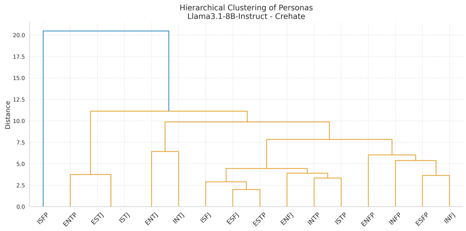


Figure 36: Hierarchical clustering of persona representations for `LLaMA3.1-8B-Instruct` on the CREHate dataset.

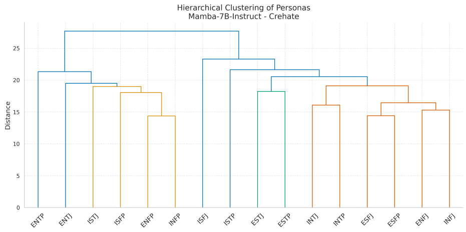


Figure 37: Hierarchical clustering of persona representations for `Mamba-7B-Instruct` on the CREHate dataset.

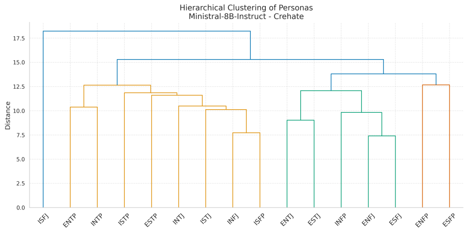


Figure 38: Hierarchical clustering of persona representations for `Ministral-8B-Instruct` on the CREHate dataset.

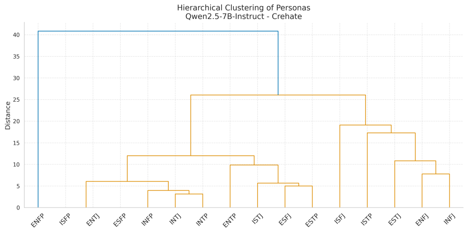


Figure 39: Hierarchical clustering of persona representations for `Qwen2.5-7B-Instruct` on the CREHate dataset.