Hate Personified: Investigating the role of LLMs in content moderation

Sarah Masud[*1], **Sahajpreet Singh**[*2], **Viktor Hangya**[3],
Alexander Fraser[4], **Tanmoy Chakraborty**[2]
[1]IIIT Delhi, [2]IIT Delhi, [3]LMU, [4]TUM
sarahm@iiitd.ac.in, sahaj.phy@gmail.com, hangyav@cis.lmu.de,
alexander.fraser@tum.de, tanchak@iitd.ac.in

Abstract

For subjective tasks such as hate detection, where people perceive hate differently, the Large Language Model's (LLM) ability to represent diverse groups is unclear. By including additional context in prompts, we comprehensively analyze LLM's sensitivity to geographical priming, persona attributes, and numerical information to assess how well the needs of various groups are reflected. Our findings on two LLMs, five languages, and six datasets reveal that mimicking persona-based attributes leads to annotation variability. Meanwhile, incorporating geographical signals leads to better regional alignment. We also find that the LLMs are sensitive to numerical anchors, indicating the ability to leverage community-based flagging efforts and exposure to adversaries. Our work provides preliminary guidelines and highlights the nuances of applying LLMs in culturally sensitive cases.[1]

1 Introduction

Human evaluators from diverse backgrounds are necessary to provide coverage against hate speech, which the UN defines as "any kind of communication that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are." Variation in annotations with respect to demographics matters as it is reflective of their lived experiences. However, the background of the evaluators also contributes to annotation biases (Rottger et al., 2022; Aroyo et al., 2019a; Munn, 2020). On the other hand, the role of Large Language models (LLMs) (Zhang et al., 2023) to help content moderation is now being explored (Huang et al., 2023; Roy et al., 2023).

Motivation. In a recently proposed hate speech dataset `CREHate` (Lee et al., 2023), annotators of five countries perceived the same post (in English) differently. The authors also observed variation upon introducing the country when prompting LLM for similar annotations. Reproducing results from `CREHate`, we observe (Figure 1) a gap between the human-LLM alignment of hatefulness. *We, thus, investigate variations in LLM's output when primed with context, which, under a similar setting for humans, causes variability.* In a two-party setup consisting of a human and a zero-shot LLM as annotators, we prompt the LLM with contextual information and observe the inter-annotator variation (w.r.t human label) in the LLM's output.

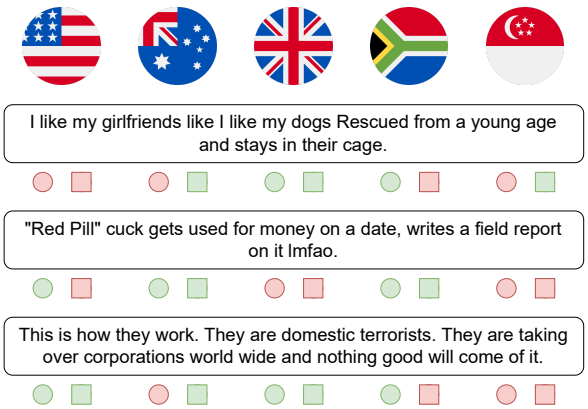


Figure 1: Annotations of hate/non-hate (red/green) for USA, Australia, UK, South Africa, and Singapore, by a) annotators from respective countries (circle) and b) prompting `GPT-3.5` with 'The following statement was made in <country>: <POST>. Is the given statement hateful?' (square). The posts and human labels are verbatim from the `CREHate` dataset (Lee et al., 2023).

Borrowing from the demographic identities of `CREHate`, we assess if LLMs qualify as a proxy for annotators of a given identity. The direct implication is obtaining cheaper and faster soft labels via LLMs for a demographic that is underrepresented among human annotators. We also investigate the role of geographical cues, similar to the use of geolocation metadata by online platforms, in annotating hate. Apart from the post's content, numerical

[*] Equal Contribution
[1]**Disclaimer:** The paper contains examples of strong and hateful language.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 15847–15863
November 12-16, 2024 ©2024 Association for Computational Linguistics

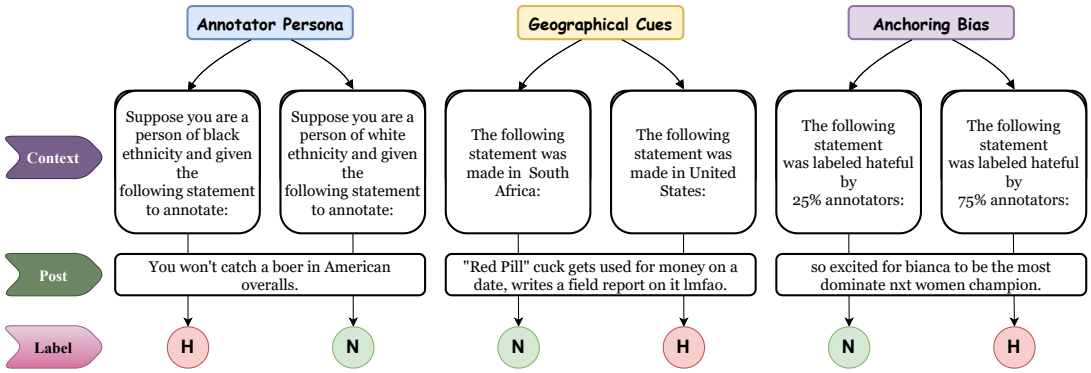


Figure 2: Overview of the research pipeline: an incoming post is prefixed with context to form the prompt for the LLM. The predicted label is then evaluated against ground truth to examine variability arising from context.

metadata (Founta et al., 2019), such as the number of likes, number of times a post has been flagged as offensive, etc, is often available internally to content moderators. We examine if such statistics can better guide LLMs to reflect a community's needs.

Prompting Format. Decision-making in the human (Slovic et al., 1977) is influenced by prior and current information. However, for general-purpose LLMs to replicate the same effect, they must be externally cajoled/cued. The prompts are formatted as 'cue + post + query', where p_{base} = *'Statement: <POST>. Is the given statement hateful?'* forms the base prompt. Further, we translate the cue and query into the respective language via Google Translate + manual verification for the multilingual datasets. RQ1 (Section 4) delves into the geographical nuances like the country or language of the post (e.g., *'The following statement was made in country x...'*) as a cue for regional context. RQ2 (Section 5) explores diverse demographic facets (e.g., *'Suppose you are a person of x ethnicity...'*) as a proxy for different human evaluator groups. Finally, RQ3 (Section 6) examines the variability introduced by a numerical context, expressed as *'x% of individuals labeled this post as hateful.'*, as a proxy for anchoring bias in LLMs. The RQs allow us to investigate the difference between implicit (base prompt) and explicit (contextual prompt) nudging (Figure 2).

Research Scope. Firstly, the aim is not to establish SOTA for a given dataset. We do not engineer the highest-performing prompt for a specific dataset. Instead, our study helps provide a general assessment across the datasets. Secondly, finetuning on a single hate speech dataset does not necessarily transfer to out-of-distribution samples (Yin and Zubiaga, 2021). Lastly, the variability and nuances probed in this study are not feasible to replicate under in-context and finetuned setups where multiple hyperparameters already impact the output. Therefore, in this study, we examine only the zero-shot prompt setting.

Observations and Implications. We perform an exhaustive analysis spanning two LLMs, five languages, and six hate speech datasets. We investigate 86 prompts (Table 12) in English and 40 multilingual prompts (Table 13). We caution against the blindsided use of LLMs for crowdsourcing in subjective tasks. In summary, this work can help practitioners gauge the LLM vs. manual effort needed in their content moderation pipeline.

- **Geographical Cues.** We observe that geographical cues lead to visible and significant increases in human-LLM agreement (Section 4). As these metadata are readily available within a platform, including them in the prompting does not require additional effort.
- **Persona Cues.** As no persona cue + LLM combination encompasses sensitivity toward all groups, practitioners employing LLMs as proxies must do so cautiously. Further, we establish that 'the manner/format' of imbibing the persona is equally crucial (Section 5).
- **Numerical Cues.** We observe that pseudo-voting values influence predicted labels, highlighting both positive (community flagging) and negative (adversarial attacks) effects. The results question the use of numerical features in zero-shot prompting (Section 6).
- **Multilinguality:** Under multilingual prompting, we observe the above patterns to persist, albeit with an expected loss in annotation agreements (Sections 4 and 5). As native speakers would prefer to engage with LLMs in their native language, this calls for improving the low-resource specificity in LLMs.

Dataset (Language) (Reference)	# Samples in original dataset			# Samples used in RQs		
	# Hate	# Non-hate	Total	# Hate	# Non-hate	Total
`HateXplain` (En) (Mathew et al., 2021)	4748	6251	10999	4748	6251	10999
`CREHate` (En) (Lee et al., 2023)	709	871	1580	709	871	1580
`MLMA` (Ar) (Ousidhoum et al., 2019)	460	915	1375	250	250	500
`MLMA` (Fr) (Ousidhoum et al., 2019)	207	821	1028	207	293	500
`HASOC-2020` (De) (Mandl et al., 2020)	146	1700	1846	146	354	500
`HASOC-2020` (Hi) (Mandl et al., 2020)	234	2116	2350	234	266	500

Table 1: Dataset statistics employed in this study. Here, Hate $\sim$ Hateful & Non-Hate $\sim$ Normal $\sim$ None.

At the intersection of annotation priming and zero-shot prompting, the current research acts as a guideline for configuring the LLM-assisted content moderation pipeline. Our study establishes the role of context in making LLM-based hate speech annotations reflect the preferences of a given vulnerable community or different cultural groups more closely. Our study illustrates that explicit cues align better with human annotations.

2 Related Work

Annotation Biases. Labeling for hate speech datasets is primarily led by human annotators (Waseem, 2016; Founta et al., 2018). Human annotations for subjective tasks (Rottger et al., 2022) are rife with ambiguity (Kanclerz et al., 2022; Aroyo et al., 2019b) and biases (Garg et al., 2023). Of particular interest is the annotation bias (Wich et al., 2021). In hate speech, annotation bias manifests due to differences among the annotator's belief (Sap et al., 2022), experience, world knowledge (Yin and Zubiaga, 2021; Abercrombie et al., 2023), and social-demographic conditioning (Orlikowski et al., 2023). Disparity in access to additional context (Ljubešić et al., 2022; İhtiyar et al., 2023) and annotation guidelines (Ross et al., 2016), in some instances, reduce the bias and, in some cases, confirm the annotator's biases. Analysis and mitigation of biases in hate speech is an active area of research (Biester et al., 2022; Wojatzki, 2018; Sap et al., 2019). Parallel research seeks to model diverse opinions (Braylan and Lease, 2020; Li et al., 2021; Weerasooriya et al., 2023) as a way to reduce the annotation bias. Variability in labeling is unavoidable, even if considered on a continuous scale (Sachdeva et al., 2022) or augmented with signals (Koufakou et al., 2020) of sentiment or emotion, both of which are hard to annotate.

LLMs for Annotations. The rise of instruction-tuned LLMs has further facilitated prompt-based labeling of hate speech (AlKhamissi et al., 2022; Yang et al., 2023; Huang et al., 2023; Roy et al., 2023). Consequently, chain-of-thought (AlKhamissi et al., 2022) and few-shot (Khullar et al., 2023) prompting are also being investigated for hate speech detection. However, the use of LLMs for annotations in NLP tasks is still in its early stages (He et al., 2023; Ostyakova et al., 2023) owing to ethical, legal, and interpretability concerns (Zini and Awad, 2022). The initial research does hint at the efficiency of zero-shot annotations via GPT-3.5 (Chung et al., 2022) for myriad standard NLP tasks (Gilardi et al., 2023; Kocoń et al., 2023). However, parallel research has shed light on adversarial attacks (Le et al., 2022; Zou et al., 2023; Nookala et al., 2023) and generative biases while prompting (Griffin et al., 2023; Lin and Ng, 2023; Wang et al., 2023; Zhao et al., 2021). It should be noted that while in-context learning is a promising area of research, the n-shot samples may exhibit repetition and sampling biases (Zhao et al., 2021; Zamfirescu-Pereira et al., 2023). Hence, our analysis concentrates on the zero-shot setting as a first step to control for additional variability.

3 Experimental Setup

This section outlines the datasets, models, and evaluation metrics employed in the study[2].

Datasets. An overview of the datasets employed in this study is provided in Table 1. All these datasets are publicly available. These datasets contain multi-class labels, including hate speech, offensive, and normal for `HateXplain`, and various mixed labels, such as abusive_hateful for multilingual datasets. To remove the subjectivity of these umbrella terms, we classify instances as *hate* when the label is either 'hate' or 'hateful,' and as *non-hate* when the label is 'normal,' 'none,' or 'non-hate.' We utilize the exact texts from the original datasets. We have not attempted to identify or remove any previously mentioned entities, such as

[2]https://github.com/sahajps/Hate-Personified

Model	# of parameters	HateXplain				CREHate			
		# Samples	# Hal	F1	IAA	# Samples	# Hal	F1	IAA
FlanT5-Small	60M	≈11k	2	0.412	0.000	≈1.5k	2	0.391	0.000
FlanT5-Base	250M	≈11k	85	0.649	0.341	≈1.5k	156	0.536	0.166
FlanT5-Large	780M	≈11k	4545	0.339	0.136	≈1.5k	572	0.411	0.187
FlanT5-XL	3B	≈11k	0	0.588	0.293	≈1.5k	4	0.638	0.292
Mistral	7B	≈11k	135	0.531	0.228	≈1.5k	198	0.568	0.303
Zephyr	7B	≈11k	3948	0.343	0.123	≈1.5k	560	0.323	0.102
Llama 3	8B	≈11k	1971	0.439	0.180	≈1.5k	679	0.357	0.150
FlanT5-XXL	11B	≈11k	0	0.731	0.476	≈1.5k	0	0.649	0.297
FlanT5-XXL	11B	500	0	0.738	0.487	500	0	0.649	0.297
GPT-3.5-Turbo*	>150B	500	0	0.780	0.576	500	2	0.758	0.517

Table 2: Performance of LLMs when prompted with p_{base} = '*<POST>. Is the given statement hateful?*'. We report the number of samples in the data set used for prompting (# samples), the number of hallucinated outputs (# hal), the rectified weighted-F1, and the rectified inter-annotator agreement (IAA). *close-sourced model.

the name of the target person in the given hate speech, etc. This decision ensures that the LLM receives the exact text for annotation as presented to the human annotator, maintaining consistency with the ground truth. For RQ1 and RQ2, we use all 1580 samples from CREHate (Lee et al., 2023). Each sample in English is labeled as hateful or not by annotators from different nations (the United States, Australia, the United Kingdom, South Africa, and Singapore). For RQ3, we employ HateXplain (Mathew et al., 2021), which contains English instances, encompassing three labels – toxic, hateful, and normal. We take samples with a majority label as hateful or normal, leading to $\approx 11k$ instances. We also investigate four datasets containing multilingual and code-mixed (with English) posts. We use Arabic (Ar) and French (Fr) datasets by Ousidhoum et al. (2019) and Hindi (Hi) and German (De) by Mandl et al. (2020). Here again, we binarize the labels wherever applicable.

Evaluation Metric. For p_{base}, we have the ground labels (majority voted gold labels). So, we employ a weighted F1 score to compare the p_{base} outputs. Among the prompt variants in RQs, where we may have a direct ground truth, we analyze the performance disparities via the Cohen-κ inter-annotator agreement (IAA) (Cohen, 1960). It has been observed that IAA and F1 are positively correlated when the dataset is not skewed (Richie et al., 2022). In our case, the skewness is controlled by having samples that are almost equal in both classes (as noted in Table 1). When choosing the IAA metric over the F1 score, our primary goal is not to emphasize high precision and recall. Instead, we aim to demonstrate how closely gold human annotations align with those generated by the LLM. This is why we also use the predicted hate-label ratio (PHLR), which represents the proportion of all samples labeled as hate. It is calculated as the ratio of the total number of predicted hate labels to the total generated labels, excluding hallucinations and empty outputs.

Rectified Scores for Hallucination. As the aim is to know if the post is hateful, for our use-case, any output not in the form of *'yes/no'*, *'hate/non-hate'*, or *'hateful/non-hateful'* can be considered as a 'hallucinated' label falling outside the range of the expected answers. We specify the prompt suffix 'answer in one word only.' We also perform a manual evaluation to access the unique outputs per setup, and where the output could be salvaged, they were updated. In line with the existing literature (Lee et al., 2023), after all filters, the outputs that still did not qualify as acceptable are discarded. Recording the number of hallucinations (discarded outputs), we introduce 'rectified F1/IAA scores', $score_{rectified} = (1 - \frac{h}{t}) \times score$, where h and t are the hallucinated and total samples, respectively. Any mention of F1 and IAA in our study means 'rectified weighted F1' and 'rectified IAA.'

Models. We begin with FlanT5 (Chung et al., 2022), Mistral (Jiang et al., 2023), Zephyr (Tunstall et al., 2023), Llama 3 (Touvron et al., 2023) and InstructGPT (Ouyang et al., 2022) variant GPT-3.5-Turbo (hereby referred to as GPT-3.5). Performance metrics on all shortlisted LLMs (with the total number of parameters in that model) are captured in Table 2 for the p_{base}. From Table 2, we observe that even with p_{base}, most models generate noisy and ill-formatted outputs. To reduce the influence of noisy results, we select only FlanT5-XXL and GPT-3.5 to analyze more intricate setups in our RQs. Both these models portray negligible hallucinations and produce relatively higher F1 and

IAA. To test FlanT5-XXL, we employ all the samples. For GPT-3.5, due to API rate limits, we test random 500 stratified samples per dataset.

We also examine language-specific LLMs to corroborate the role of zero-shot prompting setup of general-purpose LLM (GPT-3.5) with cultural context vs language-specific models. The results (Table 3) highlight that not just demographic-based training corpus but the manner of training and size of models impact multilingualism.

Prompts. Tables 12 and 13 provide the list of verbalized prompts under English and multilingual settings, respectively. For the latter, after employing Google Translate, the sample prompts were verified by two linguistic experts.

Note on Hate Speech Definition. Existing literature has observed that providing explicit definitions to LLMs may not lead to performance gain (Guo et al., 2023; Roy et al., 2023). We, thus, conduct our experiments building on the latent notions of hate speech in LLMs. For a sanity check, we prompt FlanT5-XXL and GPT-3.5 with *'Define: Hate Speech'* and obtain following responses:

- FlanT5-XXL: *"Hate Speech is a form of speech intended to incite hatred or violence against a group of people."*
- GPT-3.5: *"Hate speech refers to any form of communication, whether spoken, written, or otherwise expressed, that promotes or encourages hatred, hostility, discrimination, or violence against individuals or groups based on attributes such as race, ethnicity, religion, nationality, gender, sexual orientation, disability, or other characteristics. It often involves derogatory language, stereotypes, or dehumanizing rhetoric intended to intimidate, offend, or incite harm against the targeted individuals or communities. Hate speech is typically viewed as a violation of ethical principles and may also be prohibited by law in many jurisdictions due to its potential to inflame tensions, provoke violence, and undermine social cohesion."*

Hardware and Software(s). All experiments are conducted on an NVIDIA-A100 (80GB) GPU. Model weights are loaded from the HuggingFace pretrained models using float16 precision. For GPT-3.5, OpenAI's API was utilized. All models are employed in a zero-shot setting. The temperature variation experiment takes a total of 100 hours. The rest of the experiments collectively take ≈ 4-5 hours of GPU time.

Dataset (Lang)	Model (# Params)	# Samples	# Hal	F1	IAA
MLMA (Ar)	Mistral-Ar (7B)	500	378	0.027	0.0
	GPT-3.5 (>150B)	500	222	0.359	0.140
HASOC-2020 (Hi)	Airavata (7B)	500	477	0.046	0.0
	GPT-3.5 (>150B)	500	131	0.257	0.018

Table 3: Performance of LLMs when prompted with p_{base} = *'<POST>. Is the given statement hateful?'* in the respective language. We report the number of samples in the dataset used for prompting (#Samples), the number of hallucinated outputs (#Hal), and the rectified weighted-F1/inter-annotator agreement (IAA).

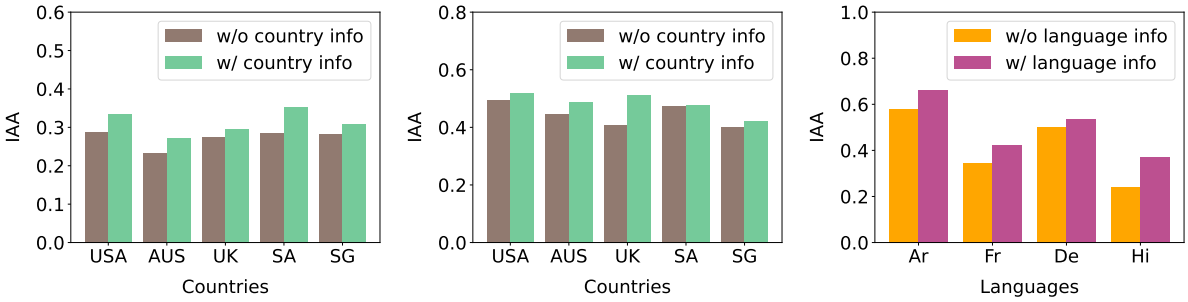
4 Do LLMs Pick on Geographical Cues?

Background. Humans from different countries are prone to variability when flagging the same post as 'hateful' or 'non-hateful' (Lee et al., 2023). Given that AI-assisted content moderation systems often have access to geolocation and language markers along with a post, we investigate whether *'akin to humans, do geographical cues influence LLM's predictions for an incoming hateful post.'*

Setup. Inspired by Lee et al. (2023), we reproduce and extend their analysis on the language tag and multilingual datasets. As described in Section 3, we modify their prompt to suit our format. We compare the change in IAA when considering the human annotator of the respective country vs. p_{base} against the human annotator vs. $p_{con} + p_{base}$ or $p_{lang} + p_{base}$. Here, p_{con} = *'The following statement was made in <country>.'*, where $country$ = *{United States (USA), Australia (AUS), United Kingdom (UK), South Africa (SA), Singapore (SG)}*. Meanwhile, p_{lang} = *'The following statement was made in <lang> language.'*, where $lang$ =*{Arabic (Ar), French (Fr), German (De), Hindi (Hi)}*. For analyzing p_{con}, we employ CREHate for both FlanT5-XXL and GPT-3.5. For p_{lang}, we employ the HASOC-2020 (German & Hindi) and MLMA (Arabic & French) on GPT-3.5's English and language-specific prompting.

Insights. We discuss the results in two broad settings: for country and language tags.

Country. From Figure 3 (b) with p_{con} + p_{base}, GPT-3.5 exhibits disparity in lack of context about social constructs of the Global South (SA, SG) compared to higher alignment with so-called Western nations (USA, UK, AUS) (Zhou et al., 2022; Li et al., 2022; Lee et al., 2023). Surprisingly, for FlanT5-XXL (Figure 3 (a)), we observe an improvement in all countries. It is further corroborated by significance testing on FlanT5-XXL, where we observe (Appendix D) that adding the country cue leads to a notable change in output. One hypoth-



(a) FlanT5-XXL (b) GPT-3.5 (c) GPT-3.5

Figure 3: [RQ1] (a-b) The IAA w.r.t human annotation for each country FlanT5-XXL and GPT-3.5, respectively, for English posts. (c) Captures each language's IAA w.r.t human labels via GPT-3.5 with posts in the language and prompts in English. Here, without (w/o) is p_{base} and with (w/) is p_{con}/p_{lang} + p_{base}.

esis for the difference in p_{con} results can be the nature of training the LLM. Due to only instruction tuning of FlanT5-XXL, it develops country-specific bias implicitly from the training data; however, for GPT-3.5 the implicit bias is augmented with explicit human feedback. It surely calls into question how LLM pretraining mechanisms impact the subjective (non-GLEU) downstream tasks such as hate speech detection (Roy et al., 2023).

Language. From Figure 3 (c), we conjecture that p_{lang} nudges GPT-3.5 to language-specific subspaces leading to a visible higher IAA with ground truth labels when we add p_{lang}. As we infer in Appendix A, the delta increase in performance in p_{lang} + p_{base} vs. p_{base} follows the same order of magnitude irrespective of whether the prompts are in English or the respective language.

Takeaways. Firstly, our observations emphasize the fact that both corpus and paradigm in training play a role in the geographical sensitivity of the LLMs. Judging why some cues and LLMs garner more improvement than others is challenging to access from prompting alone. This calls for more transparency in the LLM training to decode the geographical bias during training epochs. Secondly, as an actional insight, the findings encourage incorporating geographical cues in the zero-shot prompt to ensure higher human-LLM alignment. Interestingly, even if the prompt is in the respective language, explicitly nudging helps.

5 Can LLMs Mimic Annotator Persona?

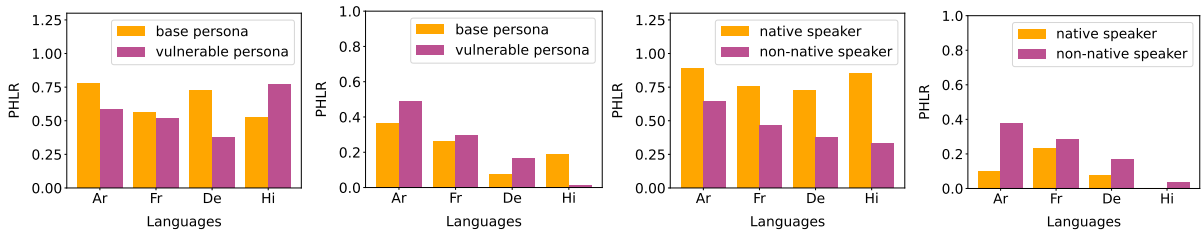
Background. Socio-cultural experiences of humans color their outlook about hate and cause variation in annotations (Sap et al., 2019; Orlikowski et al., 2023). However, without direct signals about mental state, we rely on demographic attributes as markers of human conditioning (Aroyo et al., 2019b; Sap et al., 2022) a.k.a *personas*. Meanwhile, LLMs only possess statistical socio-cultural experience. When employing LLMs for crowdsourcing, we need to assess *'Do LLM's emulation of demographics at a fine-grained level leads to variations in annotating hate?'*. We further hypothesize that differences in a vulnerable group's projection) can lead to variability in the hate perceptiveness of LLMs. We consider a vulnerable group as those who have historically been mistreated based on identity (Kulkarni et al., 2023).

Setup. Borrowing the annotator demographic list from CREHate, we define our persona attributes in terms of D={*Gender* (D_g), *Ethnicity* (D_e), *Political Orientation* (D_p), *Religion* (D_r), *Education Level* (D_q)}. Each demographic in $D_* \in D$ has further sub-classes as outlined in Table 4. For each D_*, we augment the prefix p_{trait} ='*A person who is <D*>, annotated the following statement as <H/N>.*' Operated via the <H/N> tag, we run two persona variants, calling the annotated statement either hateful p^H_{trait} or non-hateful p^N_{trait}. We also examine a third variant p^A_{trait} with the prompt p^A_{trait} ='*Suppose you are a person who is <D> and given the following statement to annotate.*' The 3 personas prompts (p^*_{trait} + p_{base}) are examined via the CREHate dataset on FlanT5-XXL and GPT-3.5.

Intrigued by the success of the language tag in RQ1 (Section 4), we deep dive into persona traits that closely represent each language's demographic. Here, $\forall l \in lang$ = {*Arabic, French, German, Hindi*}, we introduce prompts p^{L*}_{trait} = '*The following statement is in <l> language, and a <I> annotated this as <H/N>.*' In the prompt, I enlists a base/majoritarian persona vs. a vulnerable/minority persona of the respective geography examined on GPT-3.5. Steps for obtaining the following personas are provided in Appendix B.

Annotator demographics	Sub-classes	Flan-T5-XXL						GPT-3.5					
		p^H_{trait}		p^N_{trait}		p^A_{trait}		p^H_{trait}		p^N_{trait}		p^A_{trait}	
		IAA	PHLR	IAA	PHLR	IAA	PHLR	IAA	PHLR	IAA	PHLR	IAA	PHLR
Gender	Male	0.42	0.53	0.00	0.00	0.31	0.29	0.40	0.70	0.55	0.44	0.57	0.46
	Female	0.42	0.53	0.00	0.00	0.33	0.31	0.39	0.72	0.46	0.35	0.52	0.54
	Non-binary	0.42	0.42	0.01	0.01	0.32	0.29	0.31	0.77	0.45	0.38	0.53	0.58
Ethnicity	Asian	0.46	0.56	0.03	0.02	0.33	0.23	0.37	0.75	0.55	0.51	0.51	0.59
	Black	0.43	0.61	0.03	0.01	0.33	0.23	0.37	0.74	0.54	0.51	0.50	0.64
	Hispanic	0.45	0.56	0.03	0.01	0.36	0.24	0.39	0.71	0.56	0.49	0.51	0.62
	Middle Eastern	0.46	0.52	0.03	0.01	0.29	0.19	0.40	0.70	0.54	0.54	0.49	0.64
	White	0.46	0.54	0.03	0.01	0.36	0.24	0.40	0.69	0.51	0.57	0.52	0.56
Political orientation	Liberal	0.42	0.52	0.01	0.01	0.32	0.29	0.49	0.63	0.54	0.53	0.61	0.53
	Indepedent	0.43	0.58	0.00	0.00	0.32	0.29	0.46	0.65	0.58	0.48	0.54	0.50
	Conservative	0.44	0.53	0.00	0.00	0.31	0.26	0.49	0.63	0.53	0.56	0.48	0.47
Religion	Christian	0.39	0.53	0.03	0.03	0.33	0.31	0.44	0.68	0.53	0.58	0.53	0.48
	Buddhism	0.41	0.54	0.03	0.03	0.32	0.30	0.45	0.66	0.52	0.59	0.51	0.45
	Islam	0.43	0.49	0.02	0.01	0.37	0.26	0.47	0.65	0.52	0.54	0.52	0.53
	Judaism	0.45	0.48	0.02	0.01	0.30	0.22	0.46	0.68	0.50	0.60	0.52	0.55
	Hinduism	0.42	0.49	0.04	0.03	0.32	0.24	0.48	0.65	0.52	0.53	0.51	0.49
	Irreligious	0.40	0.39	0.02	0.02	0.30	0.25	0.44	0.68	0.54	0.57	0.54	0.43
Education level	<High school	0.42	0.46	0.01	0.01	0.31	0.30	0.52	0.60	0.54	0.51	0.50	0.49
	High school	0.41	0.52	0.00	0.01	0.29	0.29	0.45	0.65	0.56	0.52	0.53	0.50
	College	0.42	0.52	0.00	0.01	0.29	0.29	0.44	0.67	0.53	0.44	0.52	0.46
	Bachelor	0.42	0.52	0.00	0.00	0.28	0.29	0.43	0.68	0.56	0.47	0.50	0.44
	Master's	0.42	0.52	0.00	0.00	0.28	0.29	0.41	0.69	0.54	0.45	0.48	0.46
	PhD	0.42	0.51	0.00	0.01	0.27	0.28	0.47	0.65	0.53	0.42	0.51	0.49

Table 4: [RQ2] The inter-annotator agreement (IAA) and predicted hate label ratio (PHLR) w.r.t majority voted gold label in `CREHate`, for `FlanT5-XXL` and `GPT-3.5`. The demographic attributes (D) are compared under the hateful p^H_{trait}, non-hateful p^N_{trait}, and assumed persona p^A_{trait} settings. $\forall D_* \in D$, we combine the $p^*_{trait} + p_{base}$.



(a) `GPT-3.5` (H) (b) `GPT-3.5` (N) (c) `GPT-3.5` (H) (d) `GPT-3.5` (N)

Figure 4: [RQ2] Predicted hate label ratio (PHLR) from `GPT-3.5` comparing $p^{L_*}_{trait} + p_{base}$ for Arabic, French, German, and Hindi. (a) and (b) capture the base vs. vulnerable persona for hate/non-hate queries, respectively. (c) and (d) capture the native vs. non-native speaker persona for hate/non-hate queries, respectively.

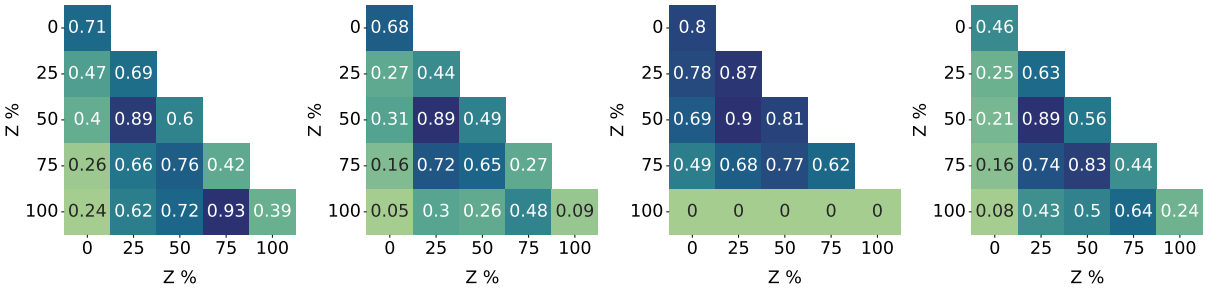
1. For Arabic, $p^{L_{Ar}}_{trait} \in$ {*Muslim/Non-mulsim*}
2. For French, $p^{L_{Fr}}_{trait} \in$ {*French/Mediterranean descent*}
3. For German, $p^{L_{De}}_{trait} \in$ {*Native/Non-native German speaker*}
4. For Hindi, $p^{L_{Hi}}_{trait} \in$ {*Upper/Lower caste*}

We examine the hateful (H) and non-hateful (N) queries here as well, both with English and multilingual prompts. As none of the datasets provide demographic-specific labels, we rely on IAA among predicted and the majority-voted gold labels for our assessment. For $p^{H/N}_{trait}$, we provide in the prompt if the persona identifies a statement as hate/non-hate speech. Our objective here is to assess the role of these traits in persuading LLMs to increase/decrease the number of hate labels in their responses. We thus utilize the Predicted Hate Label Ratio (PHLR) metric in addition to IAA.

Insights. We discuss the results broadly for English and multilingual datasets.

English: Table 4 corroborates that nudging the model to assume a persona (p^A_{trait}) is different from presenting the LLMs with a persona ($p^{H/N}_{trait}$). Further, it is evident from the predicted hate-label ratio (PHLR) that LLMs are sensitive towards some demographic subclasses more than others. Significance testing corroborates the same (Appendix D).

Under Gender demographic for `GPT-3.5`, the percentage of hate labels is higher for 'Non-binary' than 'Males' (for p^H_{trait}). It aligns with the former being a more susceptible subclass of gender in the real world. Meanwhile, for `FlanT5-XXL`, the presence of the non-hate tag p^N_{trait} dominates the demographic information in the context. However, the opposite is not valid for p^H_{trait}. Despite the ground labels associated with samples being balanced across classes, p^N_{trait} for `FlanT5-XXL` still predicts the majority of the labels as 'non-hate.'



(a) FlanT5-XXL (H) (b) GPT-3.5 (H) (c) FlanT5-XXL (N) (d) GPT-3.5 (N)

Figure 5: [RQ3] For the p^{H}_{vote} : (a) and (b) capture the IAA among various hateful voting percentages, i.e., $z\%$ for FlanT5-XXL and GPT-3.5, respectively. For p^{N}_{vote}: (c), (d) function analogously to (a), (b), respectively.
Note: diagonals in heatmaps represent IAA between p_{base} and p_{vote} + p_{base}. Meanwhile, when x and y are different, it represents IAA b/w ($P_{vote} = x + P_{base}$) and ($P_{vote} = y + P_{base}$).

Multilingual. From Figure 4 (a-b), we hypothesize that explicitly known vulnerabilities like Islamophobia (Arabic + Muslim) and Casteism (Hindi + Lower caste) are better captured by the LLM than ethnicity, leading to higher sensitivity of these pairs towards hate when prompted in English. While the patterns persist under multilingual prompting for the rest of the languages, its lack of variation in p^{H}_{trait} Hindi is puzzling (Table 6 in Appendix B).

As the highest gap in p^{*}_{trait} English is observed for German, we deep dive into this prompt (native speaker vs non-native speaker) and repeat this for all languages under English and multilingual prompting. From Figure 4 (c-d) and Table 7 in Appendix B, one can conjecture that when a native-speaking persona considers a post as hateful, the model may be contextualizing the higher acuity of the speaker to understand the geographical context.

Takeaways. Given the sensitivity of LLMs toward the combination of persona + label, we urge the practitioners not mindlessly to rely on LLM as a substitute for demographic attributes. The persistence of results in English and multilingualism is an advantage for the researchers working on multilingual datasets. It can allow faster assessment of the LLMs without requiring separate prompt engineering for both setups.

6 Are LLMs Sensitive to Anchoring Bias?

Background. Anchoring bias occurs when humans rely too heavily on the anchor information (relevant or not) to influence their decision-making. While some anchors like language and geolocation, as established in RQ1 (Section 4), are helpful for LLM-based content moderation, the influence of numerical cues under the zero-shot setting is unknown. Numerical features/cues can be defined as aggregated real-world or simulated values representing countable metadata associated with the posts. It can range from the total/average views, likes, and comments a post/user receives to the number of people who have in the past reported/flagged the post/user as hateful. The former set of metadata is publicly available and has been successfully employed in hate speech detection to improve finetuned accuracy (Founta et al., 2019; Kulkarni et al., 2023). Meanwhile, the influence of previously flagged counts is not known, as such metrics are not available to the public. Here, using voting statistics as a proxy for the crowd's opinion about the post, we are motivated to examine whether '*made-up voting percentages in the prompt lead to manipulation in LLM's output?*'

Setup. The base prompt p_{base}, is prefixed with $p^{H/N}_{vote}$ = '*The following statement was labeled <H/N> by <z> annotators.*', where $z \in \{0\%, 25\%, 50\%, 75\%, 100\%\}$. The two variants $p^{H/N}_{vote}$ capture the hateful (H) or non-hateful voting (N) label. We represent z in the percentage to give a relative sense of majority voting. At $z = 50\%$ saying, 50%, annotators consider the post is hateful, and 50% say it is non-hateful is not the same as p_{base}. We conduct this experiment with HateXplain on FlanT5-XXL and GPT-3.5 with p_{vote} + p_{base}. The $z\%$ alludes to annotators in general and not a specific persona. Further, it should be noted that these percentages are not available as a part of HateXplain, and added by us to introduce the "anchoring" information.

Insights. In line with existing evidence (Zhao et al., 2021) of majority labeling bias in few-shot learning, we also establish that LLMs are prone to labeling bias even under zero-shot settings if the context mentions voting percentages. From Figures

5 (a-b) regarding p^H_{vote}, it is evident that alignment in hate labeling is more consistent when the percentages lie closer to each other and decrease as one moves away. Succinctly, $IAA_{ij} > IAA_{ik}$ if $|z_j - z_i| < |z_k - z_i|$. A similar pattern is observed for p^N_{vote} (Figures 5 (c-d)).

Takeways. Corroborated by significance testing (Appendix D) and controlling for decoding temperature (Appendix C), we discover that while LLMs understand relative percentages, they are prone to emphasize this information over the post's content. As in our RQ, voting serves as a proxy of numerical metadata in the content moderation pipeline; it implies that corruption (accidental or intentional) can lead to instance misclassification. Despite numerical/statistical features being used earlier to enhance the finetuning in hate speech detection (Founta et al., 2019; Kulkarni et al., 2023), they cannot be directly extrapolated as features in prompting for hate speech annotations.

7 Discussion

This section summarizes the implications of the research. We also provide recommendations for LLM-assisted content moderation. Except for the `FlanT5-XXL` non-hate persona, the results across models, datasets, languages, and prompts are aligned. Cognizant of adversarial attacks on prompting, the study balances both feature importance (corroborated by significance testing) and cautions against adversary features. Despite how the perception of hatefulness manifests via p_{base}, we observe that the inclusion of cues nudges the LLMs to explore a contextual definition of hate.

Geographical Sensitivity. How the latent spaces are triggered at the mention of geographical cues (Zhou et al., 2022) is intractable from prompting alone. This becomes especially tricky for overlapping signals like 'Arabic,' 'Muslims,' and 'Islamophobia' (Figure 3 (c)). It once again highlights the need for more transparency in the training and cultural alignment of LLMs. As a quick fix, we suggest augmenting the language/country tags for an incoming post being prompted to improve alignment with people representative of a given region/geography.

Demographic Sensitivity. Our work establishes that the manner of personification of demographic attributes can lead to variation in hate labeling, with assumed persona (p^A_{trait}) being closer to the inherent knowledge and biases an LLM possesses. Here, we caution practitioners looking to adopt LLMs for crowdsourcing to experiment with different framing of the personas and identity traits.

Numerical Sensitivity. As evident from our experiments, LLMs do not have a clear way of discovering noise from the informativeness in the prompt. Our advice to practitioners is to refrain from adding any numerical statistics in the prompt unless they are quality-checked and to mask such adversarial expressions written by the authors of the input samples.

Multilingual Sensitivity. Following the same prompt format, we observe similar performance patterns across both English and Multilingual settings (refer Table 5 in Appendix B). The results are encouraging, allowing content moderators to work in the native language of the post without extraneous prompt engineering. The degradation in performance from English to Multilingual (Jin et al., 2024), however, calls for more investment in non-English training and evaluation of LLMs.

LLMs for Hate Speech Annotations. It is imperative to point out that hate speech is a human-centric phenomenon steeped in historical and cultural contexts. As such, any computational attempt to flag it can only be assistive. Our experiments over three contextual settings indicate that LLM cannot outright substitute a demographic group in the annotation process (Section 5). Further, if LLMs are prone to anchoring, that is not always beneficial. Then, our study opens up the questions around using few-shot/guideline examples for labeling hate (Sections 5 and 6).

8 Conclusion

In the current LLM space, both alignment and robustness do not have a singular non-overlapping definition, which makes our study sit at the intersection of robustness in hate speech detection as well as conceptualization of the human-LLM alignment. Over multiple RQs and prompt setups, we explore how well-suited LLMs are for assisting humans at different stages of the content moderation pipeline. Our results on demographic sensitivity, cultural priming, and anchoring bias are evident over multiple annotations on 6 datasets, 5 languages, and 2 LLMs. Our analysis reiterates using LLMs as an assistive system rather than substituting human moderators. We would like to work on more datasets and explore intersectional demographic attributes in the future.

9 Limitations

First and foremost, the list of research questions and prompts analyzed, while significant in number, is not exhaustive. Given that the adoption of LLMs in the hate speech moderation pipeline is nascent, the scope for research is more vast than what we can account for in current research. We hope our findings and primary analysis motivate future research. Besides, there is a shortage of open-source instruction-tuned multilingual LLMs, which restricts our multilingual and multicultural analysis to `GPT-3.5`. Owing to the lack of datasets with ground truth labels where multiple annotators are considered and demographic-specific annotators are released, performing one-to-one mapped inter-annotator agreement analysis between LLMs and demographic personas is challenging. The current study only utilizes textual content. In the future, we need to conduct assessments for content moderation on other modalities like memes, short videos, etc. It should be noted that geography or language only acts as a proxy for cultural and linguistic recalls and does not represent an absolute assessment for multilingual LLM evaluations. Consequently, when employing geographical or demographic attributes, we need to be cognizant of the fact that the sensitivity of the LLMs can also be a source of their implicit bias.

10 Ethical Considerations

This work does not produce any new scientific artifacts regarding datasets or proposed models, and no human evaluators were involved. Our work utilizes LLMs to annotate (Gilardi et al., 2023) hate speech, which comes with legal, technical, and ethical challenges. There is an inherent issue of magnifying human behaviors and biases by LLMs. Our findings corroborate the sensitivity of LLMs towards different demographics and identity traits. Hence, their generations/predictions should be taken cautiously. Further, this study is underpinned by gold labels obtained from human annotators, which themselves could be erroneous. The intent of releasing our prompt list is to broaden the examination of biases and fallacies an LLM is prone to. While `FlanT5-XXL` is an opensource LLM, `GPT-3.5` is not, which raises concerns around reproducibility (Ollion et al., 2024), and is a broader challenge for the NLP community.

Acknowledgements

Sarah Masud acknowledges the support of the Prime Minister Doctoral Fellowship in association with Wipro AI and Google PhD Fellowship. Viktor Hangya and Alexander Fraser acknowledge funding from the German Research Foundation (DFG; grant FR 2829/7-1). Tanmoy Chakraborty acknowledges Anusandhan National Research Foundation (CRG/2023/001351) for the financial support. The authors thank Ahmed Tarek Sobhi Abdelrahman and Nicolas Payen for validating the French, German, and Arabic prompts.

References

Gavin Abercrombie, Dirk Hovy, and Vinodkumar Prabhakaran. 2023. Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *Proc. of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 96–103, Toronto, Canada.

Badr AlKhamissi, Faisal Ladhak, Srinivasan Iyer, Veselin Stoyanov, Zornitsa Kozareva, Xian Li, Pascale Fung, Lambert Mathias, Asli Celikyilmaz, and Mona Diab. 2022. ToKen: Task decomposition and knowledge infusion for few-shot hate speech detection. In *Proc. of the 2022 Conference on EMNLP*, pages 2109–2120, Abu Dhabi, United Arab Emirates. ACL.

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019a. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, page 1100–1105, New York, NY, USA. Association for Computing Machinery.

Lora Aroyo, Anca Dumitrache, Oana Inel, Zoltán Szlávik, Benjamin Timmermans, and Chris Welty. 2019b. Crowdsourcing inclusivity: Dealing with diversity of opinions, perspectives and ambiguity in annotated data. In *Companion Proceedings of The 2019 World Wide Web Conference*, page 1294–1295, New York, NY, USA. Association for Computing Machinery.

Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. 2022. Analyzing the effects of annotator gender across NLP tasks. In *Proc. of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 10–19, Marseille, France. European Language Resources Association.

Alexander Braylan and Matthew Lease. 2020. Modeling and aggregation of complex annotations via annotation distances. In *Proc. of The Web Conference 2020*, WWW '20, page 1807–1818, New York, NY, USA. Association for Computing Machinery.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 105–114, New York, NY, USA. Association for Computing Machinery.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Comput. Surv.*, 55(13s).

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

Lewis Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly Mai, Maria Do Mar Vau, Matthew Caldwell, and Augustine Mavor-Parker. 2023. Large language models respond to influence like humans. In *Proc. of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 15–24, Toronto, Canada. ACL.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2023. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 294–297, New York, NY, USA. Association for Computing Machinery.

Musa İhtiyar, Ömer Özdemir, Mustafa Erengül, and Arzucan Özgür. 2023. A dataset for investigating the impact of context for offensive language detection in tweets. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1543–1549, Singapore. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM on Web Conference 2024*, WWW '24, page 2627–2638, New York, NY, USA. Association for Computing Machinery.

Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, Piotr Milkowski, Jan Kocon, and Przemyslaw Kazienko. 2022. What if ground truth is subjective? personalized deep neural hate speech detection. In *Proc. of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 37–45, Marseille, France. European Language Resources Association.

Aman Khullar, Daniel Nkemelu, Cuong V. Nguyen, and Michael L. Best. 2023. Hate speech detection in limited data contexts using synthetic data generation. *ACM J. Comput. Sustain. Soc.*

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating lexical features with BERT for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

Atharva Kulkarni, Sarah Masud, Vikram Goyal, and Tanmoy Chakraborty. 2023. Revisiting hate speech benchmarks: From data curation to system deployment. In *Proc. of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 4333–4345, New York, NY, USA. Association for Computing Machinery.

Thai Le, Long Tran-Thanh, and Dongwon Lee. 2022. Socialbots on fire: Modeling adversarial behaviors of socialbots via multi-agent hierarchical reinforcement learning. In *Proc. of the ACM Web Conference 2022*,

WWW '22, page 545–554, New York, NY, USA. Association for Computing Machinery.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Juho Kim, and Alice Oh. 2023. CReHate: Cross-cultural Re-annotation of English Hate Speech Dataset. *arXiv preprint arXiv:2308.16705*.

Dan Li, Zhaochun Ren, and Evangelos Kanoulas. 2021. Crowdgp: A gaussian process model for inferring relevance from crowd annotations. In *Proc. of the Web Conference 2021*, WWW '21, page 1821–1832, New York, NY, USA. Association for Computing Machinery.

Yizhi Li, Ge Zhang, Bohao Yang, Chenghua Lin, Anton Ragni, Shi Wang, and Jie Fu. 2022. HERB: Measuring hierarchical regional bias in pre-trained language models. In *Findings of the ACL: AACL-IJCNLP 2022*, pages 334–346, Online only. ACL.

Ruixi Lin and Hwee Tou Ng. 2023. Mind the biases: Quantifying cognitive biases in language model prompting. In *Findings of the ACL: ACL 2023*, pages 5269–5281, Toronto, Canada. ACL.

Nikola Ljubešić, Igor Mozetič, and Petra Kralj Novak. 2022. Quantifying the impact of context on the quality of manual hate speech annotation. *Natural Language Engineering*, pages 1–14.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *Proc. of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 29–32.

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.

Luke Munn. 2020. Angry by design: toxic communication and technical architectures. *Humanities and Social Sciences Communications*, 7(1):53.

Venkata Prabhakara Sarath Nookala, Gaurav Verma, Subhabrata Mukherjee, and Srijan Kumar. 2023. Adversarial robustness of prompt-based few-shot learning for natural language understanding. In *Findings of the ACL: ACL 2023*, pages 2196–2208, Toronto, Canada. ACL.

Étienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2024. The dangers of using proprietary LLMs for research. *Nature Machine Intelligence*, 6(1):4–5.

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

Lidiia Ostyakova, Veronika Smilga, Kseniia Petukhova, Maria Molchanova, and Daniel Kornev. 2023. ChatGPT vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions. In *Proc. of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 242–254, Prague, Czechia. ACL.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proc. of EMNLP*. ACL.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Russell Richie, Sachin Grover, and Fuchiang (Rich) Tsui. 2022. Inter-annotator agreement is not the ceiling of machine learning performance: Evidence from a comprehensive set of simulations. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 275–284, Dublin, Ireland. Association for Computational Linguistics.

Bjorn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In *NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, Bochumer Linguistische Arbeitsberichte, pages 6–9, Germany. Ruhr-Universitat Bochum.

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

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore. Association for Computational Linguistics.

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

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proc. of the 57th Annual Meeting of the ACL*, pages 1668–1678, Florence, Italy. ACL.

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 *Proc. of the 2022 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 5884–5906, Seattle, United States. ACL.

P Slovic, B Fischhoff, and S Lichtenstein. 1977. Behavioral decision theory. *Annual Review of Psychology*, 28(Volume 28, 1977):1–39.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *ArXiv*, abs/2305.17926.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proc. of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. ACL.

Tharindu Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher Homan, and Ashiqur KhudaBukhsh. 2023. Vicarious offense and noise audit of offensive speech classifiers: Unifying human and machine disagreement on what is offensive. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11648–11668, Singapore. Association for Computational Linguistics.

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

Michael Maximilian Wojatzki. 2018. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments. In *Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)*. Deutsche Nationalbibliothek.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. HARE: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore. Association for Computational Linguistics.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In *Proc. of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proc. of the 38th International Conference on Machine Learning*, volume 139 of *Proc. of Machine Learning Research*, pages 12697–12706. PMLR.

Kaitlyn Zhou, Kawin Ethayarajh, and Dan Jurafsky. 2022. Richer countries and richer representations. In *Findings of the ACL: ACL 2022*, pages 2074–2085, Dublin, Ireland. ACL.

Julia El Zini and Mariette Awad. 2022. On the explainability of natural language processing deep models. *ACM Comput. Surv.*, 55(5).

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A RQ1 Multilingual Prompting

We run additional experiments where the cues and query are in the same language as the multilingual post. The comparative results are enlisted in Table 5. Despite the expected loss in performance (Jin et al., 2024), the inclusion of p_{lang} leads to a better human-AI alignment.

B RQ2 Multilingual Persona Prompting

Personas. We start with Wikipedia to get a general sense of the demographics of each geography. From there, we narrow down the most prominent demography of the nation. We further use census and news articles and consult social experts about each geography before narrowing down the vulnerable minority based on the hypothesis that minority groups get more hate than the majority.

Prompting. From Tables 6 and 7, we again observe the same pattern as in English, thereby showcasing that the role of context is significant irrespective of the language under consideration. Although we have seen a similar pattern in PHLR, there is a significant degradation in the observed metrics compared to English.

C RQ3 Temperature Probing

We simulate multiple annotators under each z value to further corroborate the results. Due to resource constrain, we run this analysis on only `FlanT5-XXL`. We generate 100 output for each sample, by uniformly sampling 100 decoder temperature values ($t \in (0, 2)$). For reference, in p_{base}, we obtain a mean (std. dev) percentage of hate label as $0.560(\pm 0.0169)$. While the spread is low for p_{base}, we observe from Figure 6 (a) and (b) that for some z, the spread varies. It is an indicator that decoding temperature can lead to variations in the LLM's predictions when prompted on numerical anchors. We observe more spread on average when using p_{vote} as recorded in Table 8. For p^H_{vote}, with varying temperatures, as the value of z increases, the percentage of hate predicted increases, as evident from the shift towards the right for $z = 100\%$ in Figure 6 (a). The reverse trend is observed for p^N_{vote} in Figure 6 (b) where the curve for $z = 100\%$ is left shifted, leading to a decrease in the percentage of hate predicted as the majority of non-hateful increases.

D Statistical Testing on FlanT5-XXL

Process. For RQ1 (Sec. 4) and RQ3 (Sec. 6), we perform the paired t-test and report the effect size. Meanwhile, for RQ2 (Sec. 5), to capture the intra-demography disparity among the subclasses, we use ANOVA. All experiments are run via the SciPy and NumPy libraries.

Observations. For all the RQs, we observe significant disparity caused by context.

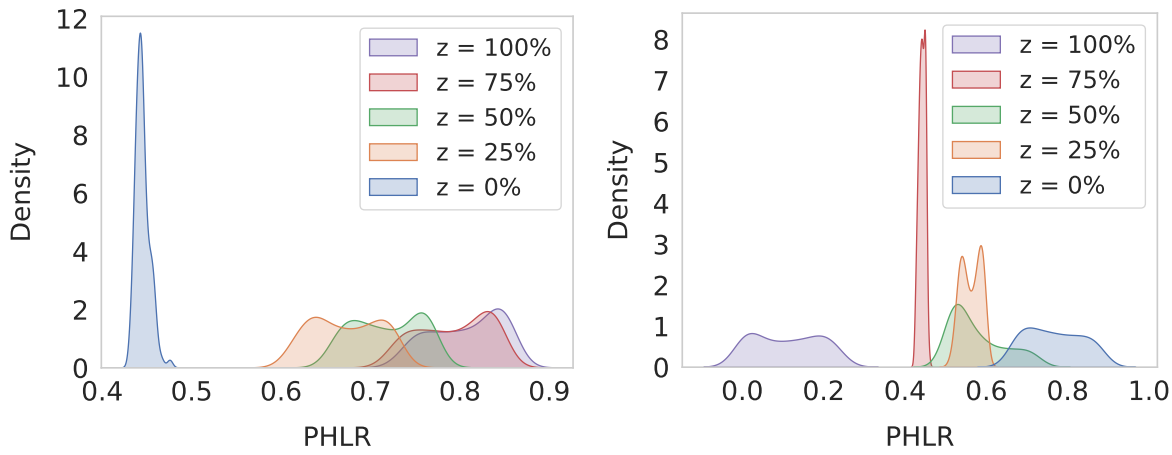
- **RQ1:** From Table 9, comparing both adjusted F1 and IAA, we observe that adding the country cue leads to a significant change in the prompted output (as captured by the higher effect size (ES) as well as the p-values). Following the reference Figure 3 (a), South Africa registers the highest impact.
- **RQ2:** From Table 10 we observe that different modes of the persona (p^H, p^N, p^A) are impacted by varying sub-classes. Interestingly, the subgroups within Religion show considerable variation for three persona prompts.
- **RQ3:** From Table 11, for both p^H and p^N, we observe significant differences (as captured by the higher effect size (ES) as well as the p-values) in performance among the various percentages in the prompt. This further corroborates the anchoring bias in LLMs employed for zero-shot labeling.

Language	Prompt in English		Prompt in same language	
	p_{base}	p_{lang}	p_{base}	p_{lang}
Arabic	0.580	0.660	0.140	0.305
French	0.344	0.425	0.272	0.356
German	0.502	0.537	0.412	0.423
Hindi	0.242	0.371	0.018	0.031

Table 5: [RQ1] Extension of Figure 3 (c) for IAA comparison of the results without and with language cues for prompts in English and the respective language.

Language	Majority or vulnerable	Prompt in English		Prompt in same language	
		p^H	p^N	p^H	p^N
Arabic	Muslim	0.778	0.364	0.992	0.737
	Non-muslim	0.586	0.492	0.525	0.483
French	French descent	0.558	0.262	0.666	0.170
	Mediterranean descent	0.520	0.298	0.649	0.176
German	Native	0.724	0.076	0.566	0.152
	Non-native	0.374	0.170	0.248	0.248
Hindi	Upper caste	0.528	0.192	0.998	0.282
	Lower caste	0.768	0.014	0.998	0.094

Table 6: [RQ2] PHLR for comparison of vulnerable persona cues. Extension of Figure 4 (a-b).



(a) `FlanT5-XXL` temperatures (H)

(b) `FlanT5-XXL` temperatures (N)

Figure 6: [RQ3] The impact of decoding temperature with varying voting percentages for $p^H_{context}$ on `FlanT5-XXL`.

Language	Speaker	Prompt in English		Prompt in same language	
		p^H	p^N	p^H	p^N
Arabic	Native	0.890	0.100	0.764	0.099
	Non-native	0.646	0.380	0.567	0.901
French	Native	0.752	0.232	0.916	0.130
	Non-native	0.462	0.286	0.702	0.106
German	Native	0.724	0.076	0.566	0.152
	Non-native	0.374	0.170	0.248	0.248
Hindi	Native	0.852	0.004	1.000	0.753
	Non-native	0.328	0.038	1.000	0.858

Table 7: [RQ2] PHLR for comparison of speaker persona cues. Extension of Figure 4 (c-d).

<z>	Distribution parameters ($\mu \pm \sigma$)	
	p^H_{vote}	p^N_{vote}
0%	0.445 ± 0.0078	0.77 ± 0.067
25%	0.67 ± 0.037	0.56 ± 0.024
50%	0.72 ± 0.037	0.57 ± 0.065
75%	0.79 ± 0.039	0.443 ± 0.0078
100%	0.81 ± 0.038	0.11 ± 0.079

Table 8: [RQ3] Mean and standard deviation of temperature distribution in Figure 6.

Country	ES F1	ES IAA
United States	1.116*	1.165*
Australia	0.731*	0.729*
United Kingdom	0.723*	0.707*
South Africa	1.244*	1.495*
Singapore	0.782*	0.834*

Table 9: [RQ1] Effect size to indicate the significance of including the country cue for `FlanT5-XXL` in English. *($p \leq 0.05$) and **($p \leq 0.001$) indicate whether the difference is significant.

Annotator demographics	p^H	p^N	p^A
Gender	0.010**	0.007	0.013*
Ethnicity	0.031**	0.027*	0.070
Political orientation	0.021*	0.011	0.007**
Religion	0.054**	0.021**	0.075**
Education level	0.013*	0.003	0.041

Table 10: [RQ2] The absolute difference between the minimum and maximum IAA obtained for the respective persona, demographic, and LLM combination. *($p \leq 0.05$) and **($p \leq 0.001$) indicate if ANNOVA is significant among the sub-classes within a demographic.

Key	Value	p^H		p^N	
		ES F1	ES IAA	ES F1	ES IAA
0	25	2.24*	1.878*	6.379	4.47*
	50	2.732*	2.362*	1.25*	0.233
	75	3.805*	3.475*	4.342*	2.661*
	100	4.037*	3.583*	0.178	1.637**
25	50	1.53*	1.334*	4.416*	3.748*
	75	2.933*	2.626*	1.065*	1.412*
	100	3.23*	2.818*	3.064*	4.91*
50	75	2.516*	2.231*	3.701*	2.6*
	100	2.738*	2.36*	0.451*	1.559*
75	100	0.875*	0.773*	2.922*	3.493*

Table 11: [RQ3] Effect size for percentage pairs 'x' (key) and 'y' (value) to indicate the significance of including the percentage cues. *($p \leq 0.05$) and **($p \leq 0.001$) indicate if the difference is significant or not.

RQ	Prompt type	Prompt(s)
Base	p_{base}	Statement: <POST>.
RQ_1	p_{con}	The following statement was made in <c>: <POST>.
	p_{lang}	The following statement was made in <l> language: <POST>.
RQ_2	$p^{H/N}_{trait}$	A <g> annotated the following statement as <H/N>: <POST>. A person of <e> ethnicity annotated the following statement as <H/N>: <POST>. A person of <p> political orientation annotated the following statement as <H/N>: <POST>. A person who follows <r> annotated the following statement as <H/N>: <POST>. An irreligious person annotated the following statement as <H/N>: <POST>. A person with an education level less than high school annotated the following statement as <H/N>: <POST>. A person with education level of <q> annotated the following statement as <H/N>: <POST>.
	p^{A}_{trait}	Suppose you are a <g> and given the following statement to annotate: <POST>. Suppose you are a person of <e> ethnicity and given the following statement to annotate: <POST>. Suppose you are a person of <p> political orientation and given the following statement to annotate: <POST>. Suppose you are a person who follows <r> and given the following statement to annotate: <POST>. Suppose you are an irreligious person and given the following statement to annotate: <POST>. Suppose you are a person with an education level less than high school and given the following statement to annotate: <POST>. Suppose you are a person with education level of <q> and given the following statement to annotate: <POST>.
	p^{L*}_{trait}	The following statement is in <l> langauge and a <B/V> annotated this as <H/N>: <POST>. The following statement is in <l> langauge and a <N/NN> speaker annotated this as <H/N>: <POST>.
RQ_3	$p^{H/N}_{vote}$	The following statement was labeled <H/N> by <z> annotators: <POST>.

Table 12: Prompts employed in this study for all settings under RQ1, RQ2, and RQ3. For `FlanT5` (all), $prompt \leftarrow prompt+$"Is the given statement hateful?" For `Llama3-8B`, `Mistral-7B`, `Zephr-7B` and `GPT-3.5`, $prompt \leftarrow prompt$ + "Is the given statement hateful? Answer in one word only." Here, $c \in$ *{United States, Australia, United Kingdom, South Africa, Singapore}*, $l \in$ *{Arabic, French, German, Hindi}*, $H/N \in$ *{hateful, non-hateful}*, $g \in$ *{male, female, non-binary}*, $e \in$ *{asian, black, hispanic, middle eastern, white}*, $p \in$ *{liberal, indepedent, conservative}*, $r \in$ *{christian, buddhism, islam, judaism, hinduism}*, $q \in$ *{high school, college, bachelor, master's, PhD}*, $B/V \in$ *{{muslim, non-muslim}, {french descent, mediterranean descent}, {native german, non-native}, {person belongs to upper caste, person belongs to lower caste}}*, $N/NN \in$ *{native, non-native}*, $z \in$ *{0%, 25%, 50%, 75%, 100%}*

Post language	Prompt in English language	Prompt in same language
Arabic	The following statement is in Arabic language and a non-muslim annotated this as non-hateful: صحيح كان كلاب ابن كلب مع شعبى بس كان صاكر دول بول البعيروكانت مملكة المنشار ترتجف خوفا اذا ظهر بخطاب بالتلفاز وال... @url. Is the given statement hateful? Answer in one word only.	العبارة التالية مكتوبة باللغة العربية وقد علق غير المسلم على ذلك بأنه غير مكروه: صحيح كان كلاب ابن كلب مع شعبى بس كان صاكر دول بول البعيروكانت مملكة المنشار ترتجف خوفا اذا ظهر بخطاب بالتلفاز وال... @url. هل البيان المعطى يكره؟ الإجابة بكلمة واحدة فقط.
French	The following statement is in French language and a french descent annotated this as hateful: @user @user l'islamisation en marche ! la pauvreté généralisée partout en france. Is the given statement hateful? Answer in one word only.	La déclaration suivante est en langue française et une descendance française l'a annotée comme haineuse: @user @user l'islamisation en marche ! la pauvreté généralisée partout en france. La déclaration faite est-elle haineuse? Répondez en un seul mot.
German	The following statement is in German language and a native german speaker annotated this as non-hateful: @neothemarius @niemehrgruen Nicht zu vergessen, einen moderaten ISlam... Is the given statement hateful? Answer in one word only.	Die folgende Aussage ist in deutscher Sprache und ein deutscher Muttersprachler hat dies als nicht hasserfüllt bezeichnet: @neothemarius @niemehrgruen Nicht zu vergessen, einen moderaten ISlam... Ist die gegebene Aussage hasserfüllt? Antworten Sie nur mit einem Wort.
Hindi	The following statement is in Hindi language and a person belongs to upper caste annotated this as hateful: RT @saitanBacha: @Kanatunga हाथ के कलर से दलित लग रहे हो 😅😅😂😂😂😂. Is the given statement hateful? Answer in one word only.	निम्नलिखित कथन हिंदी भाषा में है और ऊंची जाति के एक व्यक्ति ने इसे घृणित बताया है: RT @saitanBacha: @Kanatunga हाथ के कलर से दलित लग रहे हो 😅😅😂😂😂😂. क्या दिया गया कथन घृणास्पद है? केवल एक शब्द में उत्तर दीजिए।

Table 13: Random verbatim examples of multilingual prompts in the same language as the post (red) for the corresponding cues and queries (black) in English.