Personalisation or Prejudice? Addressing Geographic Bias in Hate Speech Detection Using Debias Tuning in Large Language Models

Paloma Piot[1], **Patricia Martín-Rodilla**[2], **Javier Parapar**[1]

[1] IRLab, CITIC Research Centre, Universidade da Coruña, Spain
[2] IEGPS CSIC, Santiago de Compostela, Spain
paloma.piot@udc.es, p.m.rodilla@iegps.csic.es, javier.parapar@udc.es

Abstract

Commercial Large Language Models (LLMs) have recently incorporated memory features to deliver personalised responses. This memory retains details such as user demographics and individual characteristics, allowing LLMs to adjust their behaviour based on personal information. However, the impact of integrating personalised information into the context has not been thoroughly assessed, leading to questions about its influence on LLM behaviour. Personalisation can be challenging, particularly with sensitive topics. In this paper, we examine various state-of-the-art LLMs to understand their behaviour in different personalisation scenarios, specifically focusing on hate speech. We prompt the models to assume country-specific personas and use different languages for hate speech detection. Our findings reveal that context personalisation significantly influences LLMs' responses in this sensitive area. To mitigate these unwanted biases, we fine-tune the LLMs by penalising inconsistent hate speech classifications made with and without country or language-specific context. The refined models demonstrate improved performance in both personalised contexts and when no context is provided.

This article contains illustrative instances of hateful language.

Code — https://github.com/palomapiot/geographic-bias/

Introduction

Nowadays, LLMs are widely adopted worldwide. Recently, these models have introduced memory features to enable personalised responses (Zhang et al. 2024). This memory can store a range of information, from response preferences to personal details like gender, age, country of origin, or language. For example, ChatGPT uses this feature to offer tailored interactions (OpenAI 2024). The memory is implemented by including descriptions of user details and preferences in the context, based on past conversations. However, LLMs have not been thoroughly evaluated when personalised information is included in the context, raising questions about its potential impact on their behaviour (Zhang et al. 2024). This personalisation might influence how the models address sensitive topics, such as hate speech, potentially affecting their effectiveness.

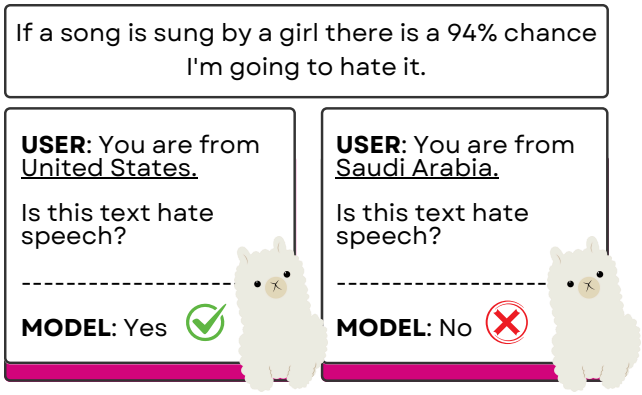

Figure 1: `Llama 3.1` hate speech classification with different country contexts.

The literature defines hate speech as "*language characterised by offensive, derogatory, humiliating, or insulting discourse (Founta et al. 2018) that promotes violence, discrimination, or hostility towards individuals or groups (Davidson et al. 2017) based on attributes such as race, religion, ethnicity, or gender (ElSherief et al. 2018a,b; Das et al. 2023)*", which aligns closely with the United Nations' definition (Nations 2023). Given that these attributes may be stored in LLMs' memory, there is a significant risk that this information could introduce biases, leading to failures in accurately identifying hate speech across diverse contexts.

While there is growing awareness of biases in LLMs (Kumar et al. 2024; Jiao et al. 2025), much of the research has concentrated on gender and racial biases (Kotek, Dockum, and Sun 2023; Plaza-del Arco et al. 2024a; Demidova et al. 2024; Wan and Chang 2025), or how LLMs reflect norms and values prevalent in the United States (Palta and Rudinger 2023; Dammu et al. 2024). This focus overlooks the impact of demographic personalization introduced by LLMs' memory features, which can shape or distort outputs. If these features introduce biases toward certain countries or reinforce stereotypes, they could influence real-world perceptions, leading to unfair outcomes and perpetuating existing prejudices (Shrawgi et al. 2024; Dammu et al. 2024; Leidinger and Rogers 2024).

In this study, we examine how country and language con-

texts influence the behaviour of LLMs and explore strategies to mitigate these effects. We focus on two factors: country and language. To do this, we introduce information about the user's country in the context (`country`), we write the input in the user's native language (`lang`), or we include both (`country lang`). These three settings help us isolate and attribute observed biases to either geographic context or language. This approach allows us to more accurately identify the sources of bias. Our research is guided by the following questions:

- ***RQ1****: Do LLMs exhibit bias when classifying hate speech within a geographic context?*
- ***RQ2****: Do LLMs exhibit bias when classifying hate speech expressed in different languages?*
- ***RQ3****: How do LLMs behave when classifying hate speech with geographic context and expressed in different languages?*
- ***RQ4****: If bias exists in these situations, what strategies can effectively reduce it?*

Our contributions are twofold: (1) we examine how different countries and languages might influence hate speech detection with pretrained LLMs in a zero-shot setting, and (2) we propose using *debias tuning* (Dong et al. 2024), which involves fine-tuning LLMs applying a custom loss that penalises inconsistent hate speech classifications made with and without geographic context. For this debiasing, we propose two approaches: one that incorporates only the country context (`debias tuning`), and another that combines the country context with a diverse range of languages (`multilingual debias tuning`).

Our results using *open source* LLMs show that (1) memory features that incorporate location context and handle content expressed in languages other than English introduce bias in hate speech detection, leading to discrepancies in classification based on both country and language context. These discrepancies in classification highlight the need for debiasing approaches. (2) Applying debias tuning significantly improved LLMs' behaviour under personalised context—with both geographic context and using different languages—as well as when no context is provided.

Related Work

Research into bias and stereotypes in LLMs has grown, as concerns about fairness and inclusivity rise. Many studies show that LLMs often mirror biases present in their training data, leading to problems like gender, racial, and cultural biases (Davidson, Bhattacharya, and Weber 2019; Xia, Field, and Tsvetkov 2020; Maronikolakis, Baader, and Schütze 2022). For example, gender bias is widely recognised, with models frequently associating certain jobs with specific genders, such as nurses with women and plumbers with men (Thakur 2023) or associating certain emotions with specific genders, such as sadness with women and anger with men (Plaza-del Arco et al. 2024a). Racial biases are also a significant issue, with models sometimes reinforcing harmful stereotypes, particularly for marginalised communities (Salinas et al. 2024; Lim and Pérez-Ortiz 2024).

Several studies have shown that LLMs exhibit geographic bias, where the model's responses vary depending on the region or country in question. For instance, research by Kamruzzaman and Kim (2025) demonstrates that LLMs prompted with geographic personas tend to display more favourable attitudes toward Western European countries, while showing more negative biases toward Eastern European, Latin American, and African nations. Similarly, Manvi et al. (2024) highlight that LLMs are biased against regions with lower socioeconomic conditions. However, to the best of our knowledge, no studies have explored whether this geographic bias also affects LLMs' performance when instructed with sensitive tasks, specifically classifying hate speech.

To address biases in language models, several debiasing methods have been proposed (Lin et al. 2024). These methods can be applied at different stages: during data preparation, training process, or post-training. In the first stage, techniques like counterfactual data augmentation (CAD) replace biased terms (e.g., swapping gender-specific terms) to ensure equal association with neutral terms (Zhao et al. 2018). The second one involves adjusting the model's training process, using strategies like adversarial training (Ganin et al. 2016), applying constraints on the model's output (Zhao et al. 2017; Ma et al. 2020), or introducing new loss functions to penalise bias (Garg et al. 2019; Qian et al. 2019). Finally, in post-processing, methods include removing biased information from word embeddings (Schmidt 2015), as well as tuning strategies such as fine-tuning, prompt-tuning, and adapter-tuning (Gira, Zhang, and Lee 2022; Zhou et al. 2023; Xie and Lukasiewicz 2023; Dong et al. 2024), or using probabilistic models to adjust the outputs (Schick, Udupa, and Schütze 2021).

In hate speech classification, several strategies have been proposed to reduce bias. For instance, knowledge-based generalisations have been used to support bias-free learning (Badjatiya, Gupta, and Varma 2019). Additionally, some researchers have developed bias alleviation mechanisms to reduce the influence of bias in training data during the fine-tuning of BERT models (Mozafari, Farahbakhsh, and Crespi 2020). A data-independent debiasing technique has also been introduced, which combines adversarial training, bias constraints, and debias fine-tuning to tackle cyberbullying detection (Yi and Zubiaga 2024). However, to the best of our knowledge, no studies have yet applied debias tuning—modifying loss functions to mitigate bias—in the context of hate speech classification with LLMs.

Experimental Setup

In this study, we investigate whether LLMs' tailored interactions introduce bias in hate speech classification. For this, we analyse how LLMs respond to hate speech contextualised to a given location and language. Specifically, we use a zero-shot approach, where the LLMs classify hate speech without prior fine-tuning on this task, and where no examples are provided to the model. To simulate LLMs' personalised information, we formulate our study by providing a user-persona with a location context (e.g. A person from United Kingdom), and the task of classifying hate speech.

This technique aligns with recent work using personas (*an entity whose viewpoints and behaviours the simulation seeks to examine and reproduce* (Zhang et al. 2024)) to examine bias and stereotypes in AI models (Gupta et al. 2024; Plaza-del Arco et al. 2024a,b). By using personas solely based on geographic identity—introducing only a country attribute without including other personal traits—we aim to uncover whether country contexts systematically prompt different classification outcomes, indicating potential bias in the models' behaviour.

Data We used MetaHate dataset in this work (Piot, Martín-Rodilla, and Parapar 2024). MetaHate comprises more than 1.2 million English labelled hate speech posts collected from 36 different datasets. Because it integrates multiple datasets, this meta-collection, is an ideal choice for validation and generalisation of our experiments. Its diversity helps ensure that modepos sls trained or evaluated on it are not overly dependent on any single dataset. Unfortunately, not all posts in MetaHate are labelled with the author's country, therefore, we could only use 24 132 instances with the country attribute. In this subset, there is a strong presence of posts from the United States, India, Australia, and the United Kingdom. To create a more balanced and comprehensive dataset, besides these countries, we will augment the dataset with underrepresented countries, ensuring a fairer representation in our study.

Country augmentation selection *Our World in Data*[1] is a research organization that provides comprehensive global data on issues such as poverty, health, education, and human rights. Their analyses include rankings of countries based on human rights, women's rights, and LGBTQ+ rights protection, with each ranking ranging from countries offering the strongest protections to those providing the least support (Herre and Arriagada 2016; Herre et al. 2023; Herre and Arriagada 2023). As mentioned, MetaHate subset lacks Non-Western countries, therefore, for our comparative analysis, we randomly selected twelve countries from either: a) the bottom 25 countries in the human rights index, b) the bottom 25 countries in LGBTQ+ legal equality index, and c) the women's rights index where laws require married women to obey their husbands. The countries chosen were Afghanistan, Belarus, Brunei, China, Cuba, Nicaragua, Nigeria, North Korea, Qatar, Russia, Saudi Arabia and Uganda. We tried to include underrepresented countries for all parts of the world to improve generalisability. Our goal is not to assess the model's alignment with the actual sociopolitical realities of these countries, but rather to examine whether LLMs reflect biases rooted in generalised Western perspectives and mainstream narratives, which may propagate reductive or prejudiced views of certain nations. With these countries included, we will now refer to our dataset as "CountryHate" to perform the experiments in this work. In the following section, we explain how we build the complete the dataset.

Models We selected five of the most advanced open-source multilingual LLMs, representing different parameter scales: `Llama-3.1-8B-Instruct` (for now onwards, `Llama 3.1`) (Dubey et al. 2024), `Mistral-Nemo-Instruct-2407` (`Nemo`) (Mistral AI team 2025), `gemma-3-27b` (`Gemma`) (Gemma Team 2025), `DeepSeek-R1-Distill-Llama-8B` (`DeepSeek`) (DeepSeek-AI 2025) and `Phi-4-mini-instruct` (`Phi 4`) (Microsoft et al. 2025). All the selected models are used in its 4-bit quantized version, provided by unsloth. `Llama 3.1` is a pre-trained model, outperforming competing models such as Mistral 7B or Gemma 7B in tasks such as commonsense understanding, mathematical reasoning tasks and general tasks. `Nemo` is an instruction-tuned model excelling in multilingual tasks and zero-shot scenarios, outperforming models like Mistal 7B in comprehension and domain-specific applications. `Gemma` is a 27-billion-parameter multimodal model supporting over 140 languages and a 128K-token context window. It achieves an MMLU-Pro score of 67.5 and a LiveCodeBench score of 29.7, indicating strong performance in reasoning and coding tasks. `DeepSeek` is a distilled version of DeepSeek's R1 model, a top-performing reasoning model, fine-tuned to improve logical inference and problem-solving abilities. `Phi 4` is a 3.8-billion-parameter language model trained on high-quality web and synthetic data, significantly outperforming recent open-source models of similar size and matching the performance of models twice its size on math and coding tasks requiring complex reasoning (Dubey et al. 2024; Mistral AI team 2025; Gemma Team 2025; DeepSeek-AI 2025; Microsoft et al. 2025). We decided to focus on open-source models as we do have no control over the responses of commercial LLMs and we do not know if there are moderation layers that are not from the model itself, but business layers (Spirling 2023; Ollion et al. 2024).

Country Language Prompt

To address ***RQ1***, ***RQ2*** and ***RQ3***, we used CountryHate data to prompt the models with located personas.

Persona The concept of "persona" or "user persona" started in marketing and design to better understand customer behaviours and motivations. Today, user-personas are widely used for role-playing and imitating human behaviour in areas like software development, UX design, and AI research. Generally, a persona refers to a fictional character created to represent different types of users that might interact with a system, or, as Zhang et al. (2024) defines it "an entity whose viewpoints and behaviour the simulation seeks to examine and reproduce". This persona can be categorised in demographic persona, character persona and individualised persona (Chen et al. 2024). We will focus on the first one to aggregate characteristics of demographic segments. In this regards, Gupta et al. (2024) propose three persona templates to guide the creation of these characters. We use one of their templates, adding geographic information, to prompt our models (see Table 1). We selected prompt B to examine how demographic personas affect classification outcomes compared to prompts without country information. Future work could explore the sensibility of LLMs when prompted with the different prompts proposed by literature.

[1] https://ourworldindata.org

Id.	Instruction
A	You are {*persona*}. Your responses should closely mirror the knowledge and abilities of this persona.
B	Adopt the identity of {*persona*}. Answer the questions while staying in strict accordance with the nature of this identity.
C	Take the role of {*persona*}. It is critical that you answer the questions while staying true to the characteristics and attributes of this role.

Table 1: Gupta et al. (2024) user personas. We used persona B in our experiments, adapting it to "Adopt the identity of someone from {*country*}".

Task prompt Our main objective is to classify hate speech, therefore, we used the following task prompt "*Determine whether a social media message is hateful or not. Respond with either a True for hate speech or False for neutral. Do not include nothing else than True or False.*". For example, for the hate speech message "*If a song is sung by a girl there is a 94% chance I'm going to hate it*", for the country Afghanistan, we will use the prompt: "*Adopt the identity of someone from Afghanistan. Answer the questions while staying in strict accordance with the nature of this identity. Determine whether a social media message is hateful or not. Respond with either a True for hate speech or False for neutral. Do not include nothing else than True or False.<Message>If a song is sung by a girl there is a 94% chance I'm going to hate it</Message>*".

Languages To investigate whether LLMs show bias when prompted in different languages, we conducted experiments using two language settings: (1) English, the original language of the messages in the CountryHate dataset, and (2) the official language of the persona's country (we will refer to this variant as **lang**). For the second setting, we translated the original posts into the official languages of the respective persona prompt before prompting the models. We opted to translate the data rather than use a different dataset to ensure a direct comparison between language settings—otherwise, the experiments would not be truly comparable, as they would be evaluating model behaviour on entirely different data rather than the same content presented in multiple languages.

Setup

CountryHate consists of 24 132 instances, which we split into two subsets: training (19 306 instances) and testing (4826 instances). We used the training subset for ***RQ4*** (Mitigate Bias), and the testing subset for studying if the LLMs exhibit a bias. We prompted each LLM thirteen times (12 country personas + 1 non-country prompt), in a zero-shot setting. This resulted in 62 738 generations—4826 without any context, and the remaining instances with country context[2]. This process was repeated for each language configuration (i.e. (1) English, (2) country persona language (**lang**)). In total, we generated 125 476 responses per model. To reduce randomness in the generation process, we set the temperature to 0, top-p to 0.1, top-k to 5, and limited the maximum token generation to 256. To summarise, our initial experiments for each model included the following variants: (1) **baseline**: no country context, (2) **country**: with country context, (3) **lang**: no country context, with posts translated into the country's language, and (4) **country lang**: both country context and posts translated into the country's language.

Output processing The prompt was crafted to guide the models to output only "True" or "False". However, some variations naturally occurred (e.g., *true*, *yes*, *vraiment* for True, and *false*, *no*, *faux* for False). We standardised these variations. Invalid responses (e.g., "*I cannot perform this action*") were removed.

Analysis Results

Next, we present the results for ***RQ1***, ***RQ2*** and ***RQ3***.

Country-specific prompts introduce bias in LLMs To answer ***RQ1*** we compare the contextualised generations with the baseline setting. We observed that the non-context variants consistently achieved higher F1 scores across all metrics for all models. Table 2 highlights this trend, showing that `Llama 3.1` **baseline** outperforms its **country** persona counterpart (F1-macro 0.7428 vs. 0.6269), and `Nemo`, `Gemma`, `DeepSeek` and `Phi 4` **baseline** similarly achieve superior results (F1-macro 0.8060 vs. 0.7512; 0.6081 vs. 0.5609; 0.7409 vs. 0.5542; and 0.6643 vs. 0.5964, respectively). These findings indicate that incorporating geographic context negatively impacts model behaviour.

	baseline		country	
Model	**F1**	**F1$_{\text{MACRO}}$**	**F1**	**F1$_{\text{MACRO}}$**
`Llama 3.1`	**0.7918**	**0.7428**	0.7401	0.6269
`Nemo`	**0.8397**	**0.8060**	0.7982	0.7512
`Gemma`	**0.6804**	**0.6081**	0.6587	0.5609
`DeepSeek`	**0.7964**	**0.7409**	0.5944	0.5542
`Phi 4`	**0.6903**	**0.6643**	0.6123	0.5964

Table 2: F1 scores of Llama 3.1, Nemo, Gemma, DeepSeek and Phi 4 base models, using the baseline (no context) and country settings.

Using personas' country languages reveals bias To answer ***RQ2***, if we compare the **baseline** variant from Table 2 and **lang** variant from Table 3, we see that for all models but `Gemma` the F1 scores are lower in the **lang** setting (e.g.

[2]Note that the distribution of posts across countries is uneven if the country is included in CountryHate.

for Nemo, the **baseline** F1 score is 0.8397, and for **lang** is 0.8261). Moreover, for ***RQ3*** we see that the use of personas' country languages further exacerbates the decline in model performance seen with geographic context alone. For example, Llama 3.1 **country lang** achieves an F1-macro score of 0.5993, a notable drop from the **country** variant's 0.6269 in Table 2. Similarly, Nemo **country lang** F1-macro score of 0.7182 is lower than the **country** variant's 0.7512. And the same behaviour is seen in the rest of the models between the **country** and **country lang** variants. This comparison underscores that incorporating both geographic context and the country's official language amplifies biases, resulting in a more pronounced reduction in overall model performance. Moreover, we noticed that for Llama 3.1, Gemma and DeepSeek, the **country lang** variant yielded a high number of invalid generations (>7000), suggesting that incorporating both geographic and language context may have introduced complexities or ambiguities that the model struggled to handle effectively.

	lang		country lang	
Model	**F1**	**F1$_{MACRO}$**	**F1**	**F1$_{MACRO}$**
Llama 3.1	**0.7880**	**0.7389**	0.7223	0.5993
Nemo	**0.8261**	**0.7910**	0.7634	0.7182
Gemma	**0.6945**	**0.6200**	0.5840	0.4945
DeepSeek	**0.7517**	**0.6926**	0.5821	0.5347
Phi 4	**0.6418**	**0.6207**	0.5341	0.5279

Table 3: F1 scores of Llama 3.1, Nemo Gemma, DeepSeek and Phi 4, on the testing subset.

Llama 3.1 is most affected by context The results in Table 4 reveal that Llama 3.1 exhibits the largest increase in False Negative Rates (FNR) when moving from the **baseline** variant (32.37%) to the **country** and **country-lang** variants (over 71%). In contrast, Nemo's FNR rises more moderately, from 16.95% to 30.66% and 26.61%, respectively, suggesting a lower sensitivity to geographic context. Gemma also shows a notable increase, with FNRs climbing from 54.11% to 81.43%. DeepSeek yielded an slightly lower FNR for the country variant compared to the baseline and lang ones, but the performance in terms of F1-score is much lower when prompted with country context. Phi 4, despite being a relatively weak performer in terms of F1-scores, maintains low FNRs across all settings, indicating minimal fluctuation due to context. These patterns highlight that while all models exhibit some level of bias, Llama 3.1 is the most affected by contextual changes.

Llama 3.1 exhibit country-specific bias As Llama 3.1 showed the higher bias within all models, we examined the differences across countries for Llama 3.1. Countries such as Brunei, Cuba, Russia, and Saudi Arabia exhibited a significantly higher FNR—approximately 140% higher than the baseline (Llama 3.1 **baseline**). This indicates that the model was more likely to miss classifying hateful messages as such in these countries. In contrast, when the country context was United Kingdom (UK) or United States (US), the false negative rate was notably lower—around 80% of the baseline FNR. Nigeria and Uganda followed with the next lowest rates, though their values remained higher than those of the UK and US (approx. 10% more). We performed a chi-squared test ($\mathcal{X}^2$) at $p > 0.01$ to determine whether there is a statistically significant difference between our baseline predictions and the contextualised predictions. As we can see in Table 5 all the countries reject the null hypothesis ($p < 0.01$): adding country context significantly affects the LLM's classification.

Model	baseline	country	lang	country lang
Llama 3.1	32.37%	71.42%	**32.22%**	71.20%
Nemo	**16.95%**	30.66%	17.48%	26.61%
Gemma	**54.11%**	75.13%	56.26%	81.43%
DeepSeek	42.36%	**31.79%**	41.13%	40.70%
Phi 4	7.67%	**6.06%**	7.60%	8.41%

Table 4: False Negative Rates (FNR) of Llama 3.1, Nemo, Gemma, DeepSeek and Phi 4, on the testing subset.

Country	**FN**	**FNR**	**p-value**
baseline	426	32.37%	–
Afghanistan	858	66.67%	$2.78e^{-65}$
Belarus	821	70.23%	$2.21e^{-63}$
Brunei	965	78.58%	$4.17e^{-114}$
China	815	64.33%	$1.98e^{-57}$
Cuba	891	79.77%	$2.16e^{-99}$
Nicaragua	936	76.47%	$3.88e^{-93}$
Nigeria	817	62.08%	$2.55e^{-60}$
North Korea	517	91.67%	$3.98e^{-67}$
Qatar	897	72.46%	$4.60e^{-90}$
Russia	958	78.91%	$3.21e^{-102}$
Saudi Arabia	845	77.52%	$6.63e^{-100}$
Uganda	815	62.02%	$6.68e^{-56}$
Australia	78	69.64%	$1.51e^{-6}$
UK	27	54.00%	$6.04e^{-82}$
US	554	59.76%	$1.53e^{-4}$

Table 5: False Negatives (FN), False Negative Rate per Hate Cases (FNR), and p-values for the selected countries and baseline, for Llama 3.1.

As Nemo was our best performing model, we decided to examine the country-specific bias. For this model, a different pattern emerges. Our target countries do not show higher false negative rates, but Australia stands out with elevated rates, although these are lower than those of Llama 3.1 (46.67% vs. 69.00%). This higher false negative rate in Aus-

tralia matches the results from Llama 3.1 and aligns with research indicating that there are strong biases against immigrants and refugees in Australia (Schweitzer et al. 2005). Because of this, when using a persona prompt with Australia, the model might not recognise hate speech against refugees or immigrants, as it has learned these prejudices about the Australian population. For the other countries, the false negative rates stay about the same.

In summary, prompting models with persona prompts leads to a performance drop, which varies depending on the specific country, not only in the incorporation of the prompt. While this work focuses on geographic and language bias, similar patterns have been observed in how LLMs represent emotions related to religion (Plaza-del Arco et al. 2024b). Future research could further explore other bias in LLMs, particularly in the context of hate speech identification.

Mitigate Bias

Based on our findings in ***RQ1***, ***RQ2*** and ***RQ3***, which revealed demographic bias in the selected LLMs, we propose *debias tuning* as a strategy to reduce this bias. The key idea of debias tuning is to align hate speech classification outcomes with contextual information and without it, while reducing the impact of biases in the context (Dong et al. 2024). Prior research has shown the effectiveness of prompting techniques for mitigating biases such as reducing gender bias, limiting the generation of hate speech or even preventing a model from performing certain tasks (Dwivedi, Ghosh, and Dwivedi 2023; Piot and Parapar 2025). In this work, we present a debias tuning method to reduce inconsistencies between country-context and non-context predictions, ensuring more consistent responses across locations.

Setup We fine-tuned Llama 3.1, Nemo and Phi 4 models using the training subset. This model selection allowed us to assess the generalizability of the approach while keeping the computational cost manageable. For our fine-tuning, we provide the models with the persona prompt, the social media text and the gold label. In order to study the generability of the models, on both country and language dimensions, we decided to perform the finetuning with only four countries from our candidates list. These are Afghanistan (Persian), Brunei (Malay), Qatar (Arabic) and Saudi Arabia (Arabic). We fine-tuned the models in two language settings: (1) English (debias tuning), and (2) the official languages of the personas' countries (multilingual debias tuning).

Custom loss To introduce our proposed debias, we have implemented a custom loss. Our loss applies a consistency penalty if either the prediction is invalid (i.e., not in True, False), or if the non-context prediction matches the gold label, but the context prediction does not. Let x be the input text, x_c the input text with country context text and y the gold label. We use the cross-entropy function as our loss function, where $\mathcal{L}_{\text{class}} = CrossEntropy(logits(x), y))$ is the classification loss for the non-country input, and $\mathcal{L}^c_{\text{class}} = CrossEntropy(logits(x_c), y))$ is the loss for the country-context text. The average classification loss is computed as:

$$\mathcal{L}_{\text{avg\_class}} = \frac{\mathcal{L}_{\text{class}} + \mathcal{L}^c_{\text{class}}}{2} \quad (1)$$

Then, to compute the loss with the consistency penalty, we include Equation 1 as:

$$\mathcal{L}_{\text{loss}} = \mathcal{L}_{\text{avg\_class}} + \alpha \cdot \mathcal{L}_{\text{avg\_class}} \quad (2)$$

Note that in Equation 2, α is > 0 if $\hat{y} = y$ and $\hat{y}_c \neq y$ or if $\hat{y} = None$ or $\hat{y}_c = None$. Otherwise, this term is 0.

Debias fine-tuning We fine-tuned Meta-Llama-3.1-8B-Instruct-bnb-4bit, Mistral-Nemo-Instruct-2407-bnb-4bit and Phi-4-mini-instruct-unsloth-bnb-4bit from unsloth using 4-bit precision for memory efficiency. Models are initialised with a sequence length of 256 tokens and fine-tuned with LoRA for one epoch, using an effective batch size of 16 (batch size 4, gradient accumulation 4). We employ the AdamW optimizer in 8-bit precision with a learning rate of $2e^{-6}$, weight decay of 0.01, and gradient clipping at 0.3. Training includes a 3% warm-up and linear decay for the learning rate.

With our models fine-tuned (debias-llama, debias-nemo, debias-phi, debias-llama-lang, debias-nemo-lang, and debias-phi-lang), we proceed to evaluate them. We used the testing subset and the same persona prompt, the twelve selected countries and the original author's country. Next we present the results.

Debias Tuning Results

Debias tuning outperforms base models Comparing the F1 scores from the debias models to the base models reveals that for the **baseline** all debias-llama, debias-nemo and debias-phi achieve higher F1 scores (see Table 6, row baseline). With **country** context, debias-llama notably increases the F1 scores, specially the F1-macro (from 0.6269 to 0.7169). debias-nemo and debias-phi models achieve similar but slightly lower F1 scores improvements, compared to its base version. This outcome may be due to the fact that Nemo and Phi 4 did not exhibit as much bias in its base version, indicating that its initial robustness may limit the extent of improvement possible through debiasing, but still getting improved results.

Mixed outcomes in multilingual debiasing The results in Table 6, rows baseline and country, show that the debias-llama-lang model consistently outperforms its counterparts across both the **baseline** and **country** variant. In the **baseline** setting, debias-llama-lang achieves an F1-micro score of 0.8142, surpassing Llama 3.1 (0.7918) and the English-tuned-only debias-llama (0.8001). Notably, in the **country** variant, debias-llama-lang records an F1-micro of 0.7937, getting closer to the **baseline** (0.8142). This is a significant improvement compared to the original Llama 3.1, where the **country** variant (0.7401) was 0.0538 points lower than the **baseline** version (0.7918). This reduction in the performance gap highlights that the classification bias introduced by geographic context

Variant	Metric	Llama 3.1	debias llama	debias llama-lang	Nemo	debias nemo	debias nemo-lang	Phi 4	debias phi	debias phi-lang
baseline	F1	0.7918	0.8001	**0.8142**	0.8397	**0.8518**	0.8413	0.6903	**0.6905**	**0.6905**
	$F1_{MACRO}$	0.7428	0.7655	**0.7765**	0.8060	**0.8155**	0.7906	0.6643	**0.6645**	**0.6645**
country	F1	0.7401	0.7792	**0.7937**	0.7982	**0.8009**	0.7929	0.6123	**0.6154**	0.6122
	$F1_{MACRO}$	0.6269	0.7169	**0.7323**	0.7512	**0.7523**	0.7264	0.5964	**0.5988**	0.5962
lang	F1	0.7880	0.8098	**0.8184**	0.8261	**0.8431**	0.8416	0.6418	0.6441	**0.6493**
	$F1_{MACRO}$	0.7389	0.7661	**0.7814**	0.7910	**0.8057**	0.7907	0.6207	0.6229	**0.6269**
country lang	F1	0.7223	0.7679	**0.7972**	0.7634	0.7831	**0.7894**	**0.5341**	0.5298	0.5316
	$F1_{MACRO}$	0.5993	0.7182	**0.7411**	0.7182	**0.7344**	0.7242	**0.5279**	0.5244	0.5260

Table 6: F1 scores of base models and debias models, with the non-context and country variants, on the testing subset.

has been effectively mitigated in the Llama debiased multilingual model, allowing for consistent performance across contexts. However, the results show that the debias-nemo-lang model did not perform as well as the English-only debias-nemo version or the base model. For the **baseline**, debias-nemo achieved higher F1 (0.8518 vs. 0.8413) and F1-macro (0.8155 vs. 0.7906) scores. Additionally, in the country setting, debias-nemo outperformed debias-nemo-lang in both metrics (F1: 0.8009 vs. 0.7929, F1-macro: 0.7523 vs. 0.7264), but both debias methods performed better than their base version. For Phi 4, we see a similar behaviour as for Nemo, with the debias-phi method yielding the best results for the **baseline** and **country** variants. These findings suggest that multilingual debiasing is highly effective in reducing bias for models like Llama, which exhibited significant bias with location context. However, in the Nemo and Phi variants, where the initial bias was less pronounced, the improvements were not as substantial.

Debias tuning reduces language bias Table 6, rows lang and country lang, demonstrates that debias tuning effectively reduces language bias, improving F1 scores across all metrics and settings for both Llama 3.1 and Nemo. For Llama 3.1, the debiased variants show marked improvements, with the **country lang** F1-macro score increasing from 0.5993 to 0.7182 and the **lang** score rising from 0.7389 to 0.7661. Nemo also benefits from debiasing, albeit to a lesser extent. The debias-nemo **lang** F1-macro improving from 0.7910 to 0.8057 and the **country lang** variant seeing a smaller increase from 0.7182 to 0.7344. For Phi, multilingual learning was less effective compared to the other models, with improvements observed only in the **lang** setting. These results indicate that debias tuning improves the overall performance, with particularly strong gains for Llama, in multilingual contextual settings.

Multilingual debias tuning excels in diverse language inference Table 6, rows lang and country lang, highlight the strong performance of the debias-llama-lang model, which consistently outperforms other Llama variants. In the **lang** setting, debias-llama-lang achieves an F1 score of 0.8184, surpassing the original Llama 3.1 (0.7880) and the English-only debias-llama (0.8098). Similarly, in the **country lang** variant, it records an F1 score of 0.7972, significantly outperforming its counterparts. These results brings closer F1 scores for the non-context and context variants, showing that our proposed models mitigated biases introduced by contextual, personalised prompts. For the debias-nemo-lang model, the **country lang** variant shows improved F1 scores, increasing from 0.7634 to 0.7894. However, in the **lang** setting, the F1 and F1-macro scores are slightly higher for debias-nemo (F1 0.8431, F1-macro 0.8057) compared to debias-nemo-lang (F1 0.8416, F1-macro 0.7907). debias-phi-lang achieves the highest F1 score (0.6493) in the **lang** setting, while in the **country lang** setting, the base model performs best. However, the F1 scores across all three variants are similarly low and close to each other. These findings confirm that multilingual debiasing not only reduces bias effectively but also improves model performance in multilingual contexts, making it the most reliable approach among the tested variants. These results also underscore the importance of choosing a strong LLM, as Phi shows limited performance and smaller gains. Moreover, although debias-llama-lang yields the best improvement, debias-nemo-lang still performs better in its **baseline** setting. However, in the **country lang** variant, originally, Nemo outperformed Llama 3.1, but now the debias models for Llama outperform Nemo's, showing the gains in multilingualism for this model.

Debias tuning reduces country-specific bias We analysed the evolution of FNRs across the base and debias models. As we can see in Figure 2, upper row, debias tuning significantly reduces country-specific bias in Llama 3.1 by lowering FNRs compared to baseline configurations. In the Llama 3.1 model, countries like Afghanistan (66.67%), Brunei (78.58%), North Korea (91.67%) and Saudi Arabia (77.52%) had high FNRs, while the United States (59.76%) and the United Kingdom (54.00%) performed moderately better. debias-llama reduced FNRs in high-bias regions, such as Afghanistan (33.10%), Brunei (31.94%) and North Korea (35.65%), achieving more equitable detection.

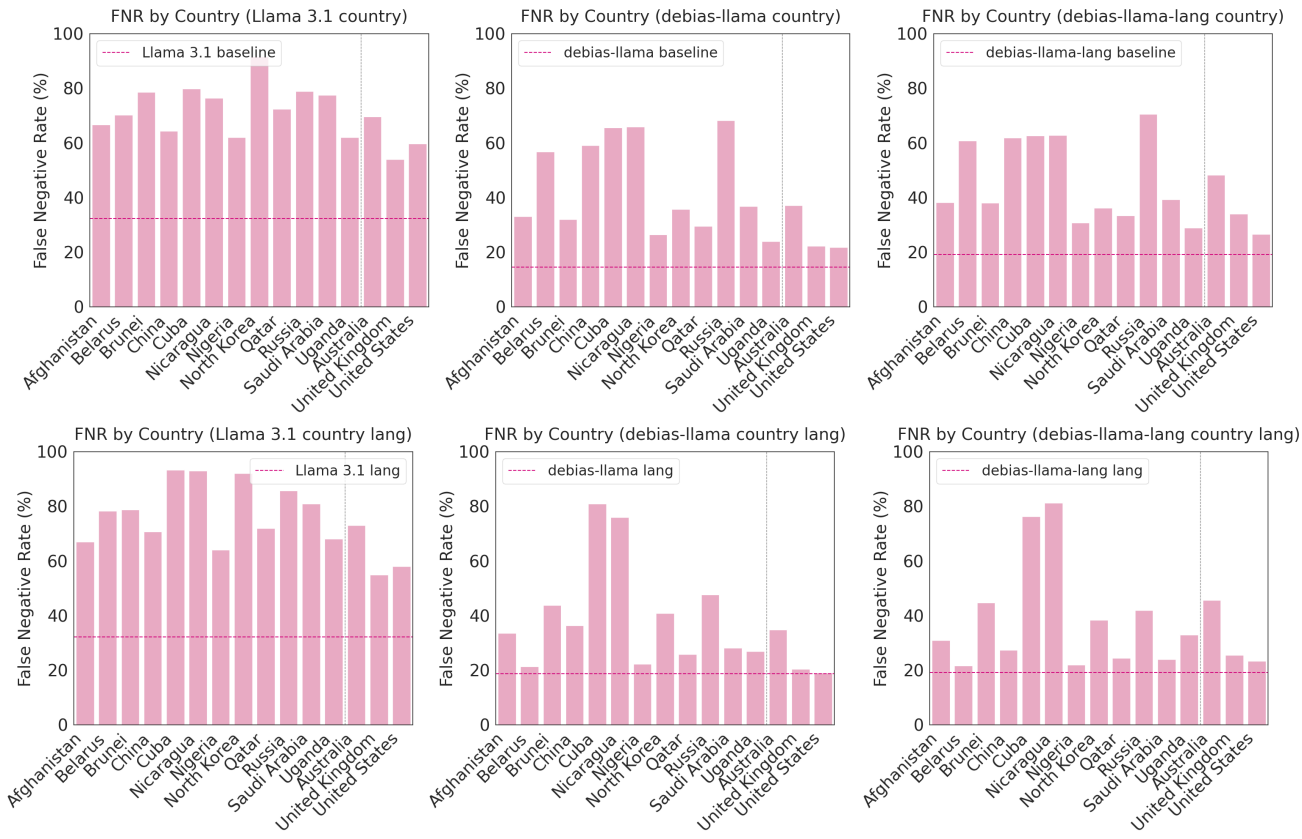

Figure 2: (Up): FNR across countries for (1) Llama 3.1, (2) debias-llama and (3) Debias Llama Lang, for English. (Down): FNR across countries for (4) Llama 3.1, (5) debias-llama and (6) Debias Llama Lang, for persona's country languages (lang).

Other countries such as Cuba, Nicaragua or Russia, show improved results (new FNRs 65.58%, 65.94% and 68.30%, old FNRs 93.23%, 92.95% and 85.77% respectively) but to a smaller extent. The `debias-llama-lang` model maintained the results but showed slightly higher FNRs in some regions, like Afghanistan (38.20%), Saudi Arabia (39.31%) and Russia (70.59%). Table 7 further illustrates the effectiveness of debias tuning. We performed a chi-square test ($\mathcal{X}^2$) at $p > 0.01$ to evaluate the significance of false negatives across the countries compared to the United Kingdom (the country with the lowest FNR). The p-values indicate that while the `Llama 3.1` model shows significant differences across all our target countries, the `debias-llama` model eliminates significant differences in many of the countries, and raised the p-values in the rest of them, reflecting reduced disparity. This reduction in significance indicates that the debiasing models successfully mitigate country-specific bias, creating a more equitable performance across the targeted countries. Nonetheless, to improve generalisation and robustness, we recommend extending the debiasing process to include a broader set of countries.

Debias tuning further mitigates bias in multiple languages Debias using personas' official languages showed similar, but improved trends, as seen in Figure 2, bottom row. For the `Llama 3.1` model, Afghanistan (66.97%), Brunei (78.95%), and Saudi Arabia (80.96%) had high FNRs, while the United Kingdom (54.90%) and the United States (55.10%) fared better. `debias-llama` reduced FNRs significantly, with Afghanistan at 33.62%, Saudi Arabia at 28.11%, and the United States at 18.90%, almost matching the baseline (18.72%). The two countries that showed limited improvement were Cuba and Nicaragua, where FNRs decreased, but to a smaller extent (from 79.77% and 76.4%, to 62.70% and 62.76%, respectively). The `debias-llama-lang` model further improved performance in regions like Afghanistan (30.97%) and Saudi Arabia (24.01%), though Australia and Brunei remained high at 49.04% and 44.65%, respectively. We observe the same pattern with Cuba and Nicaragua as with `debias-llama`, highlighting the model's difficulty in handling Spanish text when it has not been debiased for this language. Still, these results highlight again debias tuning's effectiveness in reducing geographic bias and improving fairness, even in multilingual contexts, where the models could generalize to unseen languages, although results for some languages, like Spanish, could still be improved.

Effectiveness of debias tuning on unincluded countries In Figure 2, we see that the FNRs are reduced for countries not included in the debias tuning. For some countries, such as Nigeria, North Korea, and Uganda, the `Llama`

	Llama 3.1		debias-llama	
Country	**P-value**	**Sig.**	**P-value**	**Sig.**
Afghanistan	$5.38e^{-5}$	Yes	0.0084	No
Belarus	$2.21e^{-64}$	Yes	$1.08e^{-6}$	Yes
Brunei	$2.26e^{-7}$	Yes	0.0096	No
China	$1.98e^{-57}$	Yes	$1.56e^{-6}$	Yes
Cuba	$2.16e^{-99}$	Yes	$2.48e^{-6}$	Yes
Nicaragua	$3.88e^{-93}$	Yes	$1.89e^{-6}$	Yes
Nigeria	0.0003	Yes	0.0589	No
North Korea	$3.98e^{-67}$	Yes	$1.24e^{-4}$	Yes
Qatar	$1.07e^{-5}$	Yes	0.0401	No
Russia	$3.21e^{-102}$	Yes	$1.40e^{-9}$	Yes
Saudi Arabia	$2.79e^{-6}$	Yes	0.0091	No
Uganda	0.0002	Yes	0.0392	No

Table 7: Significance (Sig.) results for pairwise comparisons between target countries and United Kingdom.

3.1 model resulted in an FNR of around 62%, while the debias-llama model lowered this to approx. 25%. But, for other countries, such as Cuba or Russia, the reduction was not as notable. This highlights the effectiveness of our method, as disparities among countries are less pronounced despite these countries not being specifically included in the debiasing process. Still, challenges remain to achieve a model were there are no big differences among countries. This behaviour is consistent across both the debias models and the debias language models.

Error Analysis

While our method effectively reduces biases, the F1-scores remain at around 0.79 for Llama 3.1 and Nemo when prompted with country context. For Phi 4, a lighter, less powerfull model, the F1-scores remain at 0.60. To identify remaining classification issues, we conducted an error analysis on a sample of 100 instances per error type for Llama 3.1 and Nemo models, as these models showed the most promising performance for potential deployment. This analysis focused on common misclassifications, examining both false negatives and false positives for the debias models on English texts. We used error categories from van Aken et al. (2018) to guide our evaluation.

Error Classes of False Negatives

Hate speech without swear words Identifying implicit hate and hate speech without swear words is a well-known challenge (Davidson et al. 2017; Piot and Parapar 2025). In our manual evaluation, 14% of the posts fall into this category, with the majority targeting women. This highlights the limitations of existing models, which often rely on overtly offensive language or specific keywords to detect hate speech. Developing models capable of accurately identifying implicit hate speech requires incorporating a deeper understanding of context, intent, and subtle biases present in language. Such improvements are crucial to effectively addressing these hidden forms of harm.

The women skaters can't fall and make it look graceful like the men

Rhetorical questions A common strategy in hate speech is the use of rhetorical questions (Parvaresh and Harvey 2023), often indicated by multiple question marks or exclamation points. Although only 2% of our manual sample exhibited this pattern, it poses a challenge for state-of-the-art models to detect and requires further investigation.

Are there any good female comedians?? Or comediennes if you prefer?

Metaphors and comparisons Another type of implicit hate speech involves the use of metaphors and comparisons. Zhang and Luo (2019) noted that understanding these texts requires complex reasoning as well as cultural and social knowledge. In our subsample, 2% of the posts followed this pattern. While this is not the most common type of error, we emphasise the importance of developing models that can effectively identify implicit hate speech in all its forms.

Releasing private Sony e mails to hurt people is the same as releasing nude photos of Jennifer Lawrence. Why are they ok t...

Sarcasm and irony Detecting sarcasm and irony is a well-known challenge in NLP, especially in hate speech, where they can mask hateful intent. In our manual sample, sarcasm and irony appear in 3% of the posts. While not common, this poses a significant challenge as these strategies often convey the opposite of their literal meaning.

I guess I would be a safer driver too if I did ten under the speed limit

Doubtful labels In this class, we include data points where we question whether the original label was correct, based on our definition of hate speech. For example, we noticed that phrases like "I hate" are often labelled as hate speech in the original dataset, even when the target did not fit the definition. In fact, 74% of the posts we analysed fall into this category. This raises concerns about potentially incorrect labels, underscoring the challenges of annotating hate speech.

I hate racist people very much.

Error Classes of False Positives

Regular use of swear words When analysing false positives, we found that half of our sample falls into this category, which includes various non-hateful uses of language. This encompasses affectionate swear words, such as saying "*You are a badass!*", as well as words like *hate*, *abuse*, or *harassed* when used in non-hate contexts, as well as negative words such as *dead* or *gun* that appear in neutral or unrelated contexts. Additionally, instances of negative or critical statements, such as "*This is his worst performance*", but misclassified as hate, are included. These examples highlight the importance of understanding context to differentiate hate from non-hate language.

WTF score did you expect serving liver??

Quotations or references A major challenge for hate speech detection models is distinguishing between actual hate speech and quotes or references to hate speech. These often include quotation marks or are explicitly contextualised, such as "*he said that...*". In our analysis, we found that 14% of the posts fall into this category. This highlights the need for models to better understand context and intent when identifying hate speech.

Yes, and "girls suck at basketball" is such an unconventional sentiment.

Idiosyncratic or rare words In our manual evaluation, we found that 11% of the posts fell into this category. These posts included rare words, misspellings, abbreviations, slang, words in other languages, or highly descriptive hashtags. This issue emphasises the need for hate speech detection models to improve their handling of less common language elements, as failing to do so can result in overlooking harmful content or mislabelling non-hateful expressions.

What does small town India think about #MeToo? Kaam ke liye sweetoo, kaam ke baad #MeToo !

Doubtful labels We found fewer doubtful labels compared to false negatives. In our sample, 8% of the posts were categorised as doubtful labels, with most of them falling under the category of implicit hate. This suggests that implicit hate remains a challenge for accurate classification, even among false positives.

call me sexist but I hate audiobooks read by women

In summary, most false negatives fall under doubtful labels, underscoring the challenges of annotating hate speech. Conversely, false positives often stem from misclassifying regular swear words, highlighting the need for methods to differentiate between insults and hate speech. Additionally, 5% of false negatives and 17% of false positives are misclassifications that do not fit into established categories.

Conclusions

This study shows that LLMs' memory features can introduce personalisation based on demographic attributes, which affects sensitive topics like hate speech. By using location-specific personas to simulate this feature, we observed systematic differences in classification outcomes depending on geographical context. We also examined whether the models displayed bias based on the language used in prompts, finding that this also influenced their behaviour. To address these variations, we applied a fine-tuning strategy that incorporates a consistency-based penalty in the custom loss function. This method aligns predictions with and without country context, leading to improved F1 scores and reducing bias in country-specific settings, thereby enhancing model performance across different contexts. Additionally, we conducted a detailed error analysis of the misclassifications made by the debias models. Future research may investigate further debiasing techniques and expand this approach to other types of social bias that could arise from the memory personalisation features of LLMs, ultimately contributing to more inclusive and robust AI systems.

Acknowledgments

The authors thank the funding from the Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351. The authors thank the financial support supplied by the grant PID2022-137061OB-C21 funded by MICIU/AEI/10.13039/501100011033 and by "ERDF/EU". The authors also thank the funding supplied by the Consellería de Cultura, Educación, Formación Profesional e Universidades (accreditations ED431G 2023/01 and ED431C 2025/49) and the European Regional Development Fund, which acknowledges the CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01).

References

Badjatiya, P.; Gupta, M.; and Varma, V. 2019. Stereotypical Bias Removal for Hate Speech Detection Task using Knowledge-based Generalizations. In *The World Wide Web Conference*, WWW '19, 49–59. New York, NY, USA: Association for Computing Machinery. ISBN 9781450366748.

Chen, J.; Wang, X.; Xu, R.; Yuan, S.; Zhang, Y.; Shi, W.; ...; and Xiao, Y. 2024. From Persona to Personalization: A Survey on Role-Playing Language Agents. arXiv:2404.18231.

Dammu, P. P. S.; Jung, H.; Singh, A.; Choudhury, M.; and Mitra, T. 2024. "They are uncultured": Unveiling Covert Harms and Social Threats in LLM Generated Conversations. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 20339–20369. Miami, Florida, USA: Association for Computational Linguistics.

Das, M.; Raj, R.; Saha, P.; Mathew, B.; Gupta, M.; and Mukherjee, A. 2023. HateMM: A Multi-Modal Dataset for Hate Video Classification. *Proceedings of the ICWSM 2023*, 17: 1014–1023.

Davidson, T.; Bhattacharya, D.; and Weber, I. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, 25–35. Florence, Italy: ACL.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the ICWSM 2017*, 11(1): 512–515.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Demidova, A.; Atwany, H.; Rabih, N.; Sha'ban, S.; and Abdul-Mageed, M. 2024. John vs. Ahmed: Debate-Induced Bias in Multilingual LLMs. In *Proceedings of The Second Arabic Natural Language Processing Conference*, 193–209. Bangkok, Thailand: ACL.

Dong, X.; Wang, Y.; Yu, P. S.; and Caverlee, J. 2024. Disclosure and Mitigation of Gender Bias in LLMs. arXiv:2402.11190.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; ...; and Zhao, Z. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Dwivedi, S.; Ghosh, S.; and Dwivedi, S. 2023. Breaking the Bias: Gender Fairness in LLMs Using Prompt Engineering and In-Context Learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 15(4).

ElSherief, M.; Kulkarni, V.; Nguyen, D.; Yang Wang, W.; and Belding, E. 2018a. Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media. *Proceedings of the ICWSM 2018*, 12(1).

ElSherief, M.; Nilizadeh, S.; Nguyen, D.; Vigna, G.; and Belding, E. 2018b. Peer to Peer Hate: Hate Speech Instigators and Their Targets. *Proceedings of the ICWSM 2018*, 12(1).

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Founta, A.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; ...; and Kourtellis, N. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *Proceedings of the ICWSM 2018*, 12(1).

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1): 2096–2030.

Garg, S.; Perot, V.; Limtiaco, N.; Taly, A.; Chi, E. H.; and Beutel, A. 2019. Counterfactual Fairness in Text Classification through Robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, 219–226. New York, NY, USA: Association for Computing Machinery. ISBN 9781450363242.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gemma Team. 2025. Gemma 3.

Gira, M.; Zhang, R.; and Lee, K. 2022. Debiasing Pre-Trained Language Models via Efficient Fine-Tuning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 59–69. Dublin, Ireland: ACL.

Gupta, S.; Shrivastava, V.; Deshpande, A.; Kalyan, A.; Clark, P.; Sabharwal, A.; and Khot, T. 2024. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

Herre, B.; and Arriagada, P. 2016. Human Rights. *Our World in Data*. Https://ourworldindata.org/human-rights.

Herre, B.; and Arriagada, P. 2023. LGBT+ Rights. *Our World in Data*. Https://ourworldindata.org/lgbt-rights.

Herre, B.; Samborska, V.; Arriagada, P.; and Ritchie, H. 2023. Women's Rights. *Our World in Data*. Https://ourworldindata.org/women-rights.

Jiao, J.; Afroogh, S.; Xu, Y.; and Phillips, C. 2025. Navigating LLM ethics: advancements, challenges, and future directions. *AI and Ethics*, 5(6): 5795–5819.

Kamruzzaman, M.; and Kim, G. L. 2025. Exploring Changes in Nation Perception with Nationality-Assigned Personas in LLMs. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 3660–3678. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Kotek, H.; Dockum, R.; and Sun, D. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, 12–24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701139.

Kumar, A.; Murthy, S. V.; Singh, S.; and Ragupathy, S. 2024. The Ethics of Interaction: Mitigating Security Threats in LLMs. arXiv:2401.12273.

Leidinger, A.; and Rogers, R. 2024. How Are LLMs Mitigating Stereotyping Harms? Learning from Search Engine Studies. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 839–854.

Lim, S.; and Pérez-Ortiz, M. 2024. The African Woman is Rhythmic and Soulful: An Investigation of Implicit Biases in LLM Open-ended Text Generation. arXiv:2407.01270.

Lin, Z.; Guan, S.; Zhang, W.; Zhang, H.; Li, Y.; and Zhang, H. 2024. Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, 57(9).

Ma, X.; Sap, M.; Rashkin, H.; and Choi, Y. 2020. PowerTransformer: Unsupervised Controllable Revision for Biased Language Correction. In *Proceedings of the 2020 Conference EMNLP*, 7426–7441. Online: ACL.

Manvi, R.; Khanna, S.; Burke, M.; Lobell, D.; and Ermon, S. 2024. Large Language Models are Geographically Biased. arXiv:2402.02680.

Maronikolakis, A.; Baader, P.; and Schütze, H. 2022. Analyzing Hate Speech Data along Racial, Gender and Intersectional Axes. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 1–7. Seattle, Washington: ACL.

Microsoft; :; Abouelenin, A.; Ashfaq, A.; Atkinson, A.; Awadalla, H.; Bach, N.; Bao, J.; ...; and Zhou, X. 2025. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. arXiv:2503.01743.

Mistral AI team. 2025. Mistral NeMo — Mistral AI. https://mistral.ai/news/mistral-nemo. Accessed: 09/01/2025.

Mozafari, M.; Farahbakhsh, R.; and Crespi, N. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PLOS ONE*, 15(8): e0237861.

Nations, U. 2023. What is hate speech? Accessed: 15/11/2023.

Ollion, E.; Shen, R.; Macanovic, A.; and Chatelain, A. 2024. The dangers of using proprietary LLMs for research. *Nature Machine Intelligence*, 6(1): 4–5.

OpenAI. 2024. Memory and new controls for ChatGPT. https://openai.com/index/memory-and-new-controls-for-chatgpt/. [Accessed 04-12-2024].

Palta, S.; and Rudinger, R. 2023. FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In *Findings of the ACL: ACL 2023*, 9952–9962. Toronto, Canada: ACL.

Parvaresh, V.; and Harvey, G. 2023. *Rhetorical Questions as Conveyors of Hate Speech*, 229–251. Cham: Springer Nature Switzerland. ISBN 978-3-031-38248-2.

Piot, P.; Martín-Rodilla, P.; and Parapar, J. 2024. MetaHate: A Dataset for Unifying Efforts on Hate Speech Detection. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1): 2025–2039.

Piot, P.; and Parapar, J. 2025. Decoding Hate: Exploring Language Models' Reactions to Hate Speech. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *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)*, 973–990. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Plaza-del Arco, F. M.; Cercas Curry, A.; Curry, A.; Abercrombie, G.; and Hovy, D. 2024a. Angry Men, Sad Women: Large Language Models Reflect Gendered Stereotypes in Emotion Attribution. In *Proceedings of the 62nd Annual Meeting of the ACL*, 7682–7696. Bangkok, Thailand: ACL.

Plaza-del Arco, F. M.; Curry, A. C.; Paoli, S.; Cercas Curry, A.; and Hovy, D. 2024b. Divine LLaMAs: Bias, Stereotypes, Stigmatization, and Emotion Representation of Religion in Large Language Models. In *Findings of the ACL: EMNLP 2024*, 4346–4366. Miami, Florida, USA: ACL.

Qian, Y.; Muaz, U.; Zhang, B.; and Hyun, J. W. 2019. Reducing Gender Bias in Word-Level Language Models with a Gender-Equalizing Loss Function. In *Proceedings of the 57th Annual Meeting of the ACL: Student Research Workshop*, 223–228. Florence, Italy: ACL.

Salinas, A.; Penafiel, L.; McCormack, R.; and Morstatter, F. 2024. "I'm not Racist but. . . ": Discovering Bias in the Internal Knowledge of Large Language Models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Schick, T.; Udupa, S.; and Schütze, H. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the ACL*, 9: 1408–1424.

Schmidt, B. 2015. Rejecting the gender binary: a vector-space operation. *Ben's Bookworm Blog*.

Schweitzer, R.; Perkoulidis, S.; Krome, S.; Ludlow, C.; and Ryan, M. 2005. Attitudes towards refugees: The dark side of prejudice in Australia. *Australian Journal of Psychology*, 57(3): 170–179.

Shrawgi, H.; Rath, P.; Singhal, T.; and Dandapat, S. 2024. Uncovering Stereotypes in Large Language Models: A Task Complexity-based Approach. In *Proceedings of the 18th Conference of the European Chapter of the ACL*, 1841–1857. St. Julian's, Malta: ACL.

Spirling, A. 2023. Why open-source generative AI models are an ethical way forward for science. *Nature*, 616(7957): 413–413.

Thakur, V. 2023. Unveiling Gender Bias in Terms of Profession Across LLMs: Analyzing and Addressing Sociological Implications. arXiv:2307.09162.

van Aken, B.; Risch, J.; Krestel, R.; and Löser, A. 2018. Challenges for Toxic Comment Classification: An In-Depth Error Analysis. In *Proceedings of the 2nd Workshop on Abusive Language Online*, 33–42. Brussels, Belgium: ACL.

Wan, Y.; and Chang, K.-W. 2025. White Men Lead, Black Women Help? Benchmarking and Mitigating Language Agency Social Biases in LLMs. In *Proceedings of the 63rd Annual Meeting of the ACL*, 9082–9108. Vienna, Austria: ACL. ISBN 979-8-89176-251-0.

Xia, M.; Field, A.; and Tsvetkov, Y. 2020. Demoting Racial Bias in Hate Speech Detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, 7–14. Online: ACL.

Xie, Z.; and Lukasiewicz, T. 2023. An Empirical Analysis of Parameter-Efficient Methods for Debiasing Pre-Trained Language Models. In *Proceedings of the 61st Annual Meeting of the ACL*, 15730–15745. Toronto, Canada: ACL.

Yi, P.; and Zubiaga, A. 2024. ID-XCB: Data-independent Debiasing for Fair and Accurate Transformer-based Cyberbullying Detection. arXiv:2402.16458.

Zhang, Z.; and Luo, L. 2019. Hate speech detection: A solved problem? The challenging case of long tail on Twitter. *Semant. Web*, 10(5): 925–945.

Zhang, Z.; Rossi, R. A.; Kveton, B.; Shao, Y.; Yang, D.; Zamani, H.; ...; and Wang, Y. 2024. Personalization of Large Language Models: A Survey. arXiv:2411.00027.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference EMNLP*, 2979–2989. Copenhagen, Denmark: ACL.

Zhao, J.; Zhou, Y.; Li, Z.; Wang, W.; and Chang, K.-W. 2018. Learning Gender-Neutral Word Embeddings. In *Proceedings of the 2018 Conference EMNLP*, 4847–4853. Brussels, Belgium: ACL.

Zhou, F.; Mao, Y.; Yu, L.; Yang, Y.; and Zhong, T. 2023. Causal-Debias: Unifying Debiasing in Pretrained Language Models and Fine-tuning via Causal Invariant Learning. In *Proceedings of the 61st Annual Meeting of the ACL*, 4227–4241. Toronto, Canada: ACL.

Ethics Checklist

1. For most authors...

(a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, in section *Experimental Setup*

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, in section *Experimental Setup*

(e) Did you describe the limitations of your work? Yes, in section *Limitations*

(f) Did you discuss any potential negative societal impacts of your work? Yes, in section *Limitations*

(g) Did you discuss any potential misuse of your work? Yes, in section *Limitations*

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, in section *Limitations*

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we provide an anonymous link with the code and models and, upon acceptance, we will provide the real link

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, in sections *Experimental Setup* and section *Mitigate Bias*

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, in section *Computational Resources*

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, in sections *Analysis Results*, *Debias Tuning Results* and *Error Analysis*

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, in sections *Analysis Results*, *Debias Tuning Results* and *Error Analysis*

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes

(b) Did you mention the license of the assets? No, not directly in the paper, but it is mentioned in our repository

(c) Did you include any new assets in the supplemental material or as a URL? NA

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, due to space constraints, but all the assets we use have the proper consent

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, we made explicit at the beginning of the paper that the manuscript might content offensive content, because we are using hate speech datasets, containing offensive content

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA