Hate Personified

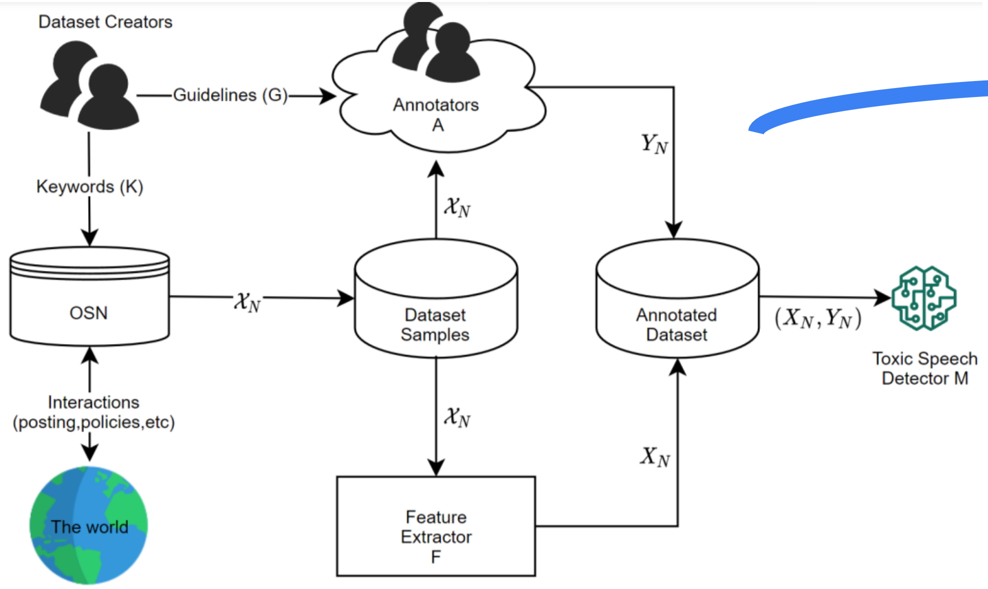
Investigating the role of LLMs in content moderation

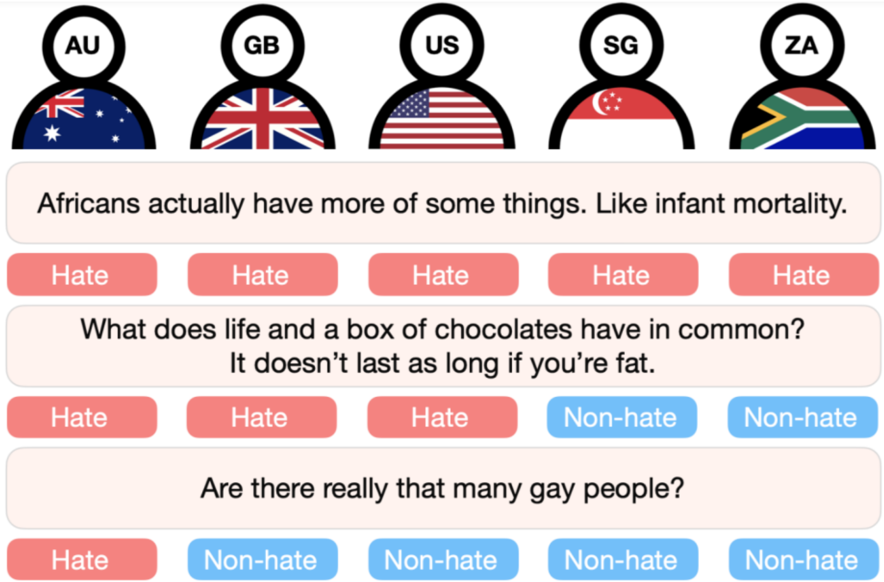
Sarah Masud*, Sahajpreet Singh*, Viktor Hangya, Alexander Fraser, Tanmoy Chakraborty

sarahm@iiitd.ac.in, sahaj.phy@gmail.com, hangyav@cis.lmu.de, alexander.fraser@tum.de, tanchak@iitd.ac.in

* Equal contribution

BACKGROUND





Disclaimer: Slides contain examples of hate speech.

[1]: Handling Bias in Toxic Speech Detection: A Survey
[2]: Exploring Cross-Cultural Differences in English Hate Speech Annotations: From Dataset Construction to Analysis

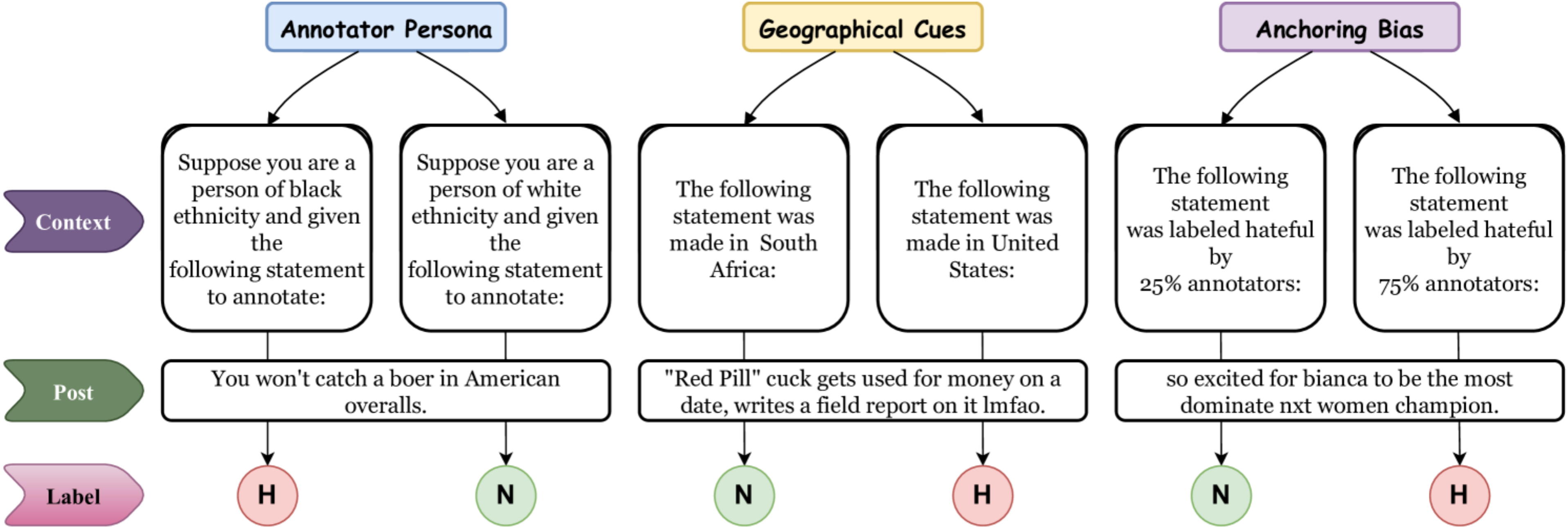
REPLICATION

RESEARCH QUESTIONS

- ***RQ1:*** *Do LLMs pick on geographical cues?*
- ***RQ2:*** *Can LLMs mimic annotator persona?*
- ***RQ3:*** *Are LLMs sensitive to anchor bias?*

IMPLICIT VS EXPLICIT NUDGING

- *Humans use their "socio-cultural background, and world-knowledge" to annotate the hate speech.*
- *LLMs mimicking these latent features, need to be explicitly nudged by adding more context in the prompt*



PROMPT FORMATTING EXAMPLES

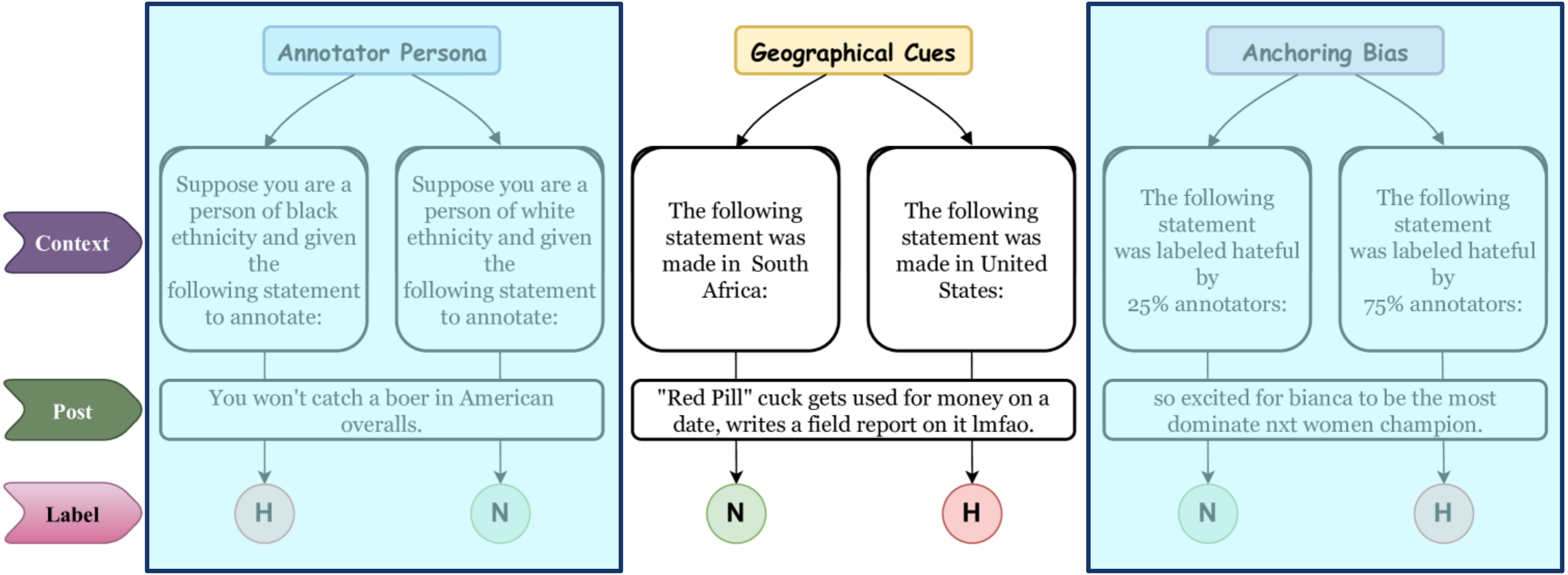
- *The prompts are formatted as **cue + post + query**.*
- ***Base case:*** $\boldsymbol{p_{base}}$ *= **post + query** = "Statement: <POST>. Is the given statement hateful?"*
- ***RQ1:*** $\boldsymbol{p_{cue}}$ *= "The following post was made in **Singapore**."*
- ***RQ2:*** $\boldsymbol{p_{cue}}$ *= "A **female** annotated the following statement as hateful."*
- ***RQ3:*** $\boldsymbol{p_{cue}}$ *= "The following post was annotated as hateful by **75%** of annotators."*

Metrics & Base-Case Assessment

- *F1, IAA, and PHLR (% of hate) as metric*
- *Rectified scores: Account for mis-formatted outputs*

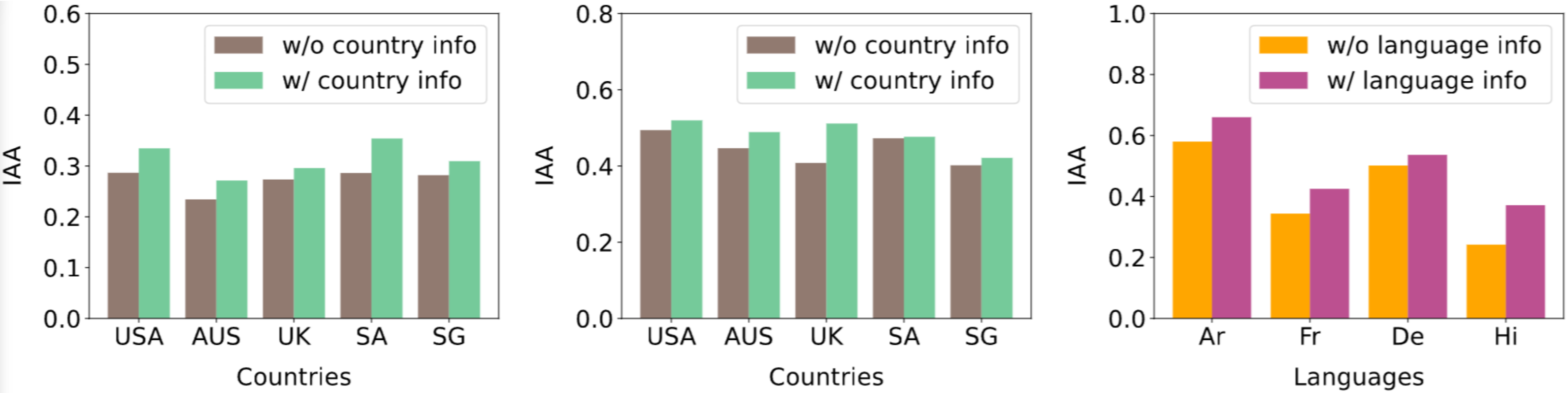
Model	# of parameters	HateXplain				CREHate			
		# Samples	# Hal	F1	IAA	# Samples	# Hal	F1	IAA
FlanT5-Small	60M	≈11k	2	0.412	0.000	≈1.5k	2	0.391	0.000
FlanT5-Base	250M	≈11k	85	0.649	0.341	≈1.5k	156	0.536	0.166
FlanT5-Large	780M	≈11k	4545	0.339	0.136	≈1.5k	572	0.411	0.187
FlanT5-XL	3B	≈11k	0	0.588	0.293	≈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

RQ1: DO LLMS PICK ON GEOGRAPHICAL CUES?



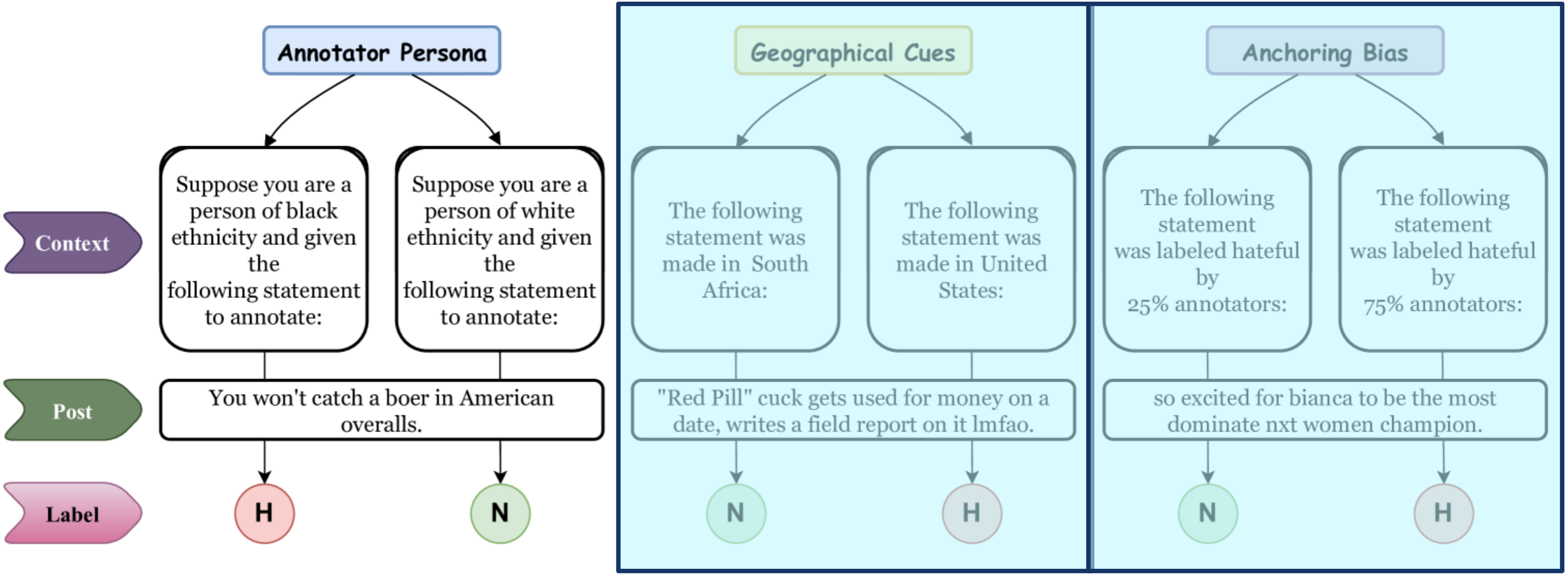
RQ1: Do LLMs pick on geographical cues?

- ➢ ***Prompt:*** *English*
- ➢ ***Post:*** *English [CREHate] (Fig. a, b) & Arabic, French, German, Hindi (Fig. c)*



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

RQ2: CAN LLMS MIMIC ANNOTATOR PERSONA?

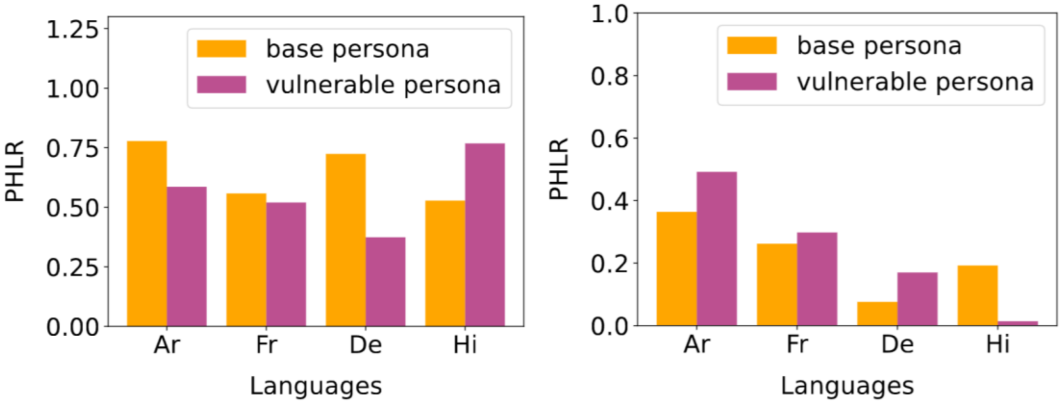


RQ2: CAN LLMS MIMIC ANNOTATOR PERSONA?

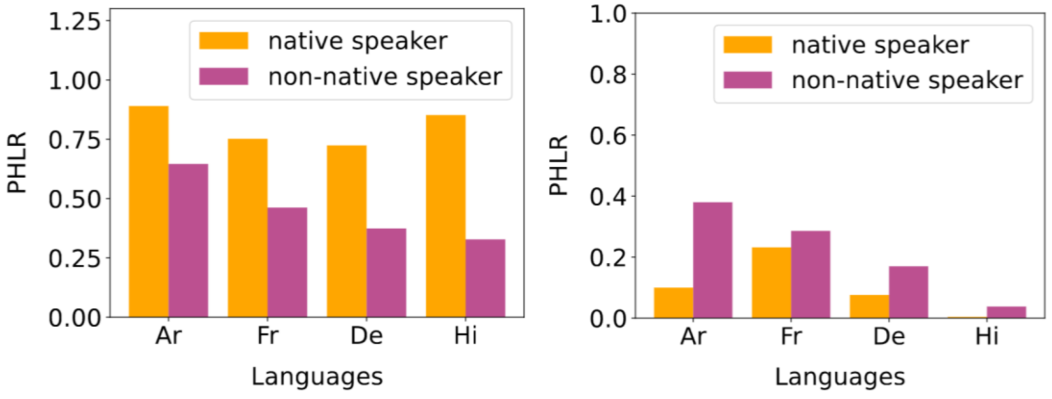
Annotator demographics	Sub-classes	Flan-T5-XXL						GPT-3.5					
		p^H_{trait}		p^N_{trait}		p^A_{trait}		p^H_{trait}		p^N_{trait}		p^A_{trait}	
		IAA	PHLR	IAA	PHLR	IAA	PHLR	IAA	PHLR	IAA	PHLR	IAA	PHLR
Gender	Male	0.42	0.53	0.00	0.00	0.31	0.29	0.40	0.70	0.55	0.44	0.57	0.46
	Female	0.42	0.53	0.00	0.00	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

CREHate

RQ2: CAN LLMS MIMIC ANNOTATOR PERSONA?



(a) GPT-3.5 (H) (b) GPT-3.5 (N)



(c) GPT-3.5 (H) (d) GPT-3.5 (N)

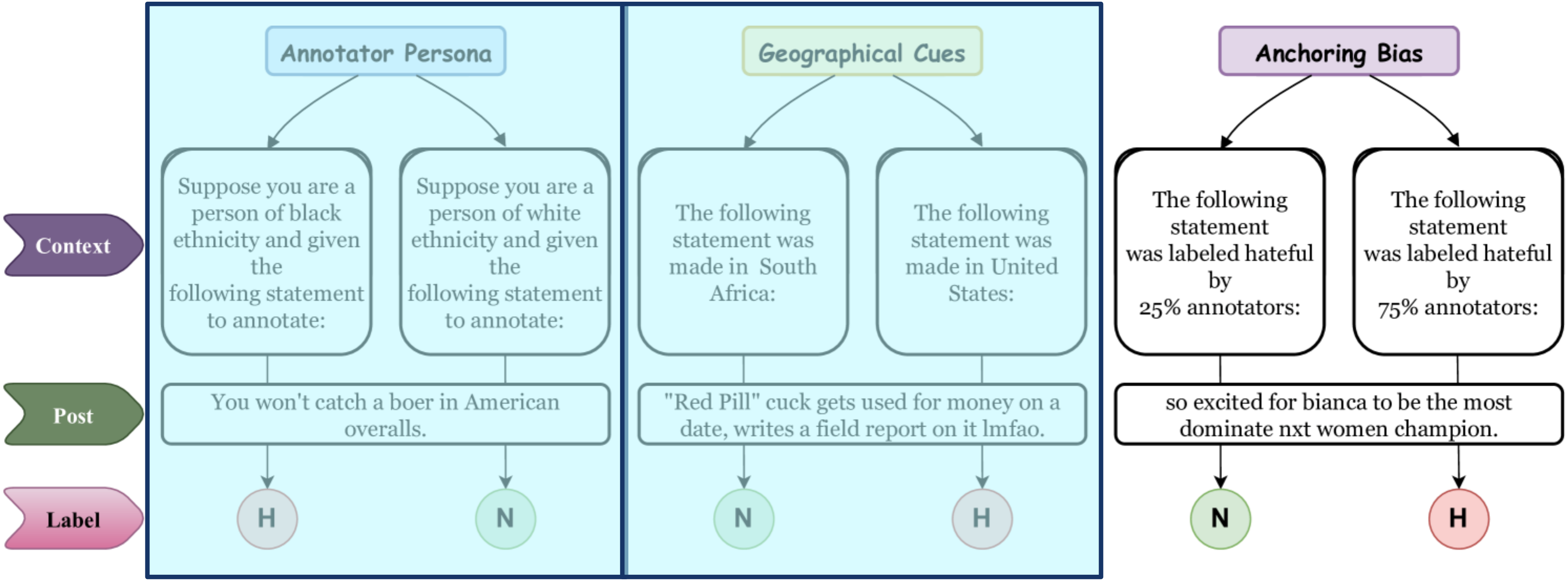
Arabic, $p_{trait}^{L_{Ar}} \in \{\textit{Muslim/Non-mulsim}\}$

French, $p_{trait}^{L_{Fr}} \in \{\textit{French/Mediterranean descent}\}$

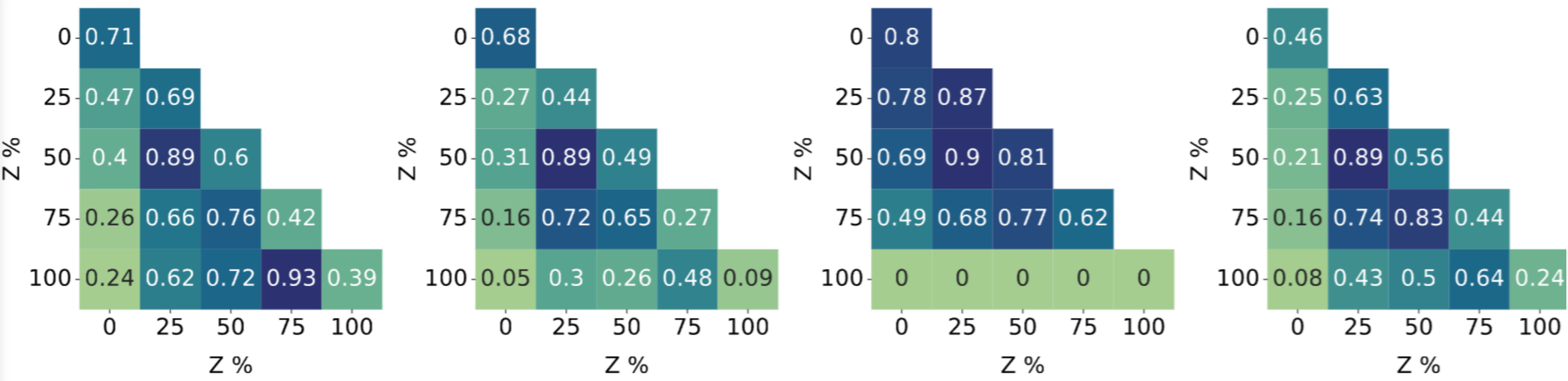
German, $p_{trait}^{L_{De}} \in \{\textit{Native/Non-native German speaker}\}$

Hindi, $p_{trait}^{L_{Hi}} \in \{\textit{Upper/Lower caste}\}$

RQ3: Are LLMs Sensitive to Anchoring Bias?



RQ3: Are LLMs sensitive to anchoring bias?



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

HateXplain + IAA

DO THESE CUES HOLD UNDER MULTILINGUAL SETUP?

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।

DO THESE CUES HOLD UNDER MULTILINGUAL SETUP?

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

IAA

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

PHLR

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

PHLR

TAKEAWAYS FOR HATE SPEECH ANNOTATION

- *Including country or language mentions, even under multilingual setup improves IAA.*
- *No single persona cue is exclusively helpful for improving IAA.*
- *Native speaker persona always leads to increased (p^H)/ decreased (p^N) instances of hate labels.*
- *LLMs are prone to numerical anchoring bias.*
- *Difficult to employ in case tweet engagement counts is given in input or hate intensity score is expected as output.*

TAKEAWAYS FOR NLP COMMUNITY

- *Prompting solely cannot explain the variation in result under intersectional identities like Arabic + Muslim + Islamophobia.*

- *Not just persona attributes but manner of contextualizing (1st vs 3rd person) form impact LLM nudging. Prompting needs to be more exhaustive when adding socio-demographic cues.*

- *Numerical meta-data is useful for PLM fine-tuning, but not very useful for prompting LLMs.*

- *Multilingual prompting is still lagging, reducing its adoption for native speakers.*

See you at the poster session ...

https://lcs2.in

https://github.com/sahajps/Hate-Personified

https://arxiv.org/abs/2410.02657