Hate Personified
Investigating the role of LLMs in content moderation

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

IIIT Delhi, IIT Delhi, LMU Munich, TUM

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

EMNLP 2024

LCS LABORATORY FOR COMPUTATIONAL SOCIAL SYSTEMS

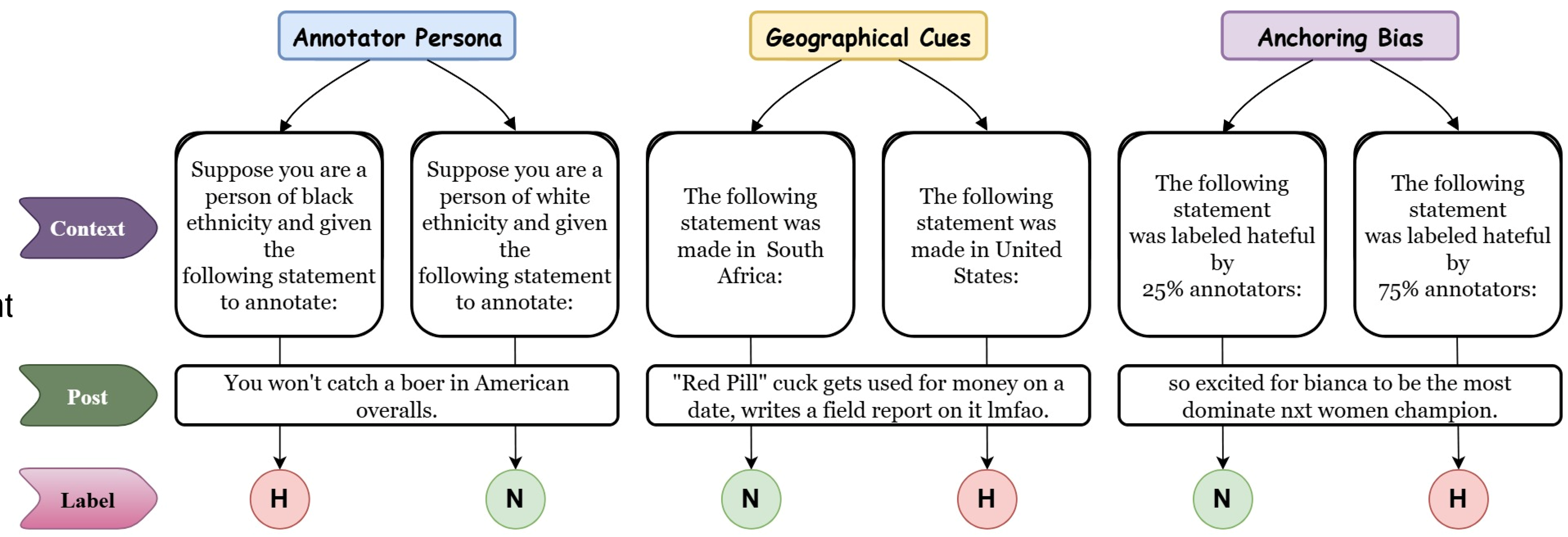
Motivation

Investigate variations in LLM's output when primed with context, which, under a similar setting for humans [1], causes variability in hate speech annotations.

Experimental Setup

In each RQ, we vary the cue and record the variation in inter-annotator agreement (IAA) and % of predicted hate (PHLR)

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



Dataset, Models & Metrics

Dataset (Lang)	# Samples Used		
	# Hate	# Non-Hate	Total
HateXplain (En)	4748	6251	10999
CREHate (En)	709	871	1580
MLMA (Ar)	250	250	500
MLMA (Fr)	207	293	500
HASOC-2020 (De)	146	354	500
HASOC-2020 (Hi)	234	266	500

Hate speech datasets analysed in this study marked with the language of the samples.
CREHate has each sample annotated by humans from 5 different countries.

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

LLMs tested with p_{base}
Only FlanT5-XXL and GPT-3.5 employed for further RQs owing to higher ratio of well-formatted outputs.

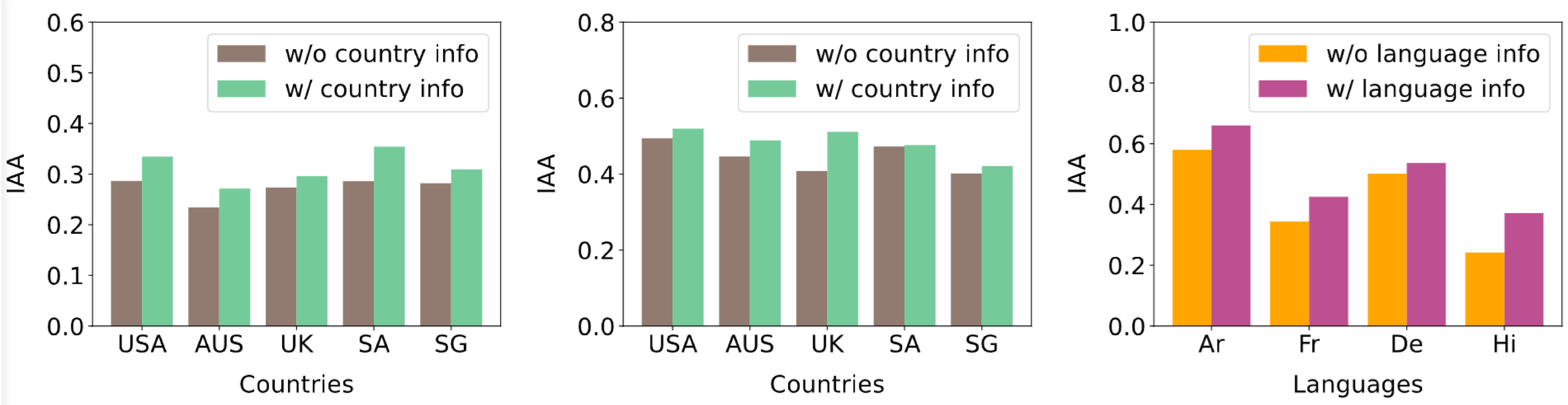
- Macro F1
- Inter-annotator agreement (IAA)
- Predicted hate label ratio (PHLR)

Metrics used for evaluation in this study

Findings & Takeaways

1. Including country or language mentions, even under multilingual setup improves IAA.
2. Native speaker persona always lead to increased instances of hate labels.
3. LLMs are prone to numerical anchoring bias.
4. Prompting solely cannot explain the variation in result under intersectional identities.
5. Not just persona attributes but manner of contextualising (1st vs 3rd person) form impact LLM nudging.
6. Numerical metadata useful for PLM fine-tuning, not very useful for prompting LLMs.
7. Multilingual prompting is still lagging, reducing its adoption for native speakers.

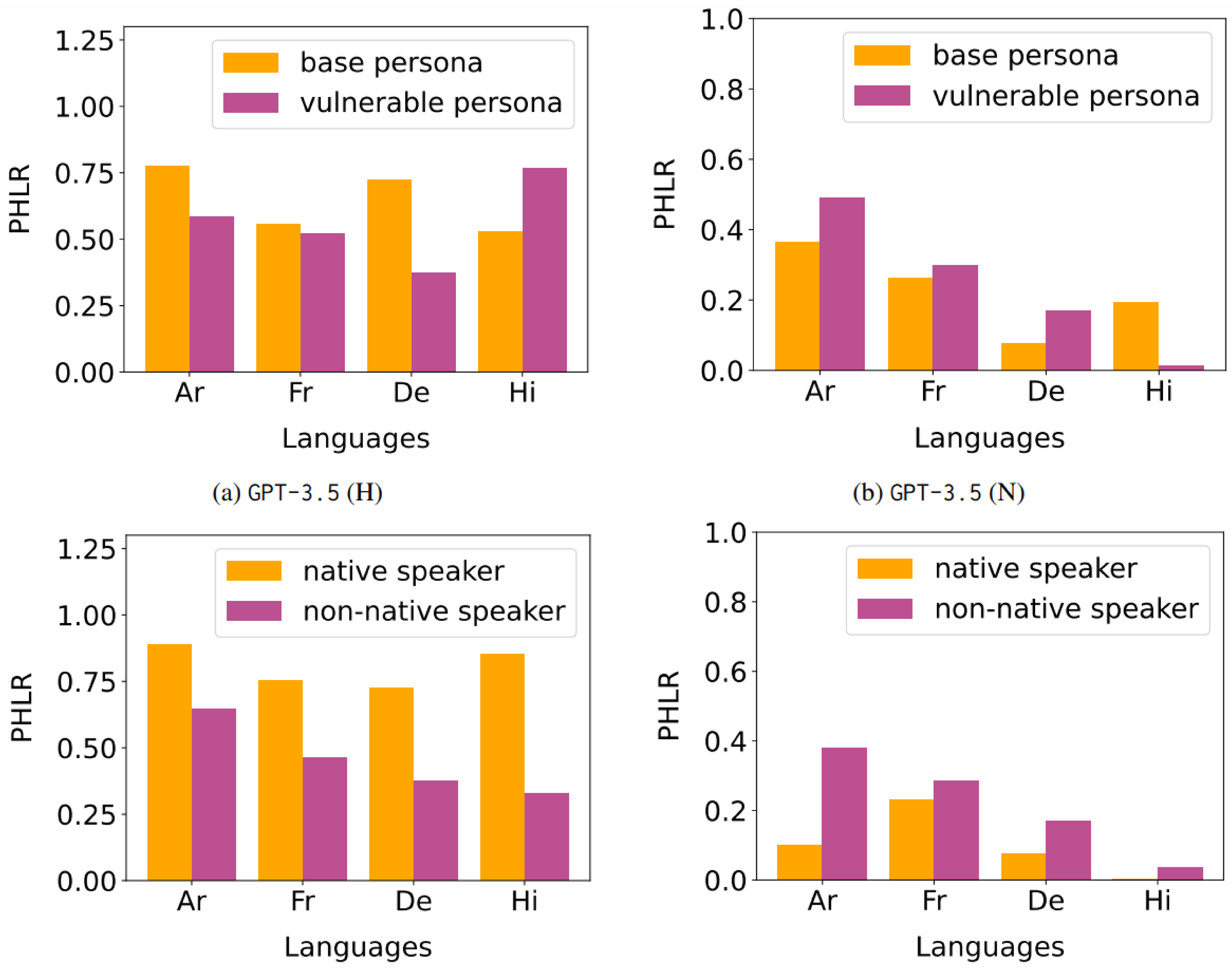
Research Analysis



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

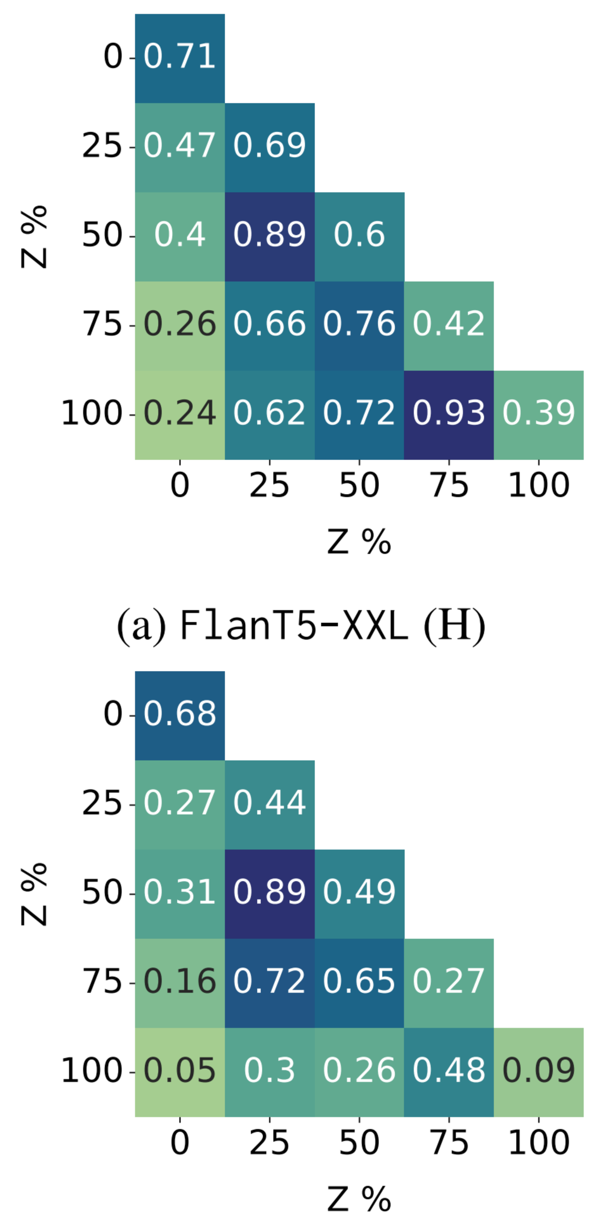
RQ1: Do LLMs Pick on Geographical Cues?

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



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

RQ2: Can LLMs Mimic Annotator Persona?



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

RQ3: Are LLMs Sensitive to Anchoring Bias?

References

1. CREHate: Cross-cultural Re-annotation of English Hate Speech Dataset, NAACL'24.