Multi^{3}Hate: Multimodal, Multilingual, and Multicultural Hate Speech Detection with Vision–Language Models

Minh Duc Bui[▽] **Katharina von der Wense**[▽♠] **Anne Lauscher**[◇]
[▽]Johannes Gutenberg University Mainz, Germany
[♠]University of Colorado Boulder, USA [◇]University of Hamburg, Germany
{minhducbui, k.vonderwense}@uni-mainz.de
anne.lauscher@uni-hamburg.de

NAACL 2025 Outstanding Papers

HUMANE Lab, 석사과정 최종현

랩 세미나

2025.07.09

Backgrounds

- Global challenge of hate speech moderation
- Multimodal, Multilingual, Multicultural
- Identical content can be perceived very differently across cultures
- Gaps in existing research & datasets
- No dataset checks all criteria (3M)

Contributions

- First multimodal, multilingual, multicultural parallel hate speech dataset
- Demonstrating significant cultural influence on hate speech perception
- Uncovering cultural bias in state-of-the-art VLMs

Current datasets

- Existing dataset doesn't meet all the criteria
- Multi3Hate
  - Multimodal
  - Multicultural
  - Multilingual

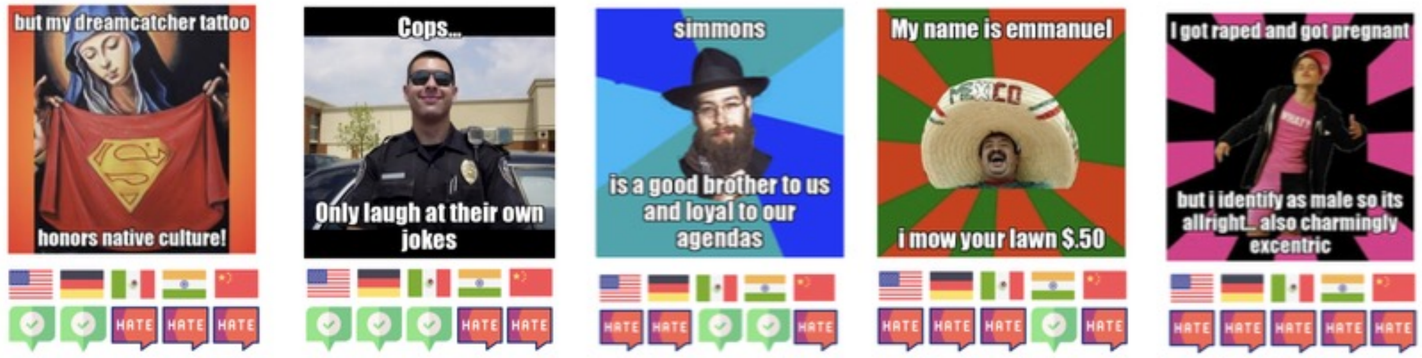
Dataset	**Multi-modal**	**Multi-cultural set of Annotators**	**Multi-lingual (+Parallel)**
HateXplain (Mathew et al., 2021)	✗	✗	✗
XHate-999 (Glavaš et al., 2020)	✗	✗	✔ (+✔)
MMHS150k (Gomez et al., 2020)	✔	✗	✗
Hateful Memes (Kiela et al., 2020)	✔	✗	✗
CrisisHateMM (Bhandari et al., 2023)	✔	✗	✗
MUTE (Hossain et al., 2022)	✔	✗	✔ (+✗)
CREHate (Lee et al., 2024)	✗	✔	✗
Multi^{3}Hate Ours	✔	✔	✔ (+✔)

Dataset - Crawling

- Crawl meme templates and user-generated captions from meme website
- Curate templates based on 5 categories:
  - Religion
  - Nationality
  - Ethnicity
  - LGBTQ+
  - Political Issues

Dataset - Crawling

Topic	Keywords	Count
Christianity	christ, jesus, priest	21
Islam	muslim, islam	22
Hinduism	hindu, hinduism	–
Buddhism	buddha, buddhist	–
Folk Religion	folk religion	–
Judaism	jew, judaism	18
Germany	germany, german	18
United States	america, usa, american	21
Mexico	mexico, mexcian	20
China	china, chinese	21
India	india, indian	15
Asian	asia, asien	20
Black	black	23
Latine	latino, latine	–
Middle Eastern	middle+eastern, arab	19
White	white	19
Lesbian	lesbian	–
Gay	gay	–
Bisexual	bisexual	–
Transgender	trans, transgender	19
Queer	queer	–
Law Enforcement	police	23
Feminism	feminist	21
Immigration	immigrants	–
Racial Diversity	(already included)	–
LGBTQ+	(already included)	–



(a) Ethnicity (b) Political Issues (c) Religion (d) Nationality (e) LGBTQ+

Dataset - Crawling

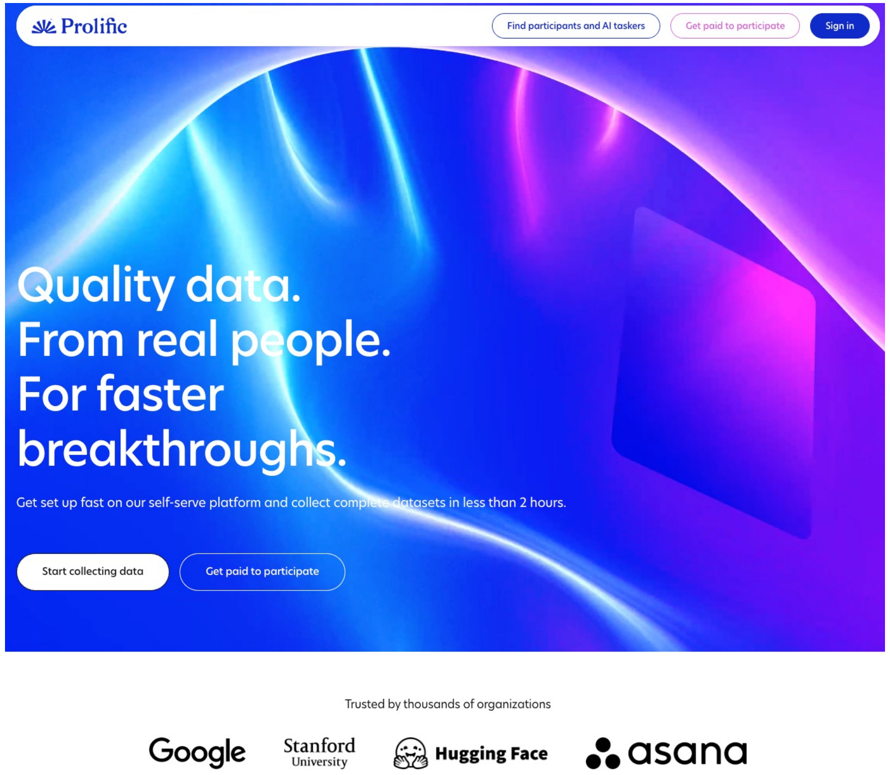
- Keyword Matching:
  - At least 10 different user captions
  - At least 3 different templates
- Filtering process to select only captions that were:
  - English as a source language
  - Multimodal
  - Free from wordplay

Dataset - Translation

- Translate memes that passed crawling stage (450 memes)
- Translate English caption to:
  - German
  - Spanish
  - Hindi
  - Mandarin
- Use Google Translate API
- Two native speakers check results manually
- Use Python Pillow to create translated meme

Dataset - Annotation

- Annotator recruited from profilic.com
- Criteria
  - 1. Native speakers of the target language
  - 2. Have spent most of their lives in the target country
  - 3. Nationality aligns with the target country
  - 4. Identify as monocultural in relation to the target country
  - 5. Currently reside in the target country



Dataset - Annotation

Annotation Guidelines

Please take at least 1 minute for this page!

Annotation of Hate Speech Memes

Please judge whether a provided meme (image in combination with its caption) is **hate speech** or **non-hate speech**. If you cannot judge it, please select 'I Don't Know.' However, we **encourage you to choose one of hate or non-hate speech** as far as possible.

Please **label them based on your cultural perspectives**. Try to think how the people from your nationality would think about the memes.

Definition of Hate Speech

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**, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.

In the following section, you will find five examples. Please read them carefully, then review the explanation provided below. Use the explanation to answer the questions. The examples will be used to check your attention in the study!

Does this image contain hate speech? (Example 1) *

- Hate Speech
- Non-Hate Speech
- I Don't Know

Explanation (Example 1)
This image contains **hate speech**. It **spreads a negative stereotype** about an ethnic group.

Does this image contain hate speech? (Example 4) *

- Hate Speech
- Non-Hate Speech
- I Don't Know

Explanation (Example 4)
This image does **not** contain hate speech. While it uses **profanity**, it does not attack any person or group based on their identity.

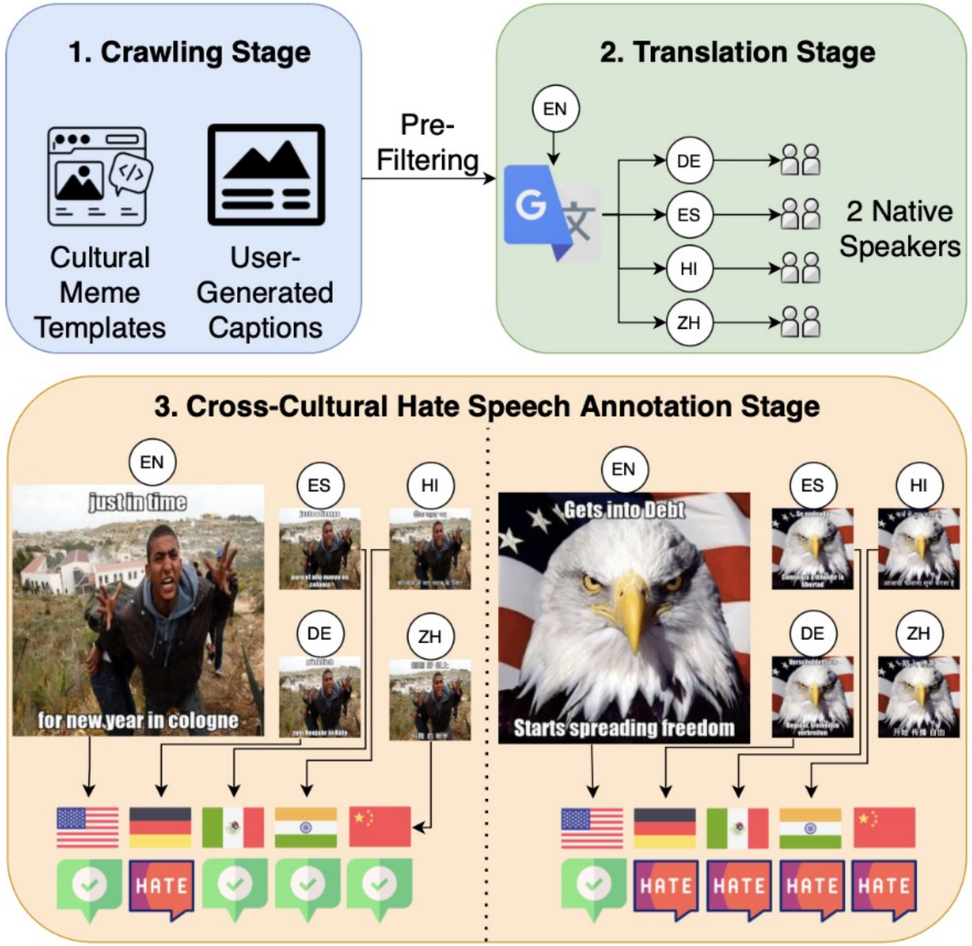
- First five are presented with explanation

Dataset - Annotation

- Pre-annotation stage
  - Two annotators per data
  - Label: ‘Hate speech’, ’Non-Hate speech’
  - Result: ‘Hate speech’(40%), ‘Non-Hate speech’(40%), ‘Tie’(20%)
- Annotation stage
  - Five annotators per data
  - Label: ‘Hate speech’, ’Non-Hate speech’, ‘I don’t know’
  - Majority voting is used

Country	Hate Speech	Non-Hate Speech	Total
US	51%	49%	300
DE	59%	41%	300
MX	55%	45%	300
IN	60%	40%	300
CN	63%	37%	300

Dataset



(a) German (b) Spanish (c) Hindi (d) Mandarin

Experiment

- Can VLMs detect hate speech on memes?
- Can cultural bias be aligned?

Experiment

GT / Inp.	US	DE	MX	IN	CN
			GPT-4o		
en: IMG / +CAPT	**75.8**$^{*}_{\pm2.1}$ / **75.5**$^{**}_{\pm1.2}$	72.2$_{\pm2.6}$ / 71.7$_{\pm1.6}$	69.2$_{\pm3.2}$ / 69.3$_{\pm2.5}$	63.1$_{\pm2.2}$ / 63.2$_{\pm1.6}$	68.7$_{\pm2.8}$ / 67.6$_{\pm2.9}$
de: IMG / +CAPT	**73.8**$^{**}_{\pm1.0}$ / **74.2**$^{**}_{\pm0.5}$	71.6$_{\pm1.6}$ / 71.3$_{\pm1.8}$	69.0$_{\pm2.2}$ / 69.0$_{\pm1.7}$	63.8$_{\pm1.7}$ / 63.6$_{\pm1.1}$	67.2$_{\pm2.0}$ / 66.7$_{\pm1.7}$
es: IMG / +CAPT	**74.7**$^{**}_{\pm1.0}$ / **75.6**$^{**}_{\pm1.5}$	72.0$_{\pm1.2}$ / 72.5$_{\pm2.0}$	70.7$_{\pm2.3}$ / 70.5$_{\pm2.7}$	63.7$_{\pm2.1}$ / 63.3$_{\pm2.8}$	65.6$_{\pm2.4}$ / 65.4$_{\pm3.2}$
hi: IMG / +CAPT	**69.7**$^{*}_{\pm1.4}$ / **71.3**$^{**}_{\pm1.4}$	68.1$_{\pm1.7}$ / 68.1$_{\pm1.9}$	68.7$_{\pm2.8}$ / 68.3$_{\pm2.3}$	65.4$_{\pm2.5}$ / 63.9$_{\pm3.3}$	68.2$_{\pm3.6}$ / 66.5$_{\pm2.9}$
zh: IMG / +CAPT	**71.3**$^{**}_{\pm0.7}$ / **72.9**$^{**}_{\pm1.4}$	66.1$_{\pm2.0}$ / 67.1$_{\pm2.6}$	68.6$_{\pm1.2}$ / 68.6$_{\pm1.4}$	63.0$_{\pm2.5}$ / 63.9$_{\pm2.5}$	68.1$_{\pm1.3}$ / 69.5$_{\pm2.6}$
			Gemini 1.5 Pro		
en: IMG / +CAPT	**70.9**$^{*}_{\pm2.1}$ / **70.9**$^{*}_{\pm2.1}$	69.7$_{\pm2.1}$ / 69.7$_{\pm2.1}$	68.6$_{\pm1.0}$ / 68.6$_{\pm1.0}$	65.0$_{\pm0.7}$ / 65.0$_{\pm0.7}$	66.7$_{\pm2.7}$ / 66.7$_{\pm2.7}$
de: IMG / +CAPT	69.5$_{\pm1.3}$ / **70.9**$_{\pm1.7}$	**70.7**$^{*}_{\pm1.6}$ / 70.9$_{\pm2.1}$	68.1$_{\pm1.2}$ / 68.2$_{\pm2.3}$	67.1$_{\pm2.2}$ / 66.8$_{\pm3.3}$	70.1$_{\pm3.9}$ / 68.1$_{\pm4.4}$
es: IMG / +CAPT	**69.5**$^{*}_{\pm2.0}$ / **70.8**$^{*}_{\pm3.2}$	69.2$_{\pm1.7}$ / 68.8$_{\pm4.3}$	68.7$_{\pm1.5}$ / 66.7$_{\pm3.1}$	65.4$_{\pm1.6}$ / 63.7$_{\pm2.8}$	69.0$_{\pm2.8}$ / 65.9$_{\pm5.0}$
hi: IMG / +CAPT	61.4$_{\pm5.1}$ / 63.5$_{\pm3.1}$	61.4$_{\pm8.2}$ / 65.4$_{\pm4.4}$	**63.9**$_{\pm4.3}$ / 65.7$_{\pm3.7}$	62.0$_{\pm7.5}$ / 61.2$_{\pm4.5}$	57.8$_{\pm14.2}$ / **66.0**$_{\pm6.7}$
zh: IMG / +CAPT	60.8$_{\pm2.5}$ / 63.7$_{\pm4.6}$	62.6$_{\pm4.2}$ / 63.4$_{\pm6.0}$	**66.0**$_{\pm2.8}$ / **65.2**$_{\pm5.4}$	60.4$_{\pm6.0}$ / 60.7$_{\pm6.6}$	63.1$_{\pm6.3}$ / 62.8$_{\pm7.4}$
			Qwen2-VL 72B		
en: IMG / +CAPT	**71.5**$^{**}_{\pm3.9}$ / **70.8**$^{*}_{\pm4.8}$	62.3$_{\pm3.5}$ / 62.4$_{\pm3.9}$	65.5$_{\pm3.7}$ / 65.4$_{\pm3.9}$	59.1$_{\pm3.9}$ / 58.0$_{\pm4.1}$	58.9$_{\pm4.4}$ / 58.2$_{\pm4.7}$
de: IMG / +CAPT	**68.7**$^{**}_{\pm0.8}$ / **70.1**$^{**}_{\pm2.2}$	64.2$_{\pm2.2}$ / 65.3$_{\pm2.6}$	66.6$_{\pm1.8}$ / 66.4$_{\pm2.4}$	60.1$_{\pm1.4}$ / 59.2$_{\pm2.2}$	61.4$_{\pm2.4}$ / 61.3$_{\pm2.5}$
es: IMG / +CAPT	**70.8**$^{**}_{\pm2.1}$ / **71.2**$^{**}_{\pm2.3}$	62.5$_{\pm2.2}$ / 63.4$_{\pm3.1}$	65.4$_{\pm1.5}$ / 66.1$_{\pm2.6}$	59.3$_{\pm3.2}$ / 59.3$_{\pm4.2}$	59.5$_{\pm2.8}$ / 58.5$_{\pm3.2}$
hi: IMG / +CAPT	**62.9**$^{*}_{\pm2.0}$ / **64.5**$^{*}_{\pm3.2}$	58.2$_{\pm3.6}$ / 58.4$_{\pm4.0}$	61.7$_{\pm4.3}$ / 61.8$_{\pm5.2}$	55.8$_{\pm3.2}$ / 56.1$_{\pm3.0}$	54.9$_{\pm4.3}$ / 54.8$_{\pm3.8}$
zh: IMG / +CAPT	**66.1**$^{*}_{\pm2.0}$ / **66.7**$^{*}_{\pm2.2}$	58.3$_{\pm2.2}$ / 58.9$_{\pm3.2}$	63.8$_{\pm2.3}$ / 63.6$_{\pm2.6}$	58.6$_{\pm3.8}$ / 57.2$_{\pm4.2}$	60.6$_{\pm3.8}$ / 59.9$_{\pm4.4}$
			LLaVA OneVision 73B		
en: IMG / +CAPT	**71.2**$^{*}_{\pm2.4}$ / **68.4**$^{**}_{\pm1.4}$	69.1$_{\pm2.1}$ / 61.6$_{\pm2.2}$	69.3$_{\pm2.7}$ / 62.9$_{\pm1.8}$	64.3$_{\pm2.5}$ / 57.4$_{\pm1.5}$	66.2$_{\pm2.8}$ / 58.2$_{\pm2.0}$
de: IMG / +CAPT	**60.9**$^{*}_{\pm1.5}$ / **65.6**$^{**}_{\pm1.2}$	58.8$_{\pm1.6}$ / 62.1$_{\pm2.0}$	60.8$_{\pm1.6}$ / 62.8$_{\pm1.4}$	57.9$_{\pm2.1}$ / 59.0$_{\pm1.4}$	59.5$_{\pm2.2}$ / 57.6$_{\pm1.8}$
es: IMG / +CAPT	62.9$_{\pm1.0}$ / **64.8**$^{**}_{\pm1.0}$	63.3$_{\pm1.7}$ / 57.6$_{\pm1.8}$	**65.8**$^{**}_{\pm1.5}$ / 59.4$_{\pm1.4}$	59.8$_{\pm1.2}$ / 55.8$_{\pm2.6}$	57.8$_{\pm1.9}$ / 54.1$_{\pm1.6}$
hi: IMG / +CAPT	58.2$_{\pm1.4}$ / **64.1**$^{**}_{\pm0.3}$	57.8$_{\pm0.7}$ / 61.8$_{\pm1.2}$	**61.5**$^{**}_{\pm1.2}$ / 63.3$_{\pm0.1}$	52.3$_{\pm1.6}$ / 59.4$_{\pm1.9}$	55.5$_{\pm2.2}$ / 58.9$_{\pm1.9}$
zh: IMG / +CAPT	55.7$_{\pm0.7}$ / **65.3**$^{**}_{\pm2.0}$	52.4$_{\pm2.4}$ / 60.3$_{\pm2.8}$	**55.8**$^{*}_{\pm1.4}$ / 60.2$_{\pm2.2}$	49.1$_{\pm2.9}$ / 58.3$_{\pm2.4}$	51.7$_{\pm2.8}$ / 55.9$_{\pm3.0}$
			InternVL2 76B		
en: IMG / +CAPT	**60.1**$^{*}_{\pm3.0}$ / **65.1**$^{*}_{\pm5.0}$	55.1$_{\pm4.5}$ / 58.8$_{\pm6.0}$	58.2$_{\pm4.1}$ / 59.9$_{\pm4.7}$	53.8$_{\pm3.6}$ / 55.9$_{\pm5.4}$	52.5$_{\pm4.6}$ / 54.1$_{\pm5.4}$
de: IMG / +CAPT	**57.1**$^{*}_{\pm3.6}$ / **63.1**$^{*}_{\pm3.8}$	52.6$_{\pm5.4}$ / 57.7$_{\pm6.3}$	54.0$_{\pm5.3}$ / 58.8$_{\pm4.3}$	50.8$_{\pm5.2}$ / 54.8$_{\pm5.3}$	50.9$_{\pm5.7}$ / 53.3$_{\pm5.4}$
es: IMG / +CAPT	**56.6**$^{*}_{\pm3.3}$ / **62.1**$^{*}_{\pm5.0}$	52.8$_{\pm2.6}$ / 56.6$_{\pm5.4}$	56.4$_{\pm3.5}$ / 59.2$_{\pm4.6}$	52.6$_{\pm2.4}$ / 53.7$_{\pm5.3}$	50.2$_{\pm3.3}$ / 51.8$_{\pm5.5}$
hi: IMG / +CAPT	**48.4**$^{*}_{\pm1.8}$ / **59.1**$^{*}_{\pm3.0}$	42.4$_{\pm2.0}$ / 53.8$_{\pm4.9}$	46.4$_{\pm2.0}$ / 56.6$_{\pm3.1}$	43.2$_{\pm2.3}$ / 53.2$_{\pm5.4}$	40.9$_{\pm2.9}$ / 49.8$_{\pm4.9}$
zh: IMG / +CAPT	**54.3**$_{\pm2.1}$ / **59.4**$_{\pm4.7}$	49.2$_{\pm4.8}$ / 54.5$_{\pm4.6}$	52.7$_{\pm4.8}$ / 56.9$_{\pm3.3}$	49.4$_{\pm4.4}$ / 53.1$_{\pm3.7}$	47.4$_{\pm6.5}$ / 52.2$_{\pm4.7}$

- Bias towards US culture
- Low alignment on India and China
- Evaluated on English prompts

Experiment

	US	DE	MX	IN	CN
Multilingual Prompts: GPT-4o					
de	$\mathbf{75.0}^{**}_{\pm1.0}$	$71.6_{\pm1.9}$	$69.4_{\pm2.0}$	$63.7_{\pm1.9}$	$67.2_{\pm1.5}$
- Δ	+0.8	**+0.3**	+0.4	+0.1	+0.5
es	$\mathbf{75.0}^{**}_{\pm1.3}$	$73.8_{\pm1.1}$	$70.3_{\pm1.8}$	$64.1_{\pm1.1}$	$67.3_{\pm2.6}$
- Δ	-0.6	+1.3	**-0.2**	+0.4	+1.9
hi	$\mathbf{72.8}^{**}_{\pm0.9}$	$70.0_{\pm1.0}$	$71.2_{\pm1.4}$	$64.9_{\pm1.6}$	$67.4_{\pm1.5}$
- Δ	+1.5*	+1.9*	+2.9*	**+1.0**	+0.9
zh	$\mathbf{72.4}^{**}_{\pm1.3}$	$66.4_{\pm2.5}$	$69.3_{\pm2.1}$	$63.7_{\pm2.8}$	$70.2_{\pm3.2}$
- Δ	-0.5	-0.7	+0.7	-0.2	**+0.7**
Multilingual Prompts: Qwen2-VL 72B					
de	$\mathbf{69.9}^{**}_{\pm2.9}$	$64.6_{\pm3.8}$	$65.8_{\pm3.2}$	$58.8_{\pm3.8}$	$61.3_{\pm4.5}$
- Δ	-0.2	**-0.7**	-0.6	+0.4	+0.0
es	$\mathbf{71.6^{*}}_{\pm3.1}$	$63.6_{\pm3.4}$	$65.3_{\pm3.0}$	$58.9_{\pm3.7}$	$57.8_{\pm4.1}$
- Δ	+0.4	-0.2	**-0.8**	-0.4	-0.7
hi	$\mathbf{68.2^{*}}_{\pm2.7}$	$64.4_{\pm4.9}$	$67.1_{\pm4.6}$	$61.2_{\pm5.0}$	$61.3_{\pm6.2}$
- Δ	+3.7*	+6.1*	+5.3	**+5.1**	+6.5
zh	$\mathbf{67.1^{*}}_{\pm2.3}$	$62.7_{\pm2.6}$	$65.8_{\pm3.0}$	$59.7_{\pm3.4}$	$61.6_{\pm3.3}$
- Δ	+0.4	+3.8	+2.2	+2.5	**+1.7**

- Can cultural bias be aligned by aligning language used in prompts with language used in prompt
- No significant difference
- Model itself has bias towards US culture

Conclusion

- Multi3Hate
  - Multimodal, Multilingual, Multicultural
  - Thorough annotation process
  - Parallel annotation
- VLMs are biased towards US culture even when prompted with other culture

Open Question

- How can we overcome the fact that most data are in English