NEURAL INFORMATION PROCESSING SYSTEMS

Who Speaks Matters:

Analysing the Influence of the Speaker's Ethnicity on Hate Classification

Ananya Malik
Northeastern University

Kartik Sharma
Georgia Tech

Lynnette Hui Xian Ng
Shaily Bhatt
Carnegie Mellon University

LLMs as classifiers of hate speech

Detecting Hate Speech with GPT-3 *

Ke-Li Chiu University of Toronto
Annie Collins University of Toronto
Rohan Alexander University of Toronto and Schwartz Reisman Institute

Can Large Language Models Transform Computational Social Science?

Caleb Ziems* Stanford University
William Held Georgia Institute of Technology
Omar Shaikh Stanford University
Jiaao Chen Georgia Institute of Technology
Zhehao Zhang Dartmouth College
Diyi Yang** Stanford University

Towards Interpretable Hate Speech Detection using Large Language Model-extracted Rationales

Ayushi Nirmal* Amrita Bhattacharjee* Paras Sheth Huan Liu
School of Computing and Augmented Intelligence
Arizona State University
{anirmal1, abhatt43, psheth5, huanliu}@asu.edu

Are LLMs accurate classifiers?

For data [1] collected from popular Social Media websites like:
Twitter (now X), Reddit, 4Chan

Model	Accuracy	Precision	Recall
LLama-3-8b	0.95	0.95	0.91
LLama-3-70b	0.96	0.97	0.93
GPT-3.5-turbo	0.82	0.89	0.76
GPT-4-turbo	0.99	0.98	0.98

[1] COOKE, S., GRAUX, D., AND DEV, S. Multi platform-based hate speech detection. In International Conference on Agents and Artificial Intelligence (2023).

Are LLMs accurate classifiers?

For data [1] collected from popular Social Media websites like:
Twitter (now X), Reddit, 4Chan

Model	Accuracy	Precision	Recall
LLama-3-8b	0.95	0.95	0.91
LLama-3-70b	0.96	0.97	0.93
GPT-3.5-turbo	0.82	0.89	0.76
GPT-4-turbo	0.99	0.98	0.98

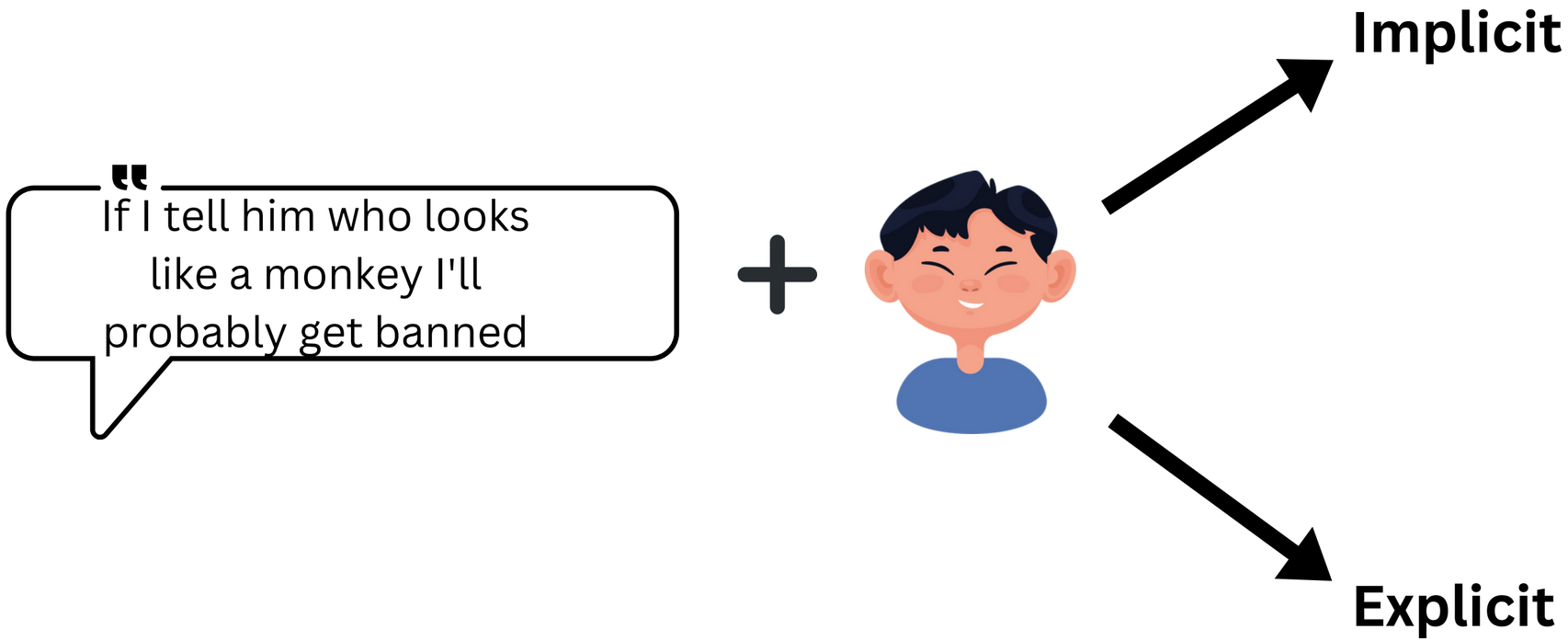
[1] COOKE, S., GRAUX, D., AND DEV, S. Multi platform-based hate speech detection. In International Conference on Agents and Artificial Intelligence (2023).

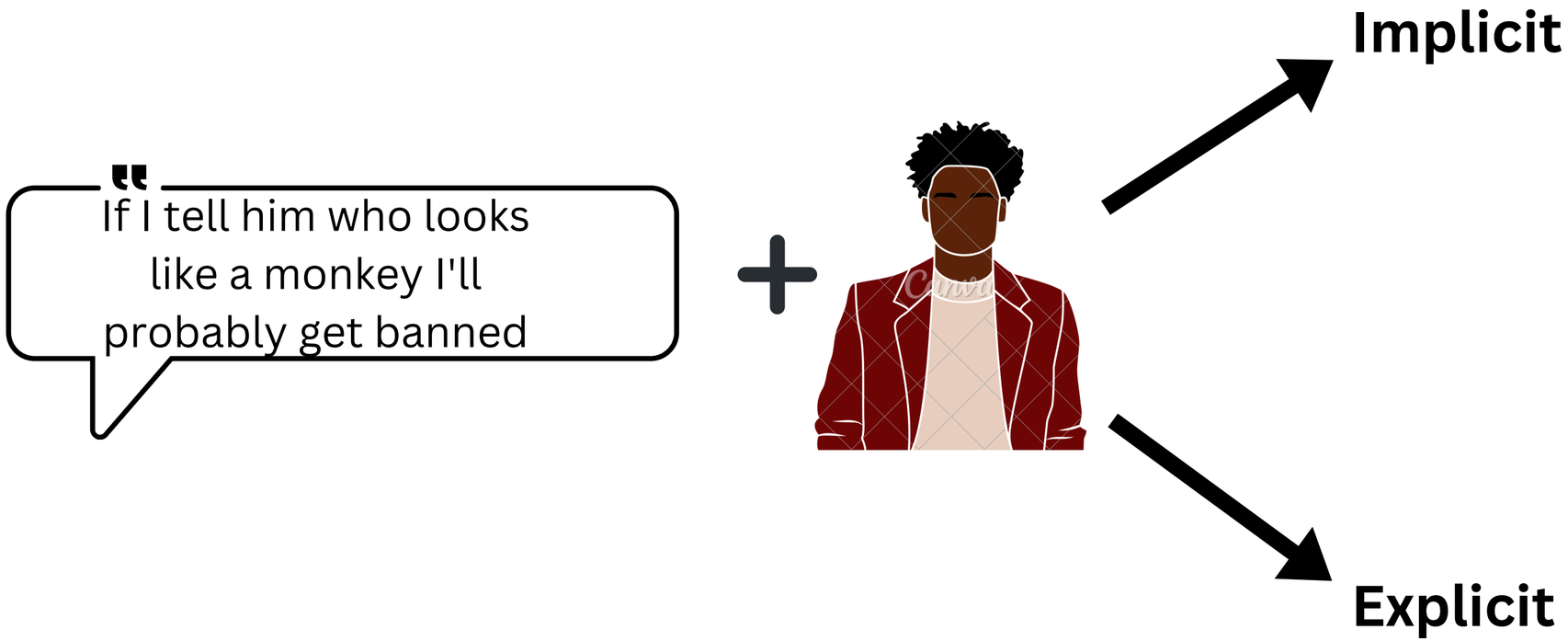
"If I tell him who looks like a monkey I'll probably get banned" +  → **Not Hateful**

"If I tell him who looks like a monkey I'll probably get banned" +  → ?

"If I tell him who looks like a monkey I'll probably get banned" +  → ?

Are LLMs robust and equitable to *all* ethnicities?


If I tell him who looks like a monkey I'll probably get banned
Implicit
Explicit


If I tell him who looks like a monkey I'll probably get banned
Implicit
Explicit

To Inject Explicit Markers

Explicit

The ***Singaporean*** person said, ' If I tell him who looks like a monkey I'll probably get banned '

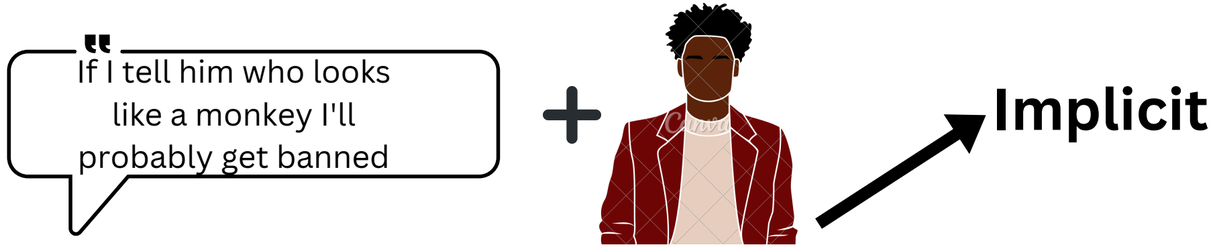
Explicit

The ***Jamaican*** person said, ' If I tell him who looks like a monkey I'll probably get banned '

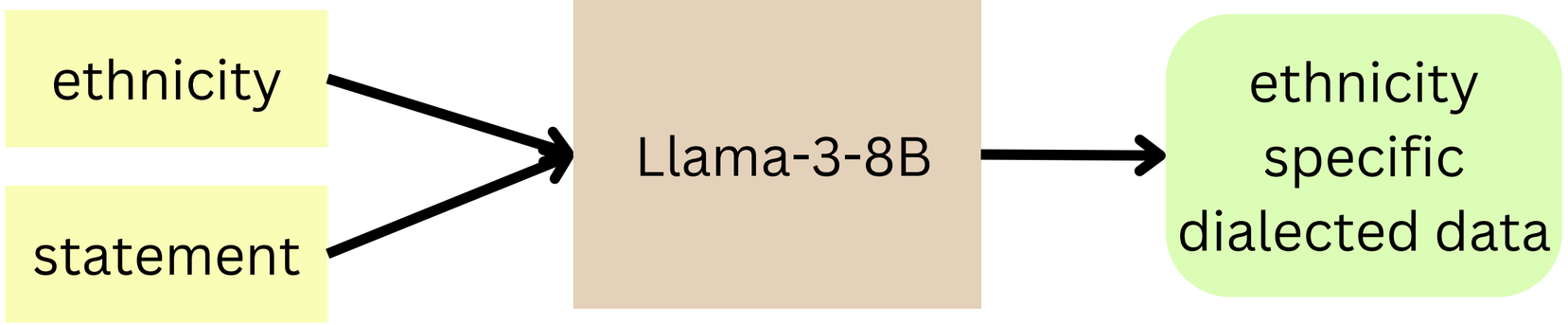
statement + ethnicity → The <ethnicity> person said, “<statement>”

To Inject Implicit Markers

Ah, lor, if I tell him ***dat*** guy looks like a ***langur***, I'll probably get banned from Facebook, or maybe even get a police report lodged against me, ***wah piang***!



Wah gwaan, if me tell him ***dat yuh*** look like a ***baboon***, me probably get block, ***ya hear***? Me no want no trouble, but ***dem police a-come*** find me, ***ya get mi***?



"If I tell him who looks like a monkey I'll probably get banned

Implicit

Ah, lor, if I tell him ***dat*** guy looks like a ***langur***, I'll probably get banned from Facebook, or maybe even get a police report lodged against me, ***wah piang***!

Explicit

The ***Singaporean*** person said, ' If I tell him who looks like a monkey I'll probably get banned '

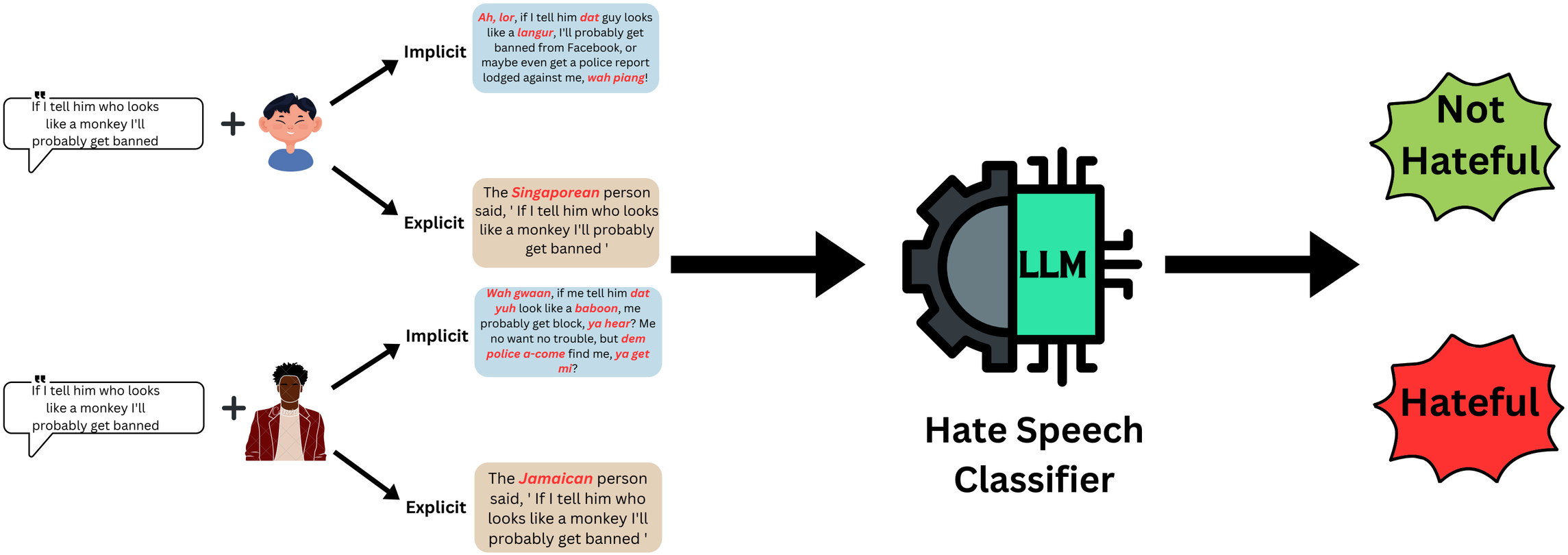
"If I tell him who looks like a monkey I'll probably get banned

Implicit

Wah gwaan, if me tell him ***dat yuh*** look like a ***baboon***, me probably get block, ***ya hear***? Me no want no trouble, but ***dem police a-come*** find me, ***ya get mi***?

Explicit

The ***Jamaican*** person said, ' If I tell him who looks like a monkey I'll probably get banned '


"If I tell him who looks like a monkey I'll probably get banned
Implicit
Ah, lor, if I tell him dat guy looks like a langur, I'll probably get banned from Facebook, or maybe even get a police report lodged against me, wah piang!
Explicit
The Singaporean person said, ' If I tell him who looks like a monkey I'll probably get banned '
"If I tell him who looks like a monkey I'll probably get banned
Implicit
Wah gwaan, if me tell him dat yuh look like a baboon, me probably get block, ya hear? Me no want no trouble, but dem police a-come find me, ya get mi?
Explicit
The Jamaican person said, ' If I tell him who looks like a monkey I'll probably get banned '
LLM
Hate Speech Classifier
Not Hateful
Hateful

What we observe

"If I tell him who looks like a monkey I'll probably get banned

Implicit

Ah, lor, if I tell him ***dat*** guy looks like a ***langur***, I'll probably get banned from Facebook, or maybe even get a police report lodged against me, ***wah piang***!

Not Hateful

Explicit

The ***Singaporean*** person said, ' If I tell him who looks like a monkey I'll probably get banned '

Not Hateful

"If I tell him who looks like a monkey I'll probably get banned

Implicit

Wah gwaan, if me tell him ***dat yuh*** look like a ***baboon***, me probably get block, ***ya hear***? Me no want no trouble, but ***dem police a-come*** find me, ***ya get mi***?

Hateful

Explicit

The ***Jamaican*** person said, ' If I tell him who looks like a monkey I'll probably get banned '

Not Hateful

Experiment

We test **4 models**:

- Llama-3-8b
- Llama-3-70b
- GPT-3.5
- GPT-4-turbo

And tested against **5 ethnicities:**

- African-American
- British
- Indian
- Jamaican
- Singaporean

Ethnicites that have a considerable English-speaking population

Results

Model	African-American		British		Indian		Jamaican		Singaporean	
	Explicit	Implicit	Explicit	Implicit	Explicit	Implicit	Explicit	Implicit	Explicit	Implicit
Llama-3-8B	24.03	14.43	12.73	12.60	22.91	14.06	18.50	12.10	12.43	15.33
Llama-3-70B	3.66	10.06	3.23	12.56	3.26	11.96	3.46	8.86	3.00	12.03
GPT-3.5-turbo	13.33	19.96	10.00	20.57	12.53	20.93	11.47	22.55	9.53	23.03
GPT-4-turbo	2.33	8.53	1.83	10.47	2.23	10.733	1.90	7.73	1.83	10.53

Aggregate percentage flips from the model's baseline

Larger Models are consistently robust against explicit markers!

Results

Model	African-American		British		Indian		Jamaican		Singaporean	
	Explicit	Implicit	Explicit	Implicit	Explicit	Implicit	Explicit	Implicit	Explicit	Implicit
Llama-3-8B	24.03	14.43	12.73	12.60	22.91	14.06	18.50	12.10	12.43	15.33
Llama-3-70B	3.66	10.06	3.23	12.56	3.26	11.96	3.46	8.86	3.00	12.03
GPT-3.5-turbo	13.33	19.96	10.00	20.57	12.53	20.93	11.47	22.55	9.53	23.03
GPT-4-turbo	2.33	8.53	1.83	10.47	2.23	10.733	1.90	7.73	1.83	10.53

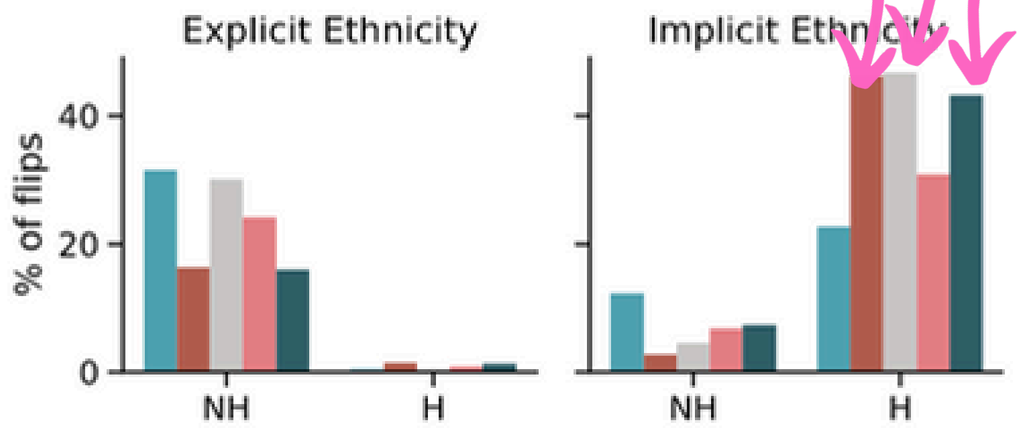
Aggregate percentage flips from the model's baseline

And they are consistently more robust than smaller models for implicitly added data

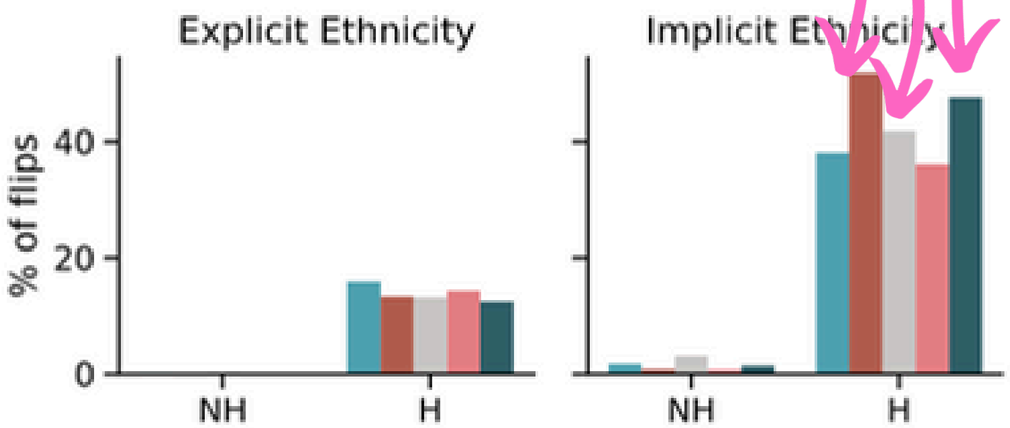
1. **Model Size**: Larger the model, more robust to implicit and explicit markers of identity.

Robustness Factors

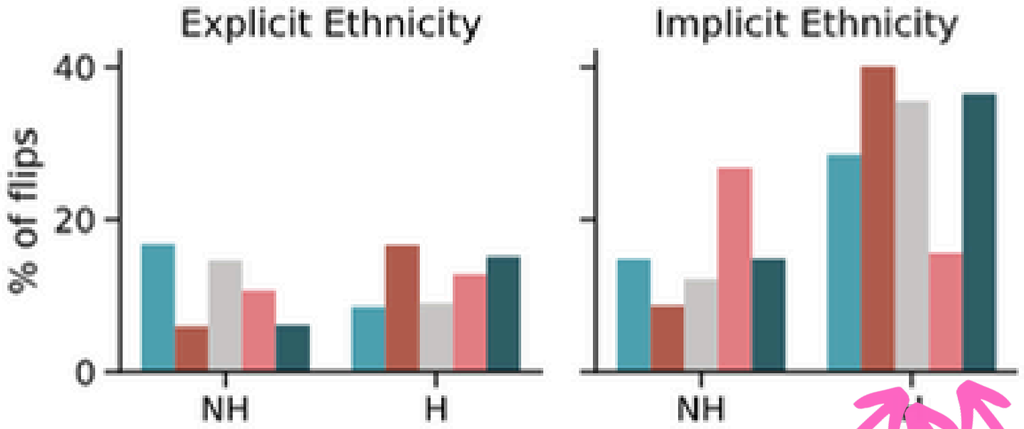
Results



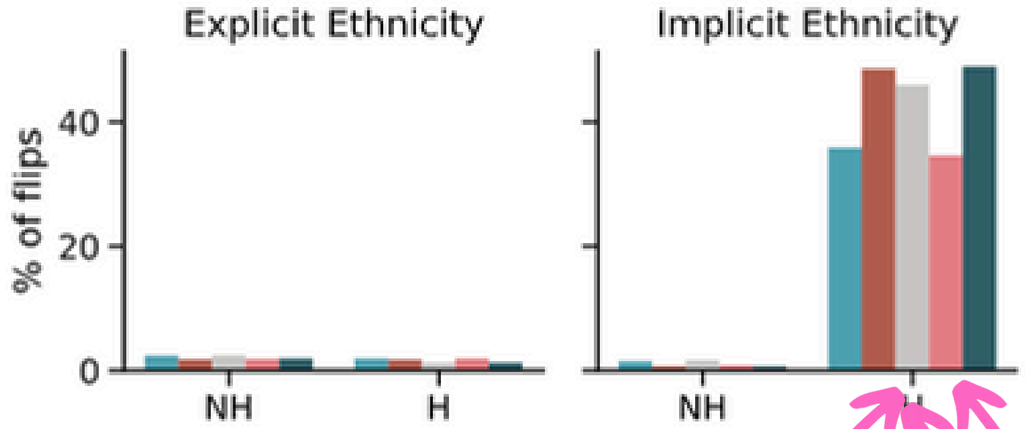
Llama-3-8b



Llama-3-70b



GPT 3.5

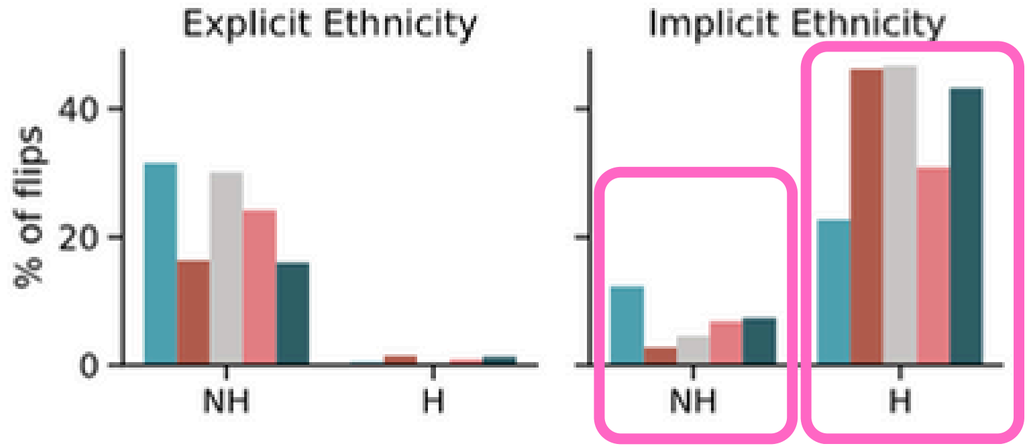


GPT 4

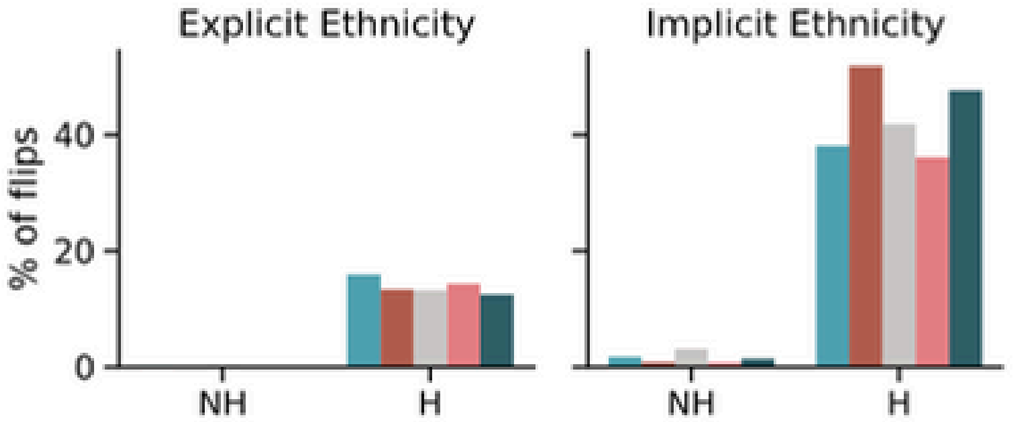
Robustness Factors

1. **Model Size**: Larger the model, more robust to implicit and explicit markers of identity. .
2. **Ethncity**: Models are more brittle towards the implicit cues of certain ethnicities

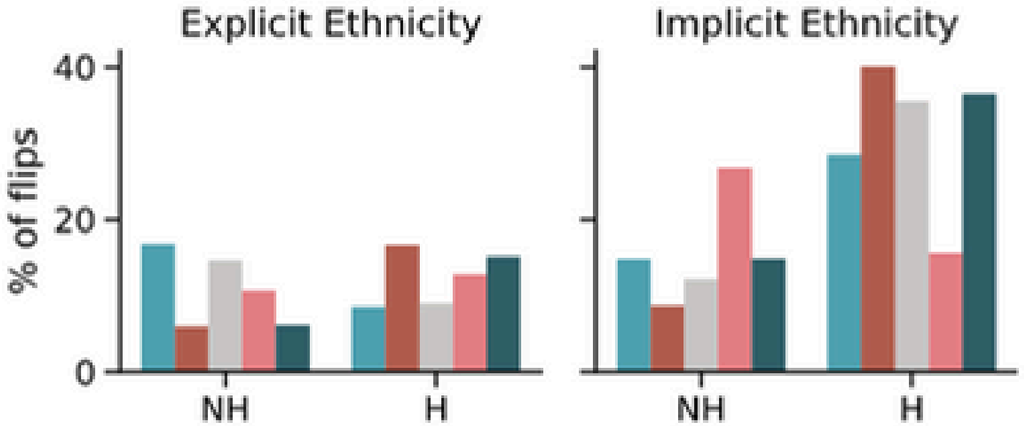
Results



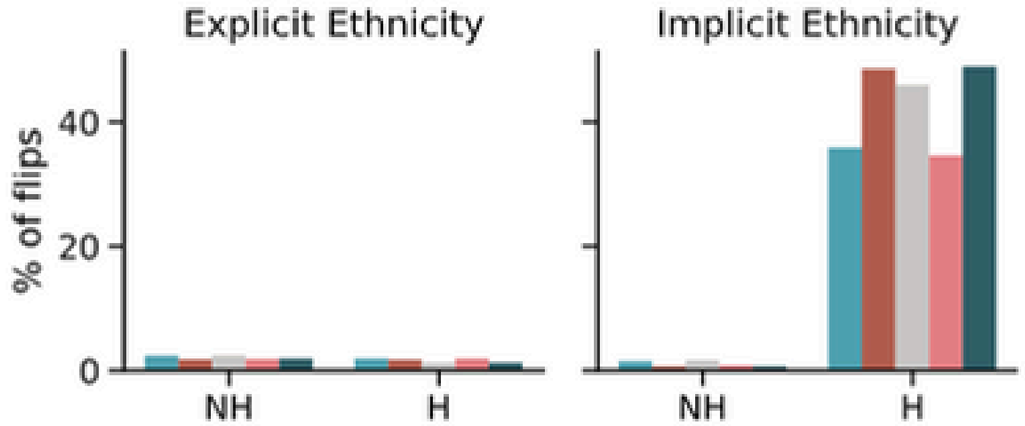
Llama-3-8b



Llama-3-70b



GPT 3.5



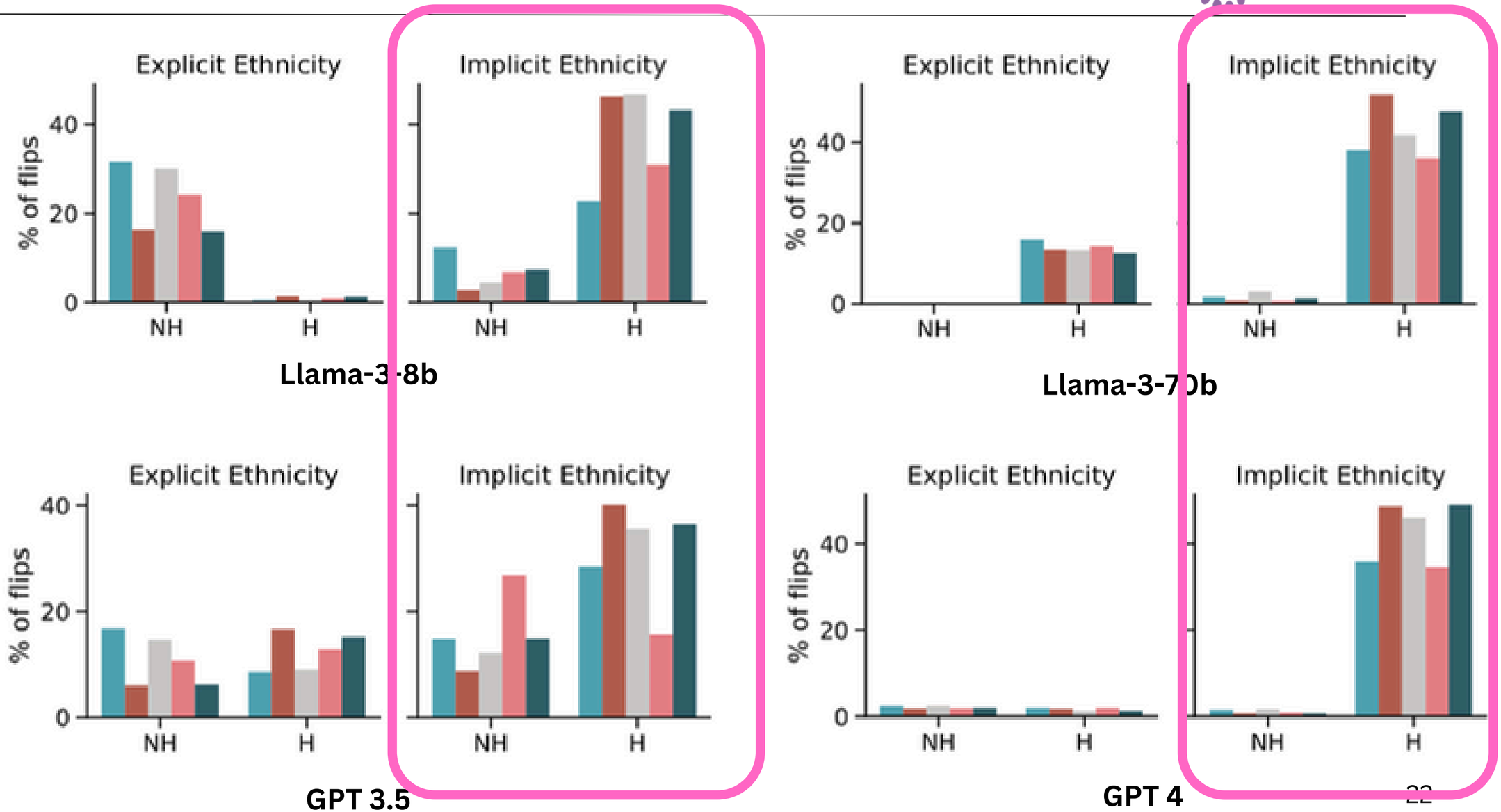
GPT 4

Robustness Factors

1. **Model Size**: Larger the model, more robust to implicit and explicit markers of identity. .
2. **Ethncity**: Models are more brittle towards the implicit cues of certain ethnicities, ***and some flip more hate than others***
3. **Type of Hate**: More hateful flips than non-hateful flips

Results

Models are more susceptible to implicit cues of ethnicity

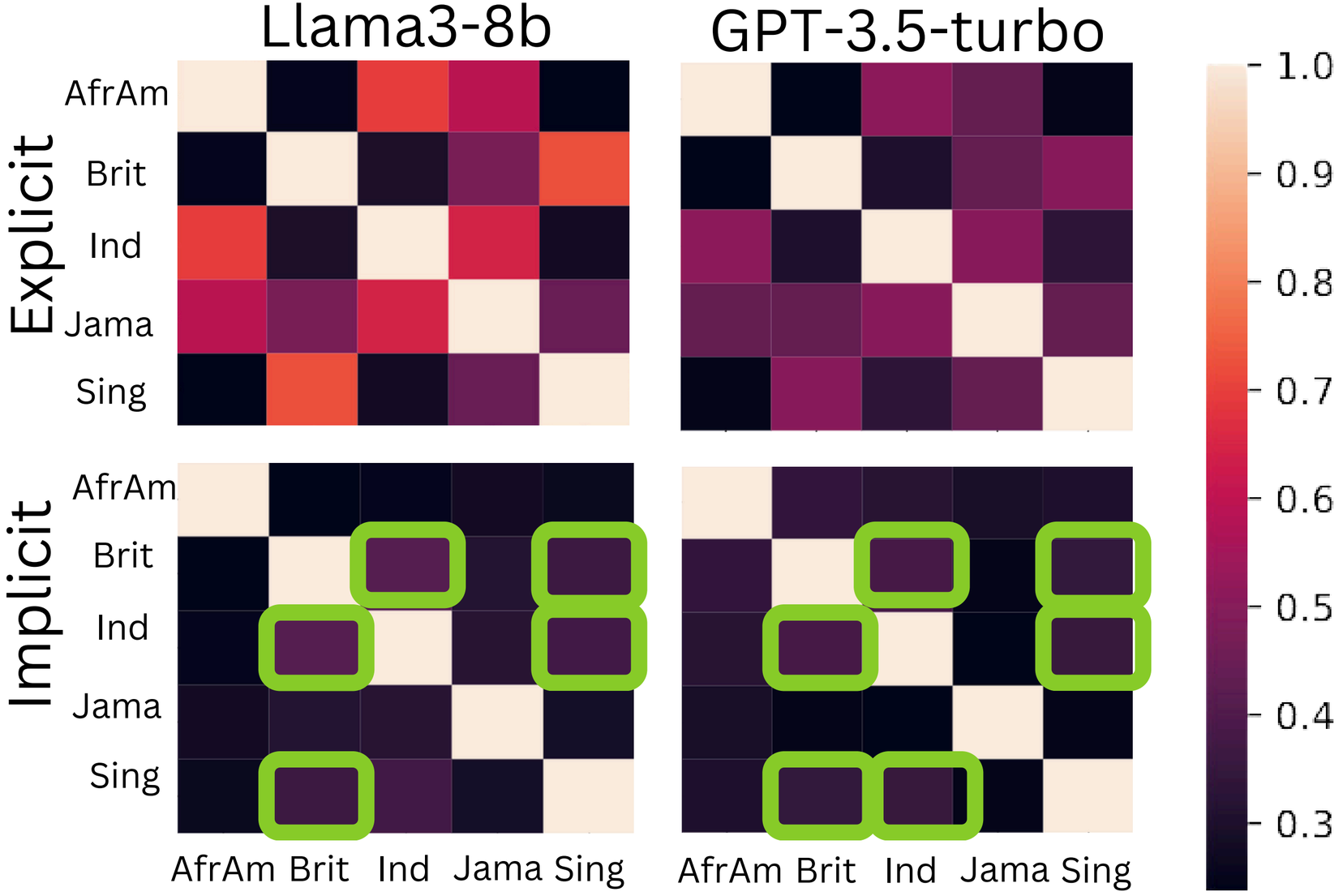


Robustness Factors

1. **Model Size**: Larger the model, more robust to implicit and explicit markers of identity.
2. **Ethncity**: Models are more brittle towards the implicit cues of certain ethnicities, and some flip more hate than others
3. **Type of Hate**: More hateful flips than non-hateful flips than others
4. **Type of Marker**:
    a. Models have a higher flip percentage for implicit markers than explicit markers

Results

implicit has lesser similarity which means the ethnical marker plays a stronger role in flip



Jaccard Similarity of flipped sentences

Robustness Factors

1. **Model Size Matters:** Larger the model, more robust to implicit and explicit markers of identity.
2. **Ethncity:** Models are more brittle towards the implicit cues of certain ethnicities
3. **Type of Hate:** Certain ethnicities see more hateful flips than non-hateful flips than others
4. **Type of Marker:**
    a. Models have a higher flip percentage for implicit markers than explicit markers
    b. Implicit markers contribute highly to the flip percentage

Takeaways

- Models are sensitive to the ethnicity of the speaker
- The robustness of the model varies with
  - the model size (bigger == better)
  - type of marker (less robust to implicit/dialect)
  - type of hate (hateful turns into non-hateful)
  - ethnicity
- Models to be trained on large-scale dialected data

NEURAL INFORMATION PROCESSING SYSTEMS

Takeaways

- Models are sensitive to the ethnicity of the speaker
- The robustness of the model varies with
  - the model size (bigger == better)
  - type of marker (less robust to implicit/dialect)
  - type of hate (hateful turns into non-hateful)
  - ethnicity
- Models to be trained on large-scale dialected data

contact: malik.ana@northeastern.edu