# All Humans are Humans, but Some Ethnicities are More Ethnic than Others

## Measuring model reporting bias with regards to marginalized ethnicities

Tom Södahl Bladsjö

January 5, 2024

## Reporting Bias

*"[...]the tendency of people to not state the obvious"*
*(Paik, Aroca-Ouellette, Roncone, & Kann, 2021)*

*The frequency with which people write about actions, outcomes, or properties is not a reflection of real-world frequencies or the degree to which a property is characteristic of a class of individuals.*
*(Chang, Ordonez, Mitchell, & Prabhakaran, 2019)*

Gricean Maxim of Quantity:

- Make your contribution as informative as is required (for the current purposes of the exchange).
- Do not make your contribution more informative than is required.

Gricean Maxim of Relevance:

- Be relevant.

(Grice, 1975)

Reporting bias can seriously impact what a model trained on text learns about the world (Gordon & Durme, 2013; Paik et al., 2021).

# A simple example

Bananas

Brown bananas

# A simple example

Bananas

Brown bananas

Reasonable. However, humans are not bananas

# Two examples from the Flickr8k dataset

A little girl in a pink dress going into a wooden cabin.

An Asian girl in a pink dress is smiling whilst out in the countryside.

All humans are human...

...but some humans are more normative than others

...attributes that are not the norm in a certain setting will be mentioned more often than attributes that are normative (and therefore too obvious to mention)

## Research questions

- **Are there between-group differences when it comes to the amount and type of information a model includes in descriptions of humans in images?**
- Based on these differences, **which group characteristic do models tend to consider "default" (that is, obvious and therefore unnecessary to mention)?**

Intuition:

- The model is generally going to be less likely to mention ethnicity than to not mention it
- ...but *how* unlikely it is to mention ethnicity will vary depending on the ethnicity in question
- Hypothesis: The model will be less likely to describe a white person as white than to describe a non-white person with their ethnicity

# The MMERB dataset[1]

- 1313 images
  - 680 images of non-white people (test group)
  - 633 images of white people (norm group)
- Each image has two contrasting captions

  One that mentions ethnicity...

  ...and one that does not

[1] https://github.com/TomBladsjo/LT-Resources-project/

Total of four test sets:

- test_mention (680 instances)
- test_no_mention (680 instances)
- norm_mention (633 instances)
- norm_no_mention (633 instances)

## Example

(a) An asian girl in a pink dress
(b) A girl in a pink dress

(a) A little white girl in a pink dress
(b) A little girl in a pink dress

## Experiment procedure

- Test an image captioning model on each of the four sets, keeping track of the order of the examples
- Compare model performance pairwise for the two captions on each image
- Compare these differences across groups

In detail:

- Measure perplexity on example sentences, normalized by sentence length (= exponentiated average cross entropy loss)
- For each group, subtract the vector of perplexities for sentences (b) from the corresponding vector for sentences (a) to obtain a vector of pairwise differences (how much more surprised was the model to see the sentence where ethnicity was mentioned)
- Perform Welch's t-test on the two groups of differences to see how much the means differ, and if the difference is significant

Models tested:

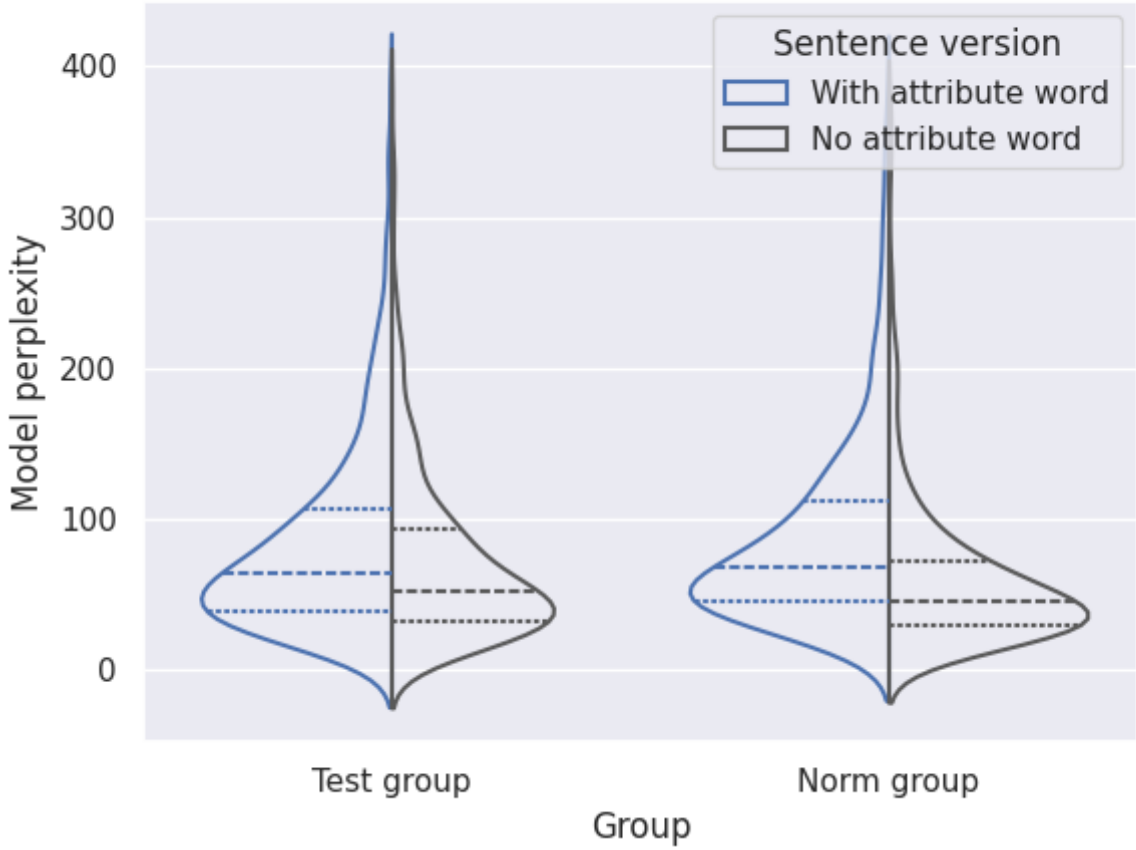
- BLIP for Conditional Generation[2] (Salesforce/blip-image-captioning-base)
- ViT+GPT2 Encoder-Decoder image captioning model[3] (nlpconnect/vit-gpt2-image-captioning)

[2] https://huggingface.co/docs/transformers/model_doc/blip

[3] https://huggingface.co/nlpconnect/vit-gpt2-image-captioning

# Results

# BLIP

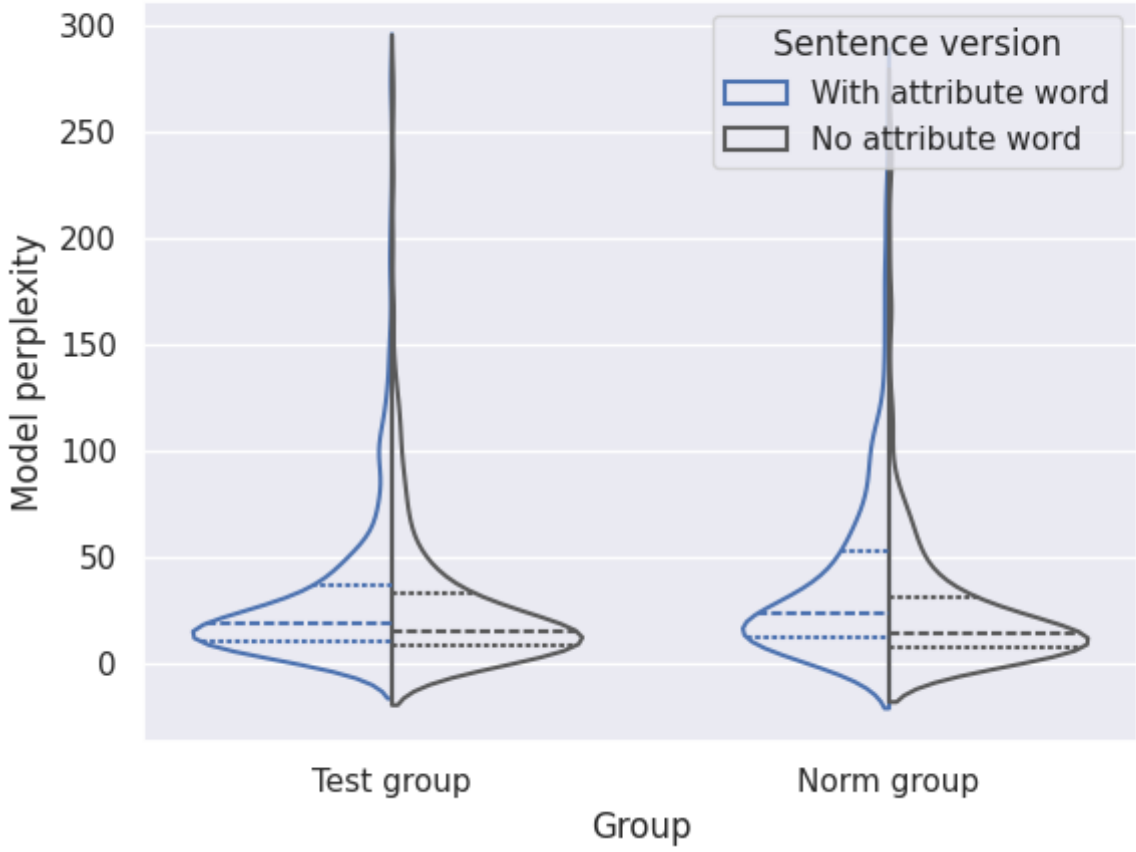



Difference in group means as measured by Welch's t-test: t-statistic = 11.4, p-value=1.13e-28

Tom Södahl Bladsjö



# ViT-GPT2




Difference in group means as measured by Welch's t-test: statistic=6.3, pvalue=3.23e-10

- Both models are generally more surprised to see the captions mentioning ethnicity than the ones that do not
- For both models, this surprise is greater for images depicting white people
- For both models, the difference is statistically significant
- For ViT-GPT2 the difference is smaller, suggesting that the model is slightly less biased

However:

- The models use different tokenization schemes and have different size vocabularies.
- This affects perplexity.

- The numbers may not be directly comparable,
- but they are reliable to measure differences within a single model.
- Both models display reporting bias with regards to marginalized ethnicities.
- This suggests that models consider whiteness to be default for people, and therefore not worth mentioning.

Possible future directions:

- Similar tests for other marginalized attributes (e.g. gender, body type, disability, queerness)
- Look into intersectional effects (is the model e.g. more likely to report ethnicity for Black women than for Black men?)
- Explore different methods and test other types of models (e.g. text/image matching models, masked language models etc)

# Thank you!

ppt template credit:
https://github.com/Urinx/LaTeX-PPT-Template

Chang, K.-W., Ordonez, V., Mitchell, M., & Prabhakaran, V. (2019). *Tutorial: Bias and fairness in natural language processing.* Recorded presentation at EMNLP 2019, hosted at UCLA NLP `http://web.cs.ucla.edu/~kwchang/talks/emnlp19-fairnlp/`.

Gordon, J., & Durme, B. (2013, 10). Reporting bias and knowledge acquisition. In (p. 25-30). doi: 10.1145/2509558.2509563

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (p. 41 - 58). Leiden, The Netherlands: Brill. Retrieved from `https://brill.com/view/book/edcoll/9789004368811/BP000003.xml` doi: 10.1163/9789004368811_003

Paik, C., Aroca-Ouellette, S., Roncone, A., & Kann, K. (2021, November). The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 823–835). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.63 doi: 10.18653/v1/2021.emnlp-main.63