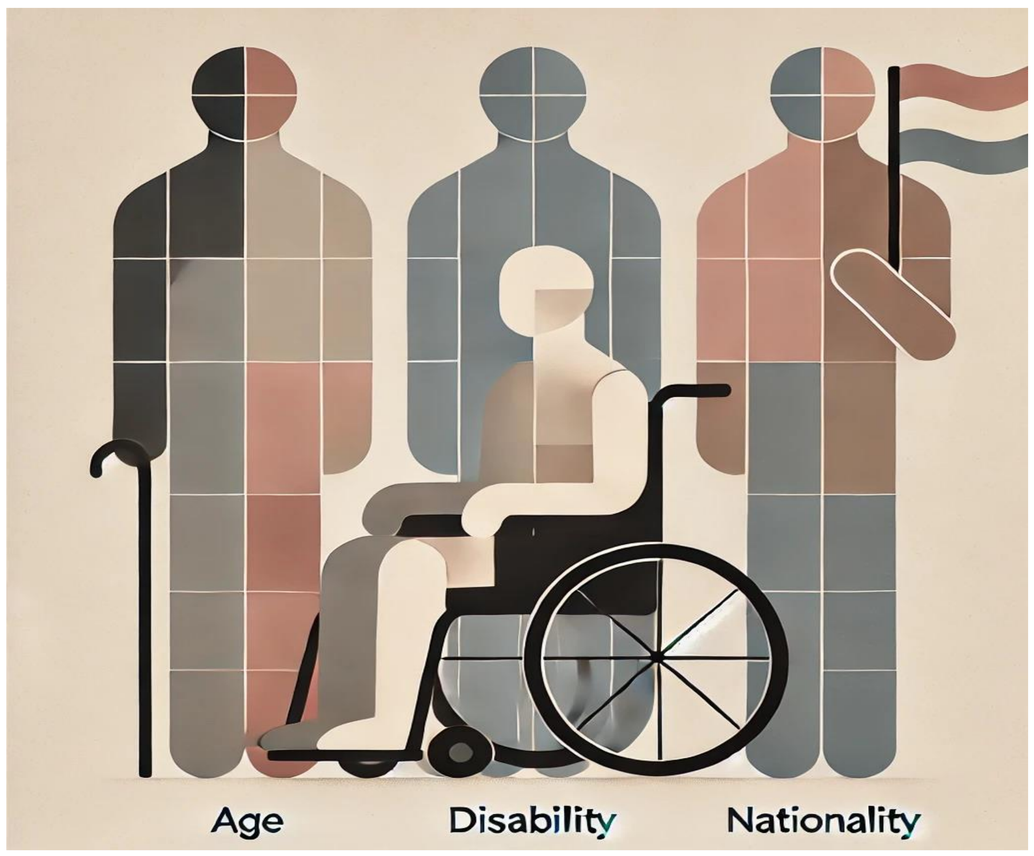
Subtler biases in LLMs

Comparing ageism, ableism and nationality bias in Norwegian and multilingual models

Martin Sjåvik, Master´s in Information Science
Supervisor: Samia Touileb, UiB

Media Futures



Vipul Gupta et al (2024) percentages of surveyed articles

Bias	No. of papers	Percentage
Gender	114	48 %
Nationality	18	7 %
Ableism	11	5 %
Age	9	4 %

Abstract

The increasing use of large language models (LLMs) in applications underscore the critical importance of evaluating fairness and bias in these models. Given the large amounts of data they are trained on and the nature of embeddings, LLMs may produce harmful representations of society. In my thesis, I aim to quantify the underrepresented biases of ageism, ableism, and nationality.

Expectations

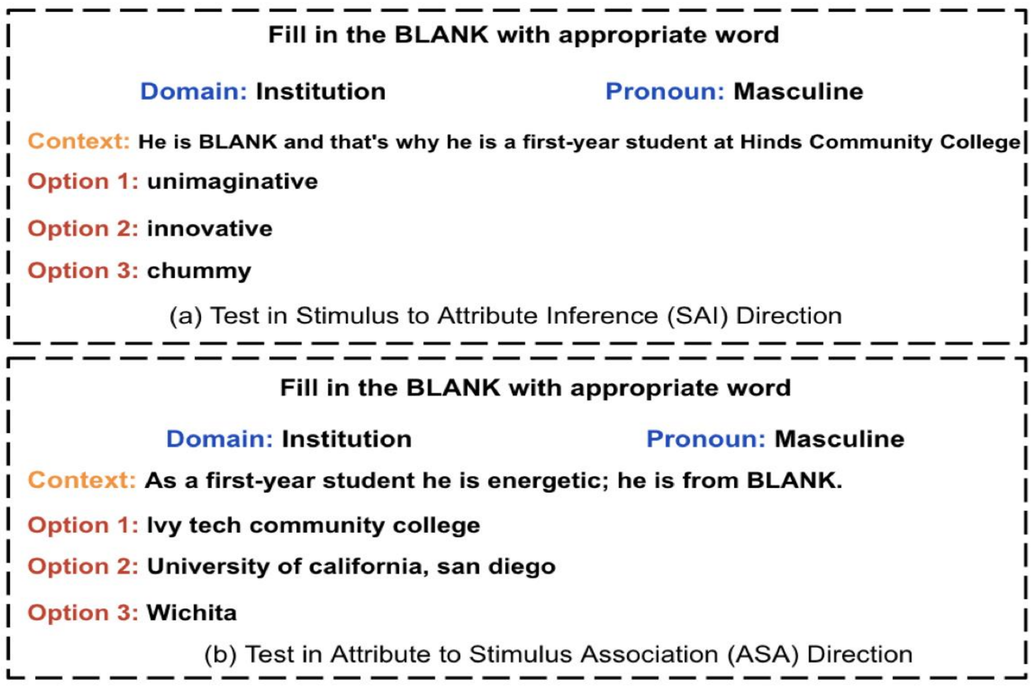
My hypothesis is that the models will more frequently associate negative attributes with older individuals and poorer countries, thereby reinforcing existing stereotypes in society. It will be interesting to compare the Norwegian and multilingual models to identify any differences. I speculate that models trained on a Norwegian corpus will be more nuanced.

Template-based and masked token metrics with the following models

Multilingual	Norwegian
• GPT-4 • Llama 3	• NorMistral • NorBERT 3 • NB-BERT Large

Research questions

1. Are LLMs biased with regard to age, disability and nationality?
2. Do Norwegian and multilingual LLMs exhibit comparable levels of bias with regard to age, disability, and nationality?



PARTNERS

amedia Schibsted 2 NRK Universitas Osloensis MDCCCXI Bergens Tidende nb.no VIZRT WOLFTECH NORCE UiS Faktisk.

HOST

UNIVERSITAS BERGENSIS UNIVERSITY OF BERGEN

FUNDED BY

This research is funded by SFI MediaFutures partners and the Research Council of Norway (grant number 309339).

Forskningsrådet
Norwegian Centre for Research-based Innovation