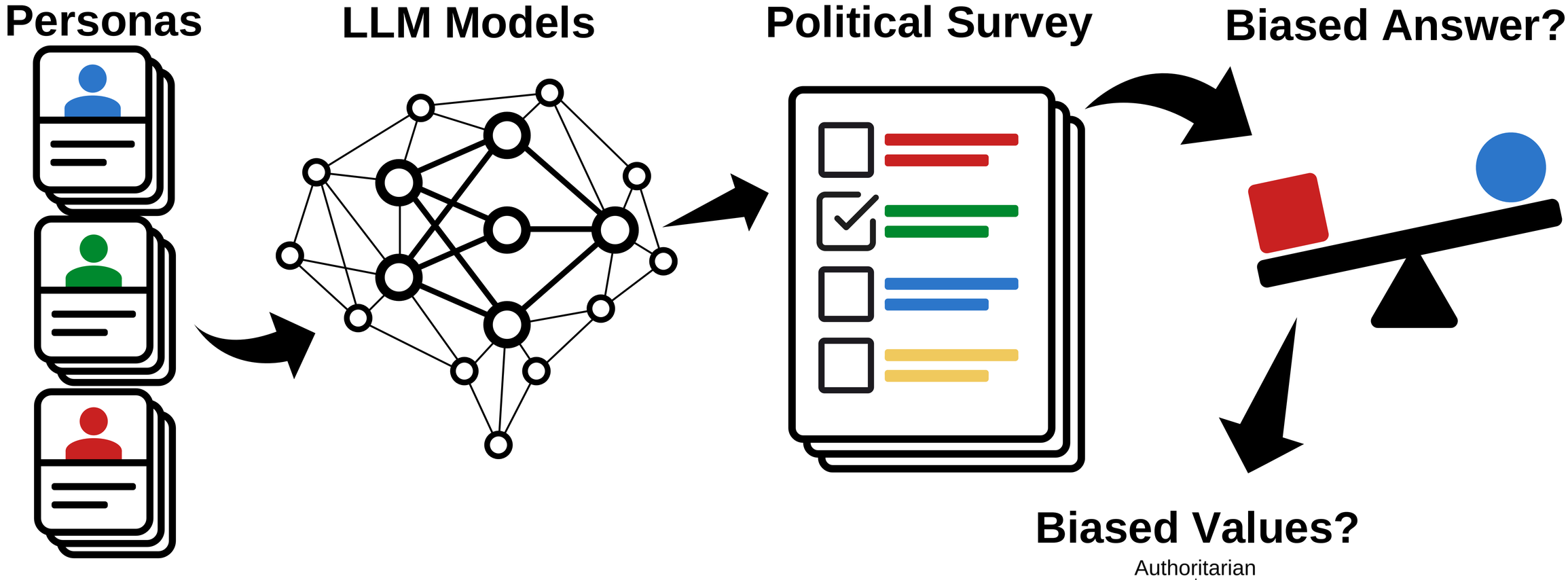
Investigating and Measuring Bias in Generative Language Models

Snorre Åldstedt, Master's in Information Science (UiB)
Supervisor: Samia Touileb (UiB)

Media Futures



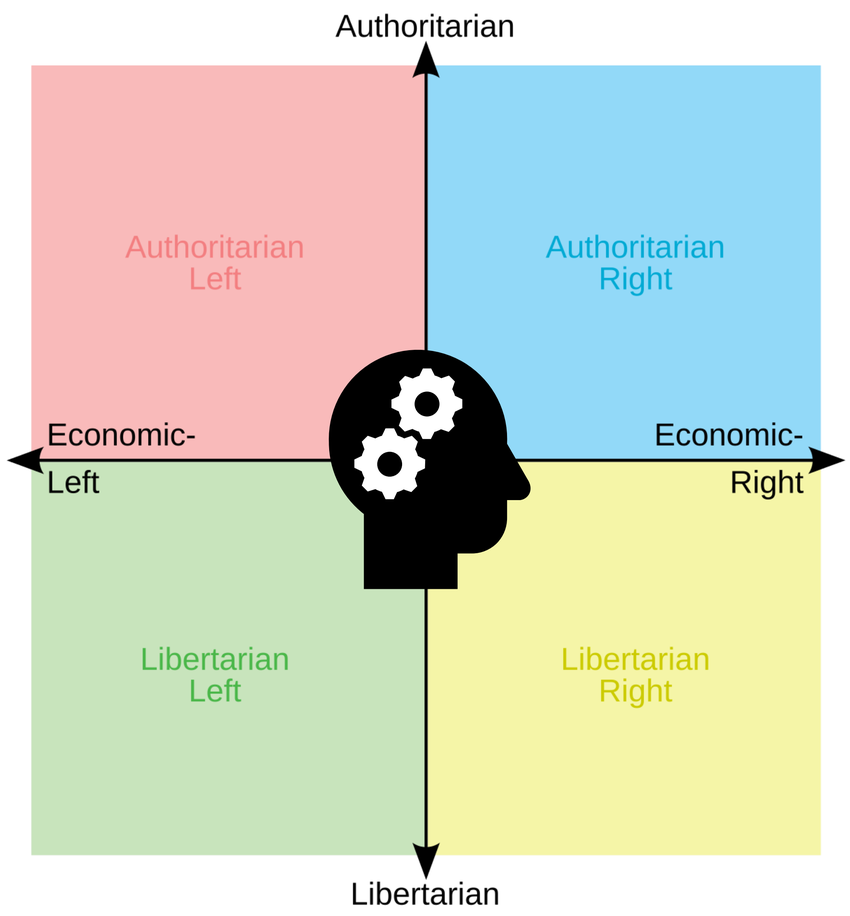
Abstract and Idea

LLMs are rapidly evolving and are integrated more and more with technology we use on a day-to-day basis. Just as we humans are biased, the generative language models we use might also be biased. The models often learn and amplify harmful bias. In this project we'll explore and try to detect these harmful social biases.

We are inspired by the work of Motoki et al., 2023 and Plaza-del-Arco et al., 2024 and will combine persona-based prompting and questionnaire to measure political ideologies.

Models
NorMistral
NorBloom
NorwAI-Mistral
NorwAI-Llama2
Viking
Llama-3

Table: Models used in this project
Blue: Norwegian, Green: Multilingual



Research Questions

1. Do Norwegian Language Models stereotype gendered-personas when it comes to political ideologies?
2. How do Norwegian Language Models compare to Multilingual-, and English models regarding political- and gender bias.

Expectations

English and multilingual models tend to be biased towards the left when it comes to political bias(Motoki et al., 2024). At the same time, they also generally stereotype genders (Plaza-del-Arco et al., 2024). Our hypothesis is that the Norwegian models will have the same tendencies as the English and the multilingual models. When mixed, we think that it'll stereotype women to the left and men to the right on the political spectrum

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