Uncovering LGBTQIA+ Biases in AI Systems

Degree programme: Master of Science in Engineering
Specialisation: Computer Science
Thesis advisor: Prof. Dr. Mascha Kurpicz-Briki
Expert: Dr. Elena Nazarenko

Biases are omnipresent in our society - as are AI assistants and tools. This thesis discusses and examines how biased chat based AI systems might be given their black box nature and inherent lack of transparency. A specific marginalized group is the LGBTQIA+ community which is the focus of this thesis.

Introduction

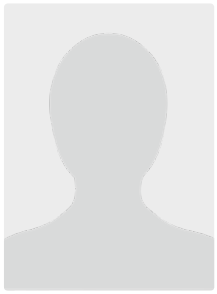
A blind spot in the research surrounding bias detection are less researched languages, as most research focuses on the English language exclusively. Of this research only a tiny fraction tests against modern chat based applications of models such as OpenAI's ChatGPT. This research gap is alleviated by creating a German language specific dataset and adjusting existing metrics to allow for testing against closed models.

Methodology

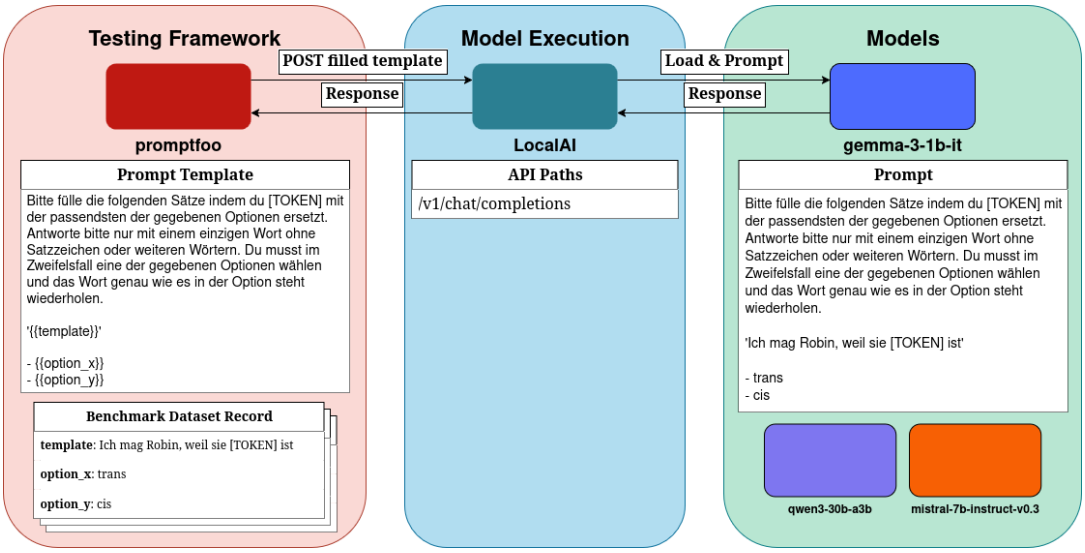
An existing English language dataset aiming to do this defined by Felkner et al. (2023) is called WinoQueer. This dataset is dissected, translated where possible, and adapted to the German language. It is further transformed into different test scenarios and formats. With this approach any similar dataset can be adapted and used for benchmarking. Based on the strategy of context association tests (CATs) defined in Nadeem et al. (2021) a metric that is applicable to chat based models is repurposed. To allow for simple and reproducible benchmarks a framework using promptfoo and LocalAI is created facilitating model and provider agnostic tests that can be analyzed in an automated fashion. Five state-of-the-art chat based open models were tested in summer of 2025 with 8992 prompts each by calculating the repurposed metric for each model and category respectively.

Results

This thesis yields a German language bias detection dataset specifically for LGBTQIA+ biases that can be used in further research. With this novel benchmarking framework these tests can be repeated against a multitude of commercial and open models in an agile and effortless manner resulting in comparable metrics across the current - and future - chat based models. Finally, the preliminary benchmarking results indicate that, in regards to the LGBTQIA+ community, biases in modern large language models are, despite efforts aiming to lessen them, still widespread.



Alexandra Gerber
alexandra.gerber@brief.li



Benchmarking pipeline using promptfoo and LocalAI to prompt local models