

Carolina Haensch (LMU Munich)

Assessing bias in LLM-generated synthetic datasets: examining LLM personas in German and European elections

The rise of large language models (LLMs) like the GPT-3 and 4 models has sparked interest in their potential for creating synthetic datasets, particularly in the realm of privacy research. This study critically evaluates the use of LLMs in generating synthetic public opinion data, pointing out the biases inherent in the data generation process. While LLMs, trained on vast internet datasets, can mimic societal attitudes and behaviors, their application in synthesizing data poses significant privacy and accuracy challenges. We investigate these issues using the case of vote choice prediction in the 2017 German federal elections and the 2024 European election.