HUMAN PREFERENCES IN LARGE LANGUAGE MODEL LATENT SPACE: A TECHNICAL ANALYSIS ON THE RELIABILITY OF SYNTHETIC DATA IN VOTING OUTCOME PREDICTION

Sarah Ball*, Simeon Allmendinger*, Frauke Kreuter, and Niklas Kühl

Problem Statement

There is an increasing usage of Generative AI models in simulating human preferences. Previous research focuses on comparing LLM predictions based on personas to a gold standard survey prediction for these personas. However, **we lack a deeper understanding** of how **"opinion formation"** works on a **technical level in LLMs** and **how reliable** the resulting **synthetic data** is **for answering human-related questions** of interest. We test this for persona-to-party mappings in the German multi-party context.

Research Questions

RQ1: How well does LLM-generated synthetic data mimic the distribution of human answers in survey-like questions for different demographic subgroups in their latent space?

RQ2: How is prompt instability reflected in the models' latent space?

Data & Models

For Germany we use the German Longitudinal Election Study (GLES) for the year 2021. We construct personas in the following form:

> *I am* {age} *years old and* {gender}*. I have* {education}*, a* {hhincome} *household net income per month, and I am* {employment}*. Ideologically, I lean towards the position* {leaning}*. I live in* {part of germany}*. If the elections were held in* {year of election}*, which party would I vote for? I vote for the party ...*

For constructing the latent space probe, we use the German Wahl-o-Mat data. The Wahl-o-Mat is an online questionnaire, which consists of short political statements based on party manifestos to which interested citizens can give their agreement (strong agree to strong disagree). For all short political statements that users see, each party provides an opinion to give more context to the question of interest. We extract this opinion for each Wahl-o-Mat item for German and European elections from 01/2021 until 12/2024.

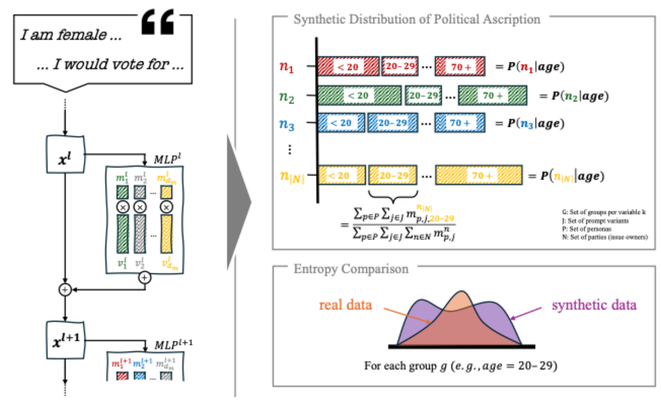
Table 1: Overview of LLM models used in the experiments.

Family	Size	Model	Reference
Llama 3.2	3B	Llama-3.2-3B-Instruct	MetaAI [2024a]
	3B	Llama-3.2-3B	MetaAI [2024a]
Llama 3.1	8B	Llama-3.1-8B-Instruct	MetaAI [2024b]
	8B	Llama-3.1-8B	MetaAI [2024b]
Llama 3	8B	Llama-3-8B-Instruct	MetaAI [2024c]
Llama 2	7B	Llama-2-7b-hf	Touvron et al. [2023]
	7B	Llama-2-7b-chat-hf	Touvron et al. [2023]
Mistral	7B	Mistral-7B-v0.1	Jiang et al. [2023]
	7B	Mistral-7B-Instruct-v0.1	Jiang et al. [2023]
Gemma	7B	Gemma-7b-it	Google [2024]
	7B	Gemma-7b	Google [2024]
Qwen	7B	Qwen2.5-7B	Yang et al. [2025]
	7B	Qwen2.5-7B-Instruct	Yang et al. [2025]

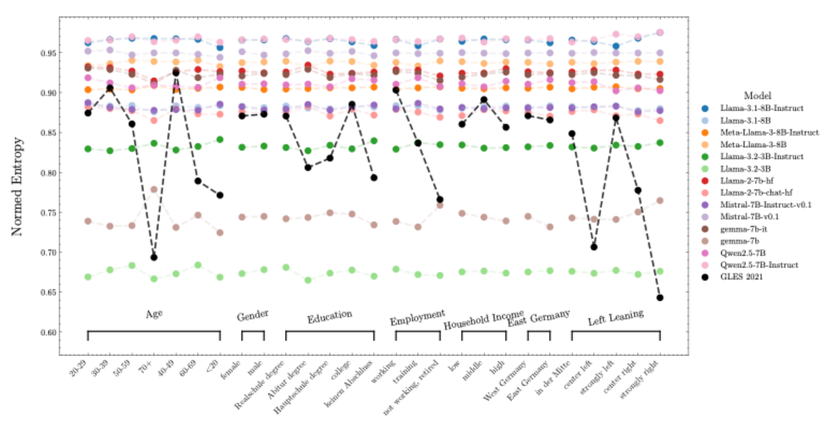
* Equal contribution, correspondence to: sarah.ball@stat.uni-muenchen.de, and simeon.allmendinger@fit.fraunhofer.de (for more, scan QR code)

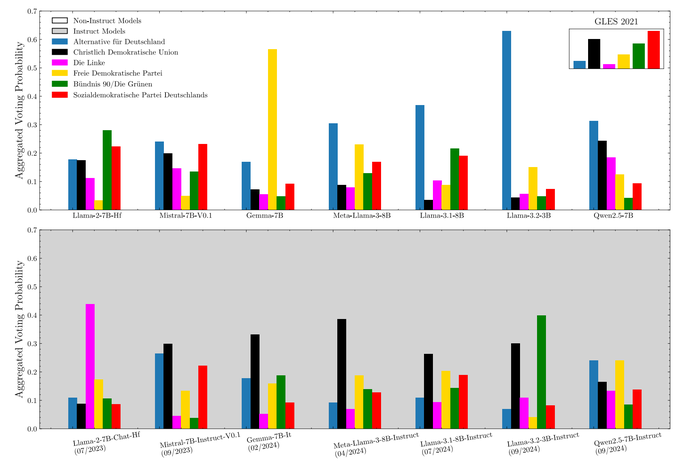
Methodology

Synthetic Respondents: Created synthetic personas and prompted LLMs to give vote choice for this persona.
Latent Space Probing: Trained Multi-Layer Perceptrons (MLPs) probes to identify model internals that are activated for specific parties and their associations.
Mapping analysis: Analyzed how different persona attributes activate the identified model internals and investigate how stable this mapping is in comparison to real world data.
Prompt Sensitivity Tests: Measured the stability of model outputs by slightly varying the prompts, while preserving meaning.



Results





LLMs fail to capture human variability: Synthetic data generated by LLMs lacks the variance and entropy seen in real survey responses, particularly across demographic subgroups.

Low differentiation in political mapping: Persona-to-party mappings in LLM latent spaces show limited differentiation, meaning LLMs struggle to anchor specific personas to specific political outcomes.

Prompt sensitivity undermines reliability: Minor, meaning-preserving variations in prompts can significantly change the model's output. This instability varies across models.

Model behavior differs by alignment: Base models (unaligned) tend to favor right-wing populist parties (e.g., AfD), while aligned models shift toward center-right or center-left parties.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN · relAI · mcml Munich Center for Machine Learning · Fraunhofer FIT · SCAN ME!