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

Sarah Ball[*1,5], Simeon Allmendinger[*2,4], Frauke Kreuter[1,3,5], and Niklas Kühl[2,4]

[1]Ludwig-Maximilian-University Munich
[2]University of Bayreuth
[3]University of Maryland
[4]Fraunhofer Institute for Applied Information Technology FIT
[5]Munich Center for Machine Learning MCML

ABSTRACT

Generative AI (GenAI) is increasingly used in survey contexts to simulate human preferences. While many research endeavors evaluate the quality of synthetic GenAI data by comparing model-generated responses to gold-standard survey results, fundamental questions about the validity and reliability of using LLMs as substitutes for human respondents remain. Our study provides a technical analysis of how demographic attributes and prompt variations influence latent opinion mappings in large language models (LLMs) and evaluates their suitability for survey-based predictions. Using 14 different models, we find that LLM-generated data fails to replicate the variance observed in real-world human responses, particularly across demographic subgroups. In the political space, persona-to-party mappings exhibit limited differentiation, resulting in synthetic data that lacks the nuanced distribution of opinions found in survey data. Moreover, we show that prompt sensitivity can significantly alter outputs for some models, further undermining the stability and predictiveness of LLM-based simulations. As a key contribution, we adapt a probe-based methodology that reveals how LLMs encode political affiliations in their latent space, exposing the systematic distortions introduced by these models. Our findings highlight critical limitations in AI-generated survey data, urging caution in its use for public opinion research, social science experimentation, and computational behavioral modeling.

1 Introduction

With the release of ChatGPT in November 2022, the world has seen a spike in interest in large language models (LLMs). Many academic disciplines, as well as the business world, wonder if and how they can integrate LLMs to their benefit. One emerging—and highly debated—topic is the usage of LLMs for (public) opinion research. The idea is that one can leverage LLMs to substitute for surveying humans. Yet, the question remains as to how valid and reliable it is to substitute humans with LLMs. Previous research mainly focuses on comparing LLM predictions based on personas to a gold standard survey prediction for these personas. The results of such analyses are mixed [Argyle et al., 2023, Kim and Lee, 2023], revealing various problems, e.g., prediction instability that occurs with slight formulation changes in the prompt [Bisbee et al., 2023] and performance differences across national and linguistic contexts [von der Heyde et al., 2024a]. While such approaches might give first insights into how well LLMs can predict general questions of interest, 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. Based on this, our article addresses two central research questions:

---

[*]Equal contribution.

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?

To address these questions, we focus on the use case of predicting election outcomes with LLMs in the German multi-party context. The election context is chosen not only for its societal relevance and its popularity as a testbed in recent research on LLM-human substitutability [Argyle et al., 2023, von der Heyde et al., 2024a, Yu et al., 2024], but also because elections are commonly used for evaluating the quality of survey data across different vendors or data collection modes, providing a rare benchmarking opportunity in survey research. We further choose the German multi-party context as it allows for multiple party comparisons, increasing the robustness of our results.

In our experiments, we analyze the latent space of LLMs, focusing on mechanistically understanding persona-to-party mappings. To do so, we develop a probe-based methodology to systematically identify model-specific value vectors—Multi-Layer Perceptrons (MLPs)—associated with political affiliations. This allows us to examine how demographic attributes—such as age, gender, and ideological leaning—interact with latent political structures within LLMs. Our results reveal that **LLMs fail to replicate the entropy observed in real-world survey data**, as their persona-to-party mappings exhibit low differentiation across demographic subgroups. We further explore prompt sensitivity by first replicating previous findings that small meaning-preserving variations in persona descriptions can alter voting predictions, underscoring the instability of LLM-generated survey data. Next, we demonstrate that, **for certain models, higher entropy in the persona-to-party mapping correlates with increased prompt sensitivity**. However, we also observe the opposite relationship in other models.

Overall, our study provides a technical foundation to assess the usability and reliability of synthetic LLM data, exposing fundamental limitations that practitioners must address before relying on LLMs for public opinion research, social science experimentation, and computational behavioral modeling. We preregistered our study on the Open Science Framework[1], and the code is available at GitHub[2].

2   Related Literature

Using LLMs as substitutes for humans. The advent of large language models (LLMs) has sparked significant interest regarding their potential to serve as substitutes for human respondents [Argyle et al., 2023]. This question is especially relevant for survey researchers in the social sciences, who are investigating whether responses generated by LLMs can reliably resemble those provided by humans in surveys [Argyle et al., 2023, Bisbee et al., 2023, Dominguez-Olmedo et al., 2025, Park et al., 2024, Qu and Wang, 2024, von der Heyde et al., 2024b, Wang et al., 2024]. Similar inquiries have emerged in fields such as market research [Brand et al., 2023, Sarstedt et al., 2024], annotation tasks [Törnberg, 2023, Ziems et al., 2024], experiments in psychology and economics [Aher et al., 2023, Xie et al., 2024], and human-computer-interaction [Hämäläinen et al., 2023, Törnberg, 2023], among others. The findings from these investigations are mixed. Some studies suggest that LLMs can reasonably approximate the average outcomes of human surveys [Argyle et al., 2023, Bisbee et al., 2023, Hämäläinen et al., 2023, Törnberg, 2023, Brand et al., 2023, Xie et al., 2024], while others highlight significant limitations, particularly in their inability to accurately represent the opinions of diverse demographic groups [Santurkar et al., 2023, von der Heyde et al., 2024b, Sarstedt et al., 2024, Qu and Wang, 2024, Dominguez-Olmedo et al., 2025]. However, a common limitation across these studies is their focus on surface-level comparisons, i.e., matching LLM output to human survey responses without delving into the underlying mechanisms of how opinions are encoded and represented within the models' latent spaces. We address this gap by studying how personas are mapped to opinions as well as what the inherent limitations are in eliciting specific knowledge from these models.

Prompt sensitivity. By systematically introducing subtle changes to the prompt format, previous studies have shown that LLM output is highly sensitive to prompt changes, thereby influencing downstream evaluations [Leidinger et al., 2023, Mizrahi et al., 2023, Chatterjee et al., 2024, Voronov et al., 2024, Zhuo et al., 2024]. Articles most closely related to our study are Sclar et al. [2023] and Zhu et al. [2023], which focus on explaining LLM prompt sensitivity next to establishing that it exists. Sclar et al. [2023] analyze how changes in the formatting of the prompt without semantic changes lead to large performance differences. They further show that prompt embeddings of different but equivalent formats are distinguishable using a trained classifier, implying that prompt formats transform the output probability distribution, yielding different predictions. Zhu et al. [2023] design "attacks" on the character, word, sentence, and semantic level to mimic user errors. They again find significant performance differences induced by the subtle prompt changes. The study further examines why LLMs are vulnerable to adversarial inputs by analyzing their attention weights

---

[1]Due to time constraints, we reduced the number of parameters to consider in our study for the preprint at hand (OSF).

[2]Codebase in GitHub Repository.

when processing both clean and adversarial prompts. The findings indicate that these adversarial prompts redirect the model's attention towards the altered elements, leading to incorrect responses. We build on these interpretability approaches and offer a different perspective of how persona-to-party mapping entropy is related to prompt sensitivity.

3 Models and Data

Model selection. For our experiments we use both base and aligned models of different model families and sizes, see Table 1. These models are developed by teams across the world and fulfill the white-box criteria, which is a requirement for studying their latent space.

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]
	8B	Llama-3-8B	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]

Real world comparison. In order to compare our model predictions to real data, we use the German Longitudinal Election Study (GLES) [GESIS – Leibniz Institute for the Social Sciences, 2024]. This representative survey captures insights about German citizens' political attitudes, preferences and behaviours and is a widely used gold standard [Schmitt-Beck et al., 2010]. As a baseline, we choose the cross-sectional survey of the year 2021[3], for which the GLES asks about voting decisions in the respective federal elections and also captures our variables of interest. To obtain a representative comparison baseline for our LLMs we weight the data with a socio-demographic weight that aligns the distributions to the marginal distributions of the 2021 Microcensus.

Personas. For the construction of the personas, we follow previous literature [von der Heyde et al., 2024a] by combining political science theory for identifying voting predictors and representative surveys for extracting plausible values for these predictors. Hence, our personas are both theory- and data-driven. Concretely, we select the variables `age`, `gender`, `education`, `hhincome`, `employment`, `political orientation`, and whether a person `lives` in East or West Germany and combine them in a prompt[4]. We vary the values for the different variables while holding the prompt structure fix. Furthermore, to account for LLM models' prompt sensitivity, we paraphrase the prompts. Thus, an example persona instantiated via an LLM prompt reads as follows:

> *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 {left leaning}. I live in {east germany}. If the elections were held in {year of election}, which party would I vote for? I vote for the party ...*

Probe generation. To train our probe, we use the German "Wahl-o-Mat" data [Bundeszentrale für politische Bildung, 2025]. 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). Based on the user's answers, the tool provides a voting recommendation. 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. The parties of interest are, in alphabetical order, the Alternative für

---

[3]In future iterations of this manuscript, we will repeat the comparison with 2025 data.
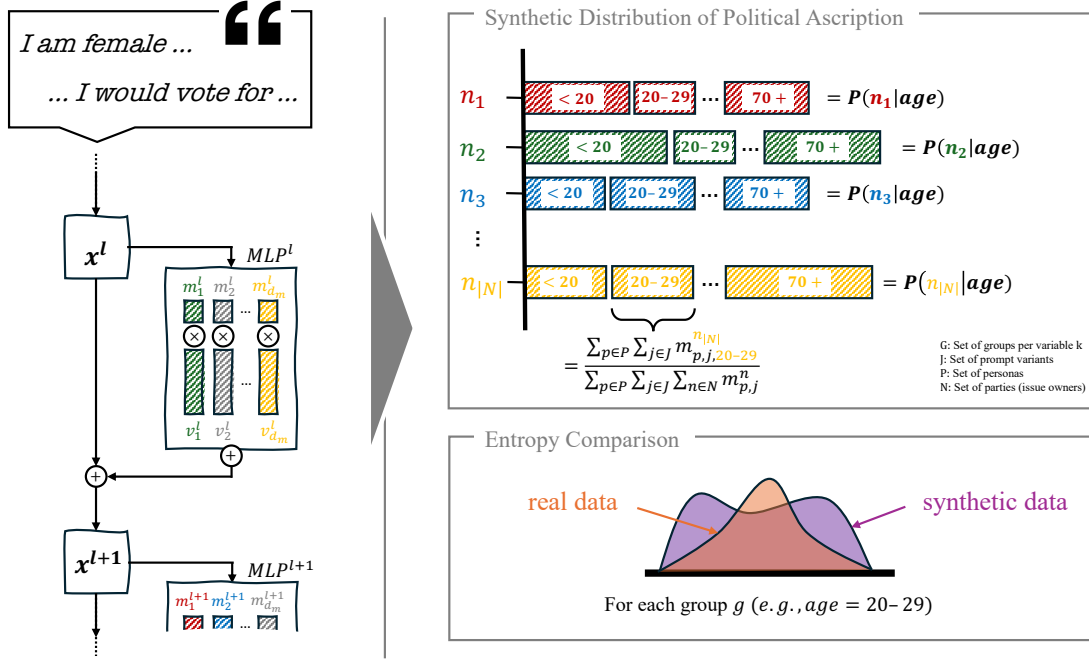[4]See Appendix–Table 2 on details for the values of the specific variables.

Figure 1: Method overview for comparing latent space persona-to-party mappings with real world voting distributions.

Deutschland (AfD), Christlich Demokratische Union (CDU), Freie Demokratische Partei (FDP), Sozialdemokratische Partei Deutschland (SPD), Bündnis 90/Die Grünen (GRÜNE), and DIE LINKE.

4 Methodology

Understanding how LLMs encode and generate synthetic survey responses necessitates to investigate persona-to-party mappings within the models' latent space. Building upon prior research, our methodology integrates trained probes to systematically identify model-specific representations of political ascriptions, thereby offering insights into the underlying value vectors. As depicted in Figure 1, our methodology depicts how LLM architectures encode voting preferences compared to historical human preferences from GLES.

4.1 Technical Preliminaries

Each transformer model [Vaswani et al., 2023] consists of transformer blocks in which multihead-attention (MHA) and multilayer perceptrons (MLP) update the residual stream representation (x_i^l) in each layer l to obtain an updated representation x_i^{l+1} (bias terms omitted for brevity) [Elhage et al., 2021]:

$$x_i^{l+1} = x_i^l + MLP^l(x_i^l + MHA^l(x_i^l)), l = 1, 2, ..., L \tag{1}$$

Based on [Geva et al., 2022] we can further decompose each MLP into two linear transformations (note that we write $x_i^l = x^l$ for brevity):

$$MLP^l(x^l) = f(W_K^l x^l)W_V^l, \tag{2}$$

where f is a non-linear activation function and $W_K^l, W_V^l \in \mathbb{R}^{d_{mlp} \times d}$. Hence, each value vector v_i^l in column i of W_V^l is weighted by a vector of coefficients $m^l := f(W_K^l x_i^l) \in \mathbb{R}^{d_m}$. Noting k_i^l as the key vector of row i in W_K^l, one can write:

$$MLP^l(x_i^l) = \sum_{i=1}^{d_m} f(x_i^l k_i^l)v_i^l = \sum_{i=1}^{d_m} m_i^l v_i^l \tag{3}$$

From this equation, [Geva et al., 2022] interpret an MLP update to the residual stream as sub-updates, consisting of weighted value-vectors. They further show that in each sub-update, v_i^l either de- or increases the probability of a token t to be generated:

$$p(t|x^l + m_i^l v_i^l, E) \propto exp(e_t \cdot x^l) \cdot exp(e_t \cdot m_i^l v_i^l), \tag{4}$$

where e_t is the embedding of token t and E the embedding matrix used to generate the first embedding of this token. Importantly, when $e_t \cdot m_i^l v_i^l < 0$, the probability of t decreases and vise versa if $e_t \cdot m_i^l v_i^l > 0$. Furthermore, $e_t \cdot v_i^l$ is static and does not depend on the input, which is why the impact of the input is determined by the scaling m_i^l, which itself is determined by the key vector, k_i^l and the residual stream representation x^l. Given this decomposition for our case at hand, we are first interested in identifying these "static" value vectors, which most increase the likelihood of outputting a token related to a party. We then analyze the scaling m_i^l of these identified value vectors, induced by the personas, which are our inputs of interest.

4.2 Constructing a probe for identifying party-related MLP value vectors

We aim to extract value vectors from the intermediate layers of LLMs, as these layers capture conceptual structures and high-level semantic representations more effectively than final layers, which are predominantly specialized for next-token prediction [Panickssery et al., 2023]. By focusing on these layers, we seek to uncover how specific residual stream patterns correlate with political biases and party affiliation in LLMs. To achieve this, we train linear probes that predict the party on the basis of the residual stream $\bar{x}^l$ of layer l. Similar to [Lee et al., 2024], these probes help identify value vectors that promote tokens linked to specific partys. The probes are trained as binary classifiers, distinguishing between residual streams corresponding to a specific party n ($y = 1$) and all others ($y = 0$). The training loss is weighted to mitigate class imbalance:

$$\mathcal{L} = -w_1 y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) \tag{5}$$

where w_1 represents the weight for the positive class. The probe function follows:

$$P(n|\bar{x}^l) = \text{softmax}(W_n \bar{x}^l), \quad W_n \in \mathbb{R}^d_m, \quad l \in [0.6L, 0.9L], \tag{6}$$

where W_n represents the learned parameters, n symbolizes a party from $N = \{n_1, n_2, \dots\}$, and $\bar{x}^l$ is the mean residual stream of a selected layer l from the interval $[0.6L, 0.9L]$. The model consists of a linear layer with dropout, optimized using Adam with cosine annealing. The training data consists of opinion statements. Each statement is paired with the corresponding party opinion to construct prompts. The language model aims to predict the party from the statement-opinion pairs, and residual streams are recorded at all layers and token positions ($=$ sequence). After training, we extract value vectors per party by identifying MLP weights most aligned with the trained probe weights. These vectors are selected based on cosine similarity, where W_{probe} represents the trained probe weights and $v_{i,l}$ denotes the value vector:

$$\cos(\theta_i^l) = \frac{W_{\text{probe}} \cdot v_i^l}{\|W_{\text{probe}}\| \|v_i^l\|}, \quad i \in [1, d_m], \quad l \in [1, L]. \tag{7}$$

We define the set of top 20 value vectors per party n as:

$$\hat{V}^n = \{v_i^l \mid \cos(\theta_i^l) \text{ is among the top 20 for } i \in [1, d_m], \text{ and } l \in [1, L]\}. \tag{8}$$

The selection criterion ensures that only the 20 most aligned value vectors are retained per layer, as we observe a drop in cosine similarity beyond this threshold.

4.3 Analyzing the mapping between personas and the identified party-related value vectors

Personas are defined by key attributes (such as age, gender, and political attitudes) with systematically varied values and paraphrased prompt variants to ensure robustness. These controlled inputs allow us to analyze how different demographic and ideological configurations affect model residual streams in response to political prompts. To investigate the interaction between personas and the identified value vectors v_i^l, we compute their residual stream of responses to persona prompts. Specifically, we measure the contribution of scaling factors $m_p \in \mathbb{R}^N$ (cf. eq. (4)) by evaluating the residual stream across all layers for each value vector v_i^l. The scaling factor of each persona prompt is computed as:

$$m_p = \sum_{i,l} m_i^l \cdot \frac{\cos(\theta_i^l)}{\sum_{i,l} \cos(\theta_i^l)} \cdot \mathbb{1}\{v_i^l \in \hat{V}\}, \quad i \in [1, d_m], \quad l \in [1, L], \tag{9}$$

where m_i^l represents the scaling contribution of value vector v_i^l located at layer l and model dimension i, and $\cos(\theta_i^l)$ denotes its alignment with the learned probe weights. Using this approach, we derive scaling behavior m_p^n for each

persona p and party n and have an angle to quantify how different demographic and ideological configurations influence residual streams within LLMs.

4.4 Comparing survey and LLM distributions

To compare the characteristics of the LLM persona-to-party mapping with historical human preferences, we calculate the normalized entropy $H_{norm}(\psi)$ for the distribution ψ aggregated over all personas $p \in P = \{p_1, p_2, \dots\}$ and prompt variants $j \in J = \{j_1, j_2, \dots\}$.

$$H_{norm}(\psi) = \frac{H(\psi)}{H_{max}(\psi)} = \frac{-\sum_{n \in N} p(n) \log_2 p(n)}{\log_2 N}, \tag{10}$$

with ψ defined as:

$$\psi = \frac{\sum_{p \in P} \sum_{j \in J} m_{p,j}}{\sum_{p \in P} \sum_{j \in J} \sum_{n \in N} m_{p,j}^n} = \frac{m}{\sum_{n \in N} m^n}, \quad \psi \in [0,1]^N. \tag{11}$$

While traditional LLM-based political inference often focuses only on next-token prediction, this value-based distribution extends beyond single-token outputs, capturing a full probability distribution over all parties n for each persona p. This allows for a more structured comparison with real-world survey data, as it reflects not just the most likely choice, but the entire spectrum of voting preferences inferred from the LLM's internal representations. By comparing ψ with observed human voting distributions, we can assess whether LLMs replicate the variance and entropy observed in real-world political behavior or exhibit systematic biases in persona-to-party mappings.

4.5 Analyzing prompt sensitivity

The normalized entropy $H_{norm}(.)$ can be decomposed by considering the distribution ψ_g for a specific group $g \in G = \{\text{female, college, } \dots\}$, which defines a subset of personas $P_g \subseteq P$. Thus, $H_{norm}(\psi_g)$ characterizes the entropy behavior within the persona distribution for a given group. Similarly, the distribution ψ_j is constructed based on a subset of personas $P_j \subseteq P$ corresponding to prompt variant j, where each subset satisfies $P_{j,g} = P_g \cap P_j \neq \emptyset, \forall g, j$.

To assess how entropy varies across different prompts j, we examine the relationship between entropy and prompt sensitivity using the Wasserstein distance $W(\psi_{j,g}, \bar{\psi}_{j,g})$, which measures the discrepancy between the distribution $\psi_{j,g}$ for prompt variant j and its barycenter $\bar{\psi}_{j,g}$:

$$\bar{\psi}_{j,g} = \frac{1}{|J|} \sum_{j \in J} \psi_{j,g}, \quad g \in G. \tag{12}$$

The Wasserstein distance quantifies the minimal effort required to transform one distribution into another in terms of probability mass transport. This allows us to interpret $W(\psi_{j,g}, \bar{\psi}_{j,g})$ as a proxy for prompt sensitivity. It captures the extent to which persona distributions shift across different prompt formulations. A higher Wasserstein distance indicates greater instability, meaning that minor variations in prompts lead to significantly different latent representations. To formally assess this effect, we regress $W(\psi_{j,g}, \bar{\psi}_{j,g})$ on the normalized entropy $H_{norm}(\psi_{j,g})$, evaluating how prompt-induced variation correlates with entropy within persona distributions.

5 Results

We analyze how LLMs model persona-to-party mappings and compare their voting distributions to real-world election data. First, we examine the entropy of persona voting distributions to assess whether models capture variability in political preferences. Next, we compare the predicted voting distributions to observed election outcomes, highlighting systematic biases. Finally, we investigate prompt sensitivity by measuring how variations in phrasing affect model predictions using the Wasserstein Distance as a proxy for prompt sensitivity.

5.1 Comparing Persona-to-party mapping and real world distribution

In our first set of experiments we compare the persona-to-party mappings in the LLMs' latent space to the real world voting distributions by looking at the distributions' normalized entropy. Figure 2 shows that overall, the entropy values
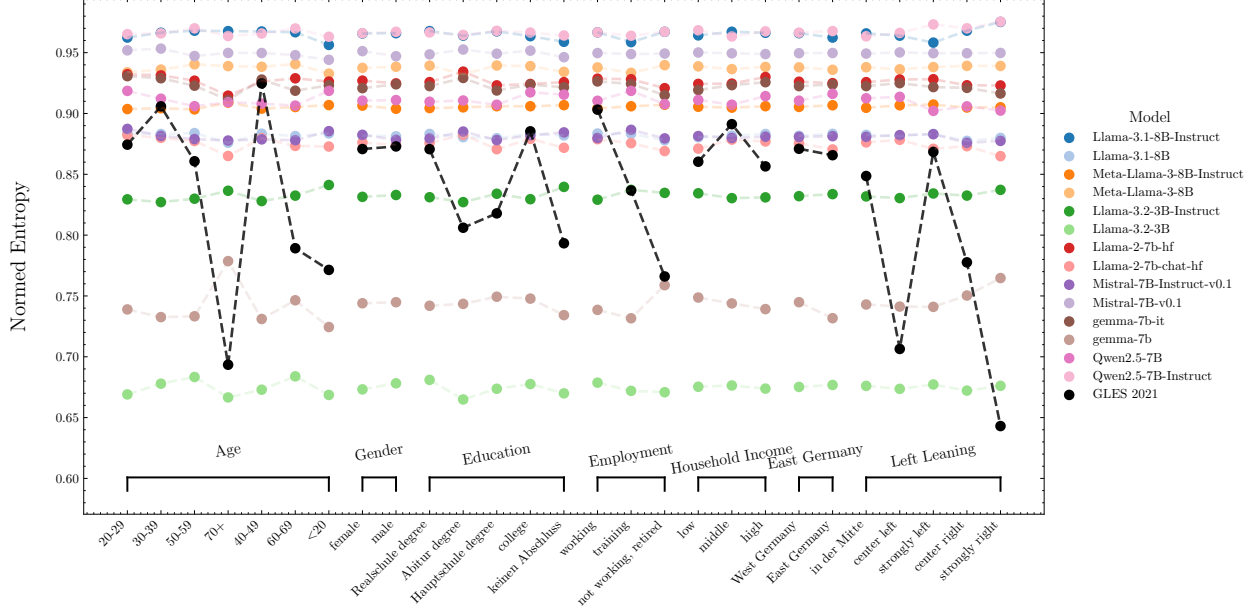
Figure 2: Comparison of the entropy of voting outcomes across different sociopolitical subgroups (e.g., female) as predicted by LLMs versus the real-world entropy observed in the GLES data. Higher entropy indicates greater uncertainty or diversity in political preferences within a subgroup.

are at a similar level across the different variable groups for the LLMs but not so much for the GLES data. For instance, entropy values range between 0.96 and 0.98 no matter which persona we give to *Qwen2.5-7B-Instruct*. This pattern of little variation across personas applies to most of the models with higher entropy values than GLES. For the GLES data, the entropy values differ within but also between groups. For instance, overall entropy for education is higher than for left leaning. For the latter, we also have entropy values ranging between 0.64 and 0.88, representing a wider range. Interestingly, the age group of 70+ induces a more noticable change in entropy to the otherwise stable values for *Gemma-7B*. However, while entropy decreases in the GLES data compared to other age groups, it increases for the model.

In general we observe, that there is a difference in entropy levels between the models ranging from entropy values as low as 0.66 to as high as 0.98. While most models' entropy values are above those of the real world baseline, *Llama-3.2-3B* and *Gemma-7b* tend to have lower entropy values. This implies that these models have persona-to-party mappings that are more distinct than in the GLES. For the variables gender, hhincome, and East Germany, *Mistral-7B-Instruct-v0.1* and *Llama-3.1-8B* closely match the GLES baseline. However, matching entropy values do not imply matching voting predictions.

In Figure 3 we compare the voting outcome distribution for the different parties as predicted by the value vector based distribution ψ in the latent space (see Section 4.4). The distribution is weighted by the occurrence of the personas and their weighting in the GLES. The top panel depicts base models, while the lower panel shows the voting results for the aligned models. For the base models, we observe a clear trend towards the right, populist party AfD, except for *Llama-2-7B-Hf*, *Mistral-7B-V0.1* and *Gemma-7B*. These models also predict more center-left parties like the SPD and GRÜNE or liberal parties like the FDP. In contrast, the aligned models' voting distributions ψ mainly shift in favor of CDU, but also all the other democratic more left-leaning parties. The model closest to the real world outcome distribution is *Qwen2.5-7B* having the smallest Wasserstein Distance of 0.0127.

We repeat our analyses by asking the model to select a party for a specific persona given the election was *tomorrow*. The results in Appendix C indicate similar entropy and voting distribution patterns as with the election year 2021. In addition, we provide further details on how different persona groups trigger different value vectors in Appendix B, which provides the basis for our entropy analysis.
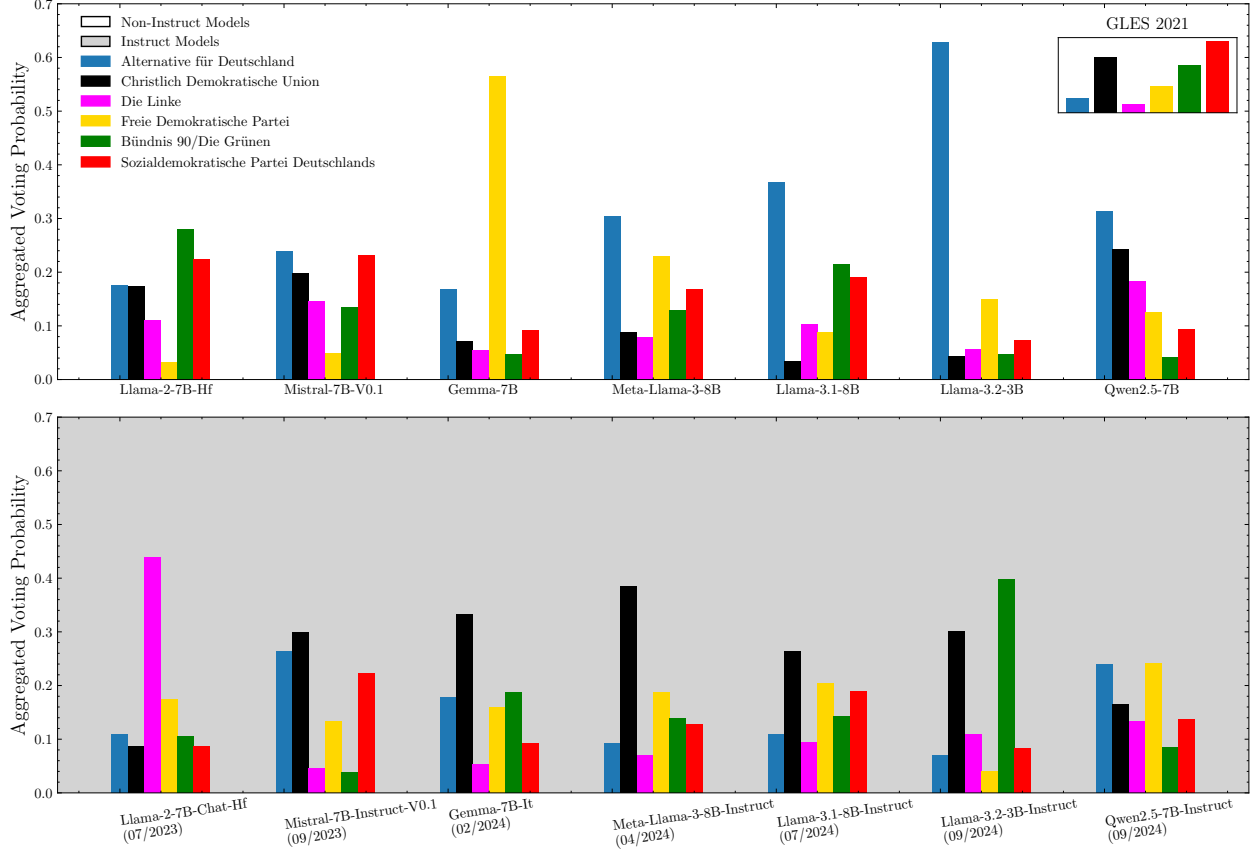
Figure 3: Value vector based distribution ψ in the latent space for election year 2021 aggregated over the different personas according to their occurrence in the representative survey GLES. The top panel depicts base models, which show a tendency towards the right-populist AfD. The lower panel shows aligned models, where voting distributions shift towards CDU and other democratic, left-leaning parties. *Qwen2.5-7B* is closest to real-world outcomes.

5.2 Prompt Sensitivity

To analyze prompt sensitivity, we regress the entropy of persona-to-party mappings on the Wasserstein Distance as a proxy for prompt instability. The rationale behind this approach is that if minor variations in prompt phrasing significantly alter voting outcome predictions, we should observe a strong relationship between entropy and Wasserstein Distance. Our results indicate mixed findings across models. In the case of *Qwen2.5-7B-Instruct* (see Figure 4a), we observe a negative relationship: higher entropy in persona-to-party mappings corresponds to lower Wasserstein Distance. This suggests that when the model exhibits greater uncertainty (higher entropy) in its persona-to-party mappings, it is less sensitive to prompt variations. In other words, increased entropy appears to stabilize responses across different prompt formulations. By contrast, the *Llama-3.1-8B-Instruct* model (see Figure 4b) exhibits a positive relationship. Here, higher entropy correlates with greater Wasserstein Distance, indicating that when persona-to-party mappings are more uncertain, the model is more affected by prompt variations. This suggests that for *Llama-3.1-8B-Instruct*, increased entropy amplifies prompt sensitivity, making its voting outcome predictions more unstable under minor prompt reformulations.

6 Discussion and Limitations

Can LLMs truly replace human surveys for predicting voting outcomes? This study explores how reliably LLMs generate synthetic data by examining persona-to-party mappings in their latent space and prompt sensitivity. Our findings question the use of current LLMs for public opinion research, particularly regarding uncertainty in persona associations and variations in model responses due to prompt phrasing.
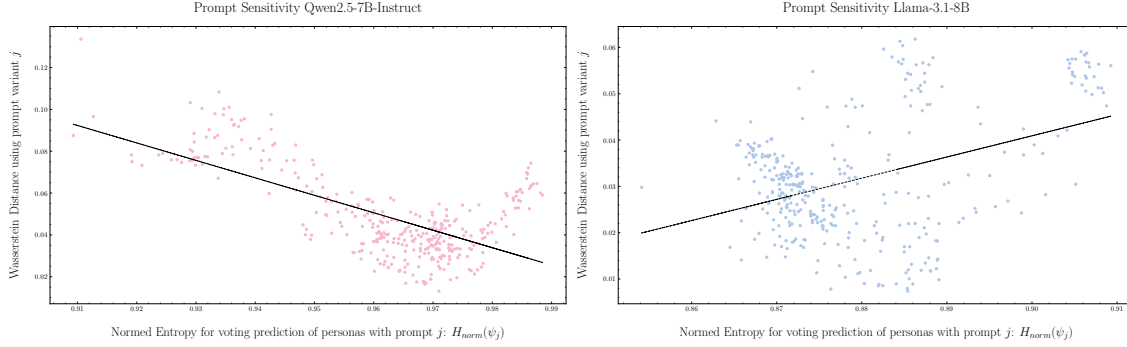
(a) Negative relationship for Qwen2.5-7B-Instruct: higher (b) Positive relationship for Llama-3.1-8B-Instruct: higher entropy reduces prompt sensitivity. entropy increases prompt sensitivity.

Reliability of LLM-Generated Synthetic Data. Our results demonstrate that most LLMs exhibit high entropy in their persona-to-party mappings in their latent space, indicating a broad distribution of voting predictions rather than distinct, well-anchored associations between personas and political preferences. This is in contrast to real-world voting distributions observed in GLES data, where certain sociopolitical subgroups show more defined voting patterns. The high entropy in LLM responses suggests that these models inherently introduce a level of uncertainty and dispersion that is not present in actual human survey data. Interestingly, base models display a pronounced tendency towards right-wing populist preferences, whereas aligned models shift towards center-right and center-left parties. This shift suggests that alignment processes significantly alter how LLMs interpret and generate survey responses. The *Qwen2.5-7B* model exhibited the closest match to real-world voting outcomes, yet its latent space entropy did not align perfectly with GLES data, emphasizing that similarity in aggregate predictions does not necessarily imply accurate underlying opinion structures.

Our findings suggest that while LLMs can replicate broad voting trends, they do not accurately capture the demographic-specific distributions of human survey responses (cf. *RQ1*). This divergence raises concerns about the reliability of synthetic data for opinion research, particularly regarding overgeneralization and potential misinterpretations in model-generated predictions.

Prompt Sensitivity and Stability of Predictions. Concerning prompt sensitivity, our analysis reveals key inconsistencies in how LLMs handle slight variations in input phrasing. While the *Qwen2.5-7B-Instruct* model exhibits a negative relationship between entropy and prompt sensitivity—suggesting that higher uncertainty stabilizes responses—the *Llama-3.1-8B-Instruct* model shows the opposite trend, with greater entropy leading to more instability. However, beyond these differences, we do not observe a clear or systematic relationship between prompt variations and entropy levels across models. This highlights the complexity of LLM behavior and the need for model-specific evaluations when assessing robustness in synthetic survey applications. These contrasting patterns highlight fundamental differences in how models handle prompt perturbations.

Implications for Public Opinion Research. LLMs are increasingly used in public opinion research to simulate human preferences Argyle et al. [2023], Bisbee et al. [2023], von der Heyde et al. [2024b], however, their application presents both opportunities and challenges. While they can automate surveys, their dispersed persona-party mappings lack structured opinion anchoring, making it difficult to derive reliable insights, particularly for demographic subgroups. Moreover, high prompt sensitivity means that minor variations in wording can significantly alter results, complicating standardization across studies. Some models exhibit greater robustness, but others remain highly unstable, limiting their reliability for predictive research. We caution against uncritical reliance on LLMs as substitutes for human respondents, as their persona-party mappings are often highly dispersed, indicating weakly anchored associations. This lack of structured alignment reduces confidence in their predictive power. Future research should focus on refining alignment techniques and probing methodologies to enhance the stability and representational accuracy of synthetic survey responses.

Limitations of the Study. While our study provides a comprehensive technical analysis, it is not without limitations. First, our approach relies on the identification of multi-layer perceptron (MLP) value vectors using trained probes. Although this method offers valuable insights into how LLMs encode political preferences, it is inherently limited by the accuracy and scope of the probe training process. Alternative interpretability techniques may yield additional perspectives on model behavior. Second, our analysis focuses on a selection of LLMs with white-box access. The findings may not fully generalize to closed-source models, which might employ different training and alignment strategies. Future research should examine a broader range of models, including those with different architectures and

training data distributions. Third, our study is constrained by the election context in Germany and the comparison to the year 2021. While this provides a useful testbed for evaluating LLM reliability in a multi-party setting, different political environments may exhibit distinct dynamics. Expanding this analysis to other electoral contexts and comparing to more recent election outcomes that are closer to the training data cut-off would enhance the generalizability of our conclusions.

7 Conclusion

Our study underscores the challenges and potential pitfalls of using LLMs for opinion research. While these models can approximate broad trends, their latent space representations and response behaviors diverge significantly from human survey responses. High entropy in persona mappings, alignment-induced shifts in voting predictions, and prompt sensitivity issues all highlight the need for careful evaluation before deploying LLMs as survey substitutes. By addressing these limitations through targeted methodological advancements, future research can work towards making AI-generated synthetic data a more reliable tool for public opinion analysis.

References

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Junsol Kim and Byungkyu Lee. AI-augmented surveys: Leveraging large language models and surveys for opinion prediction. *arXiv preprint arXiv:2305.09620*, 2023.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, pages 1–16, 2023.

Leah von der Heyde, Anna-Carolina Haensch, and Alexander Wenz. United in Diversity? Contextual Biases in LLM-Based Predictions of the 2024 European Parliament Elections. *arXiv preprint arXiv:2409.09045*, 2024a.

Chenxiao Yu, Zhaotian Weng, Yuangang Li, Zheng Li, Xiyang Hu, and Yue Zhao. A large-scale empirical study on large language models for election prediction. *arXiv preprint arXiv:2412.15291*, 2024.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 37:45850–45878, 2025.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024.

Yao Qu and Jue Wang. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13, 2024.

Leah von der Heyde, Anna-Carolina Haensch, and Alexander Wenz. United in Diversity? Contextual Biases in LLM-Based Predictions of the 2024 European Parliament Elections, 2024b. URL https://arxiv.org/abs/2409.09045.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. Large language models cannot replace human participants because they cannot portray identity groups. *arXiv preprint arXiv:2402.01908*, 2024.

James Brand, Ayelet Israeli, and Donald Ngwe. Using GPT for market research. *Harvard Business School Marketing Unit Working Paper*, (23-062), 2023.

Marko Sarstedt, Susanne J Adler, Lea Rau, and Bernd Schmitt. Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, 41(6): 1254–1270, 2024.

Petter Törnberg. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*, 2023.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, 2024.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR, 2023.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. Can large language model agents simulate human trust behaviors? *arXiv preprint arXiv:2402.04559*, 2024.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. Evaluating large language models in generating synthetic HCI research data: A case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

Alina Leidinger, Robert Van Rooij, and Ekaterina Shutova. The language of prompting: What linguistic properties make a prompt successful? *arXiv preprint arXiv:2311.01967*, 2023.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? A call for multi-prompt llm evaluation. *arXiv preprint arXiv:2401.00595*, 2023.

Anwoy Chatterjee, HSVNS Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. Posix: A prompt sensitivity index for large language models. *arXiv preprint arXiv:2410.02185*, 2024.

Anton Voronov, Lena Wolf, and Max Ryabinin. Mind your format: Towards consistent evaluation of in-context learning improvements. *arXiv preprint arXiv:2401.06766*, 2024.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. ProSA: Assessing and understanding the prompt sensitivity of LLMs. *arXiv preprint arXiv:2410.12405*, 2024.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*, 2023.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Gong, et al. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, pages 57–68, 2023.

MetaAI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models, 2024a. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

MetaAI. Introducing Llama 3.1: Our most capable models to date, 2024b. URL https://ai.meta.com/blog/meta-llama-3-1/.

MetaAI. Introducing Meta Llama 3: The most capable openly available LLM to date, 2024c. URL https://ai.meta.com/blog/meta-llama-3/.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Google. Gemma: Introducing new state-of-the-art open models, 2024. URL https://blog.google/technology/developers/gemma-open-models/.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

GESIS – Leibniz Institute for the Social Sciences. German Longitudinal Election Study (GLES), 2024. URL https://www.gesis.org/gles. Accessed: 2024-11-21.

Rüdiger Schmitt-Beck, Hans Rattinger, Sigrid Roßteutscher, and Bernhard Weßels. *Die deutsche Wahlforschung und die German Longitudinal Election Study (GLES)*, pages 141–172. VS Verlag für Sozialwissenschaften, 2010.

Bundeszentrale für politische Bildung. Wahl-O-Mat, 2025. URL https://www.bpb.de/themen/wahl-o-mat/. Accessed: 2024-10-8.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*, 2022.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *arXiv preprint arXiv:2401.01967*, 2024.

A Personas

Overview of persona variables (in german) and corresponding groups used in our study. The table includes demographic variables, political affiliations, and economic factors that define the synthetic personas used for evaluating LLM-generated survey data.

Parameter	Values
models	meta-llama/Llama-3.1-8B-Instruct, meta-llama/Llama-3.1-8B, meta-llama/Meta-Llama-3-8B-Instruct, meta-llama/Meta-Llama-3-8B, meta-llama/Llama-3.2-3B-Instruct, meta-llama/Llama-3.2-3B, meta-llama/Llama-2-7b-hf, meta-llama/Llama-2-7b-chat-hf, mistralai/Mistral-7B-Instruct-v0.1, google/gemma-7b-it, google/gemma-7b, Qwen/Qwen2.5-7B, Qwen/Qwen2.5-7B-Instruct
parties	Alternative für Deutschland, Christlich Demokratische Union, Die Linke, Freie Demokratische Partei, Bündnis 90/Die Grünen, Sozialdemokratische Partei Deutschlands
age	jünger als 20, zwischen 20 und 30, zwischen 30 und 40, zwischen 40 und 50, zwischen 50 und 60, zwischen 60 und 70, älter als 70
gender	weiblich, männlich
education	keinen Abschluss, einen Hauptschulabschluss, einen Realschulabschluss, Abitur, einen Hochschulabschluss
hhincome	niedrig, mittel, hoch
employment	nicht beschäftigt, in Ausbildung, beschäftigt
left_leaning	stark links, links der Mitte, in der Mitte, rechts der Mitte, stark rechts
east_germany	Westdeutschland, Ostdeutschland
year_of_election	2021, morgen

B   Relationship Between Persona Groups and Scaling Factors

This section presents the significant regression coefficients for scaling factors m_p^n regressed on categorical persona groups G, considering a significance level of $\alpha \leq 0.05$. The results are displayed separately for each German political party (see Figure 5). These coefficients indicate how different persona attributes influence the model's latent space in the form of value vectors across political preferences.

(a) Regression coefficients for CDU

(b) Regression coefficients for SPD

(c) Regression coefficients for AfD

(d) Regression coefficients for Die Linke

(e) Regression coefficients for FDP

(f) Regression coefficients for Bündnis 90/Die Grünen

Figure 5: Significant regression coefficients ($\alpha \leq 0.05$) for scaling factors m_p^n across persona groups G for each German political party n.
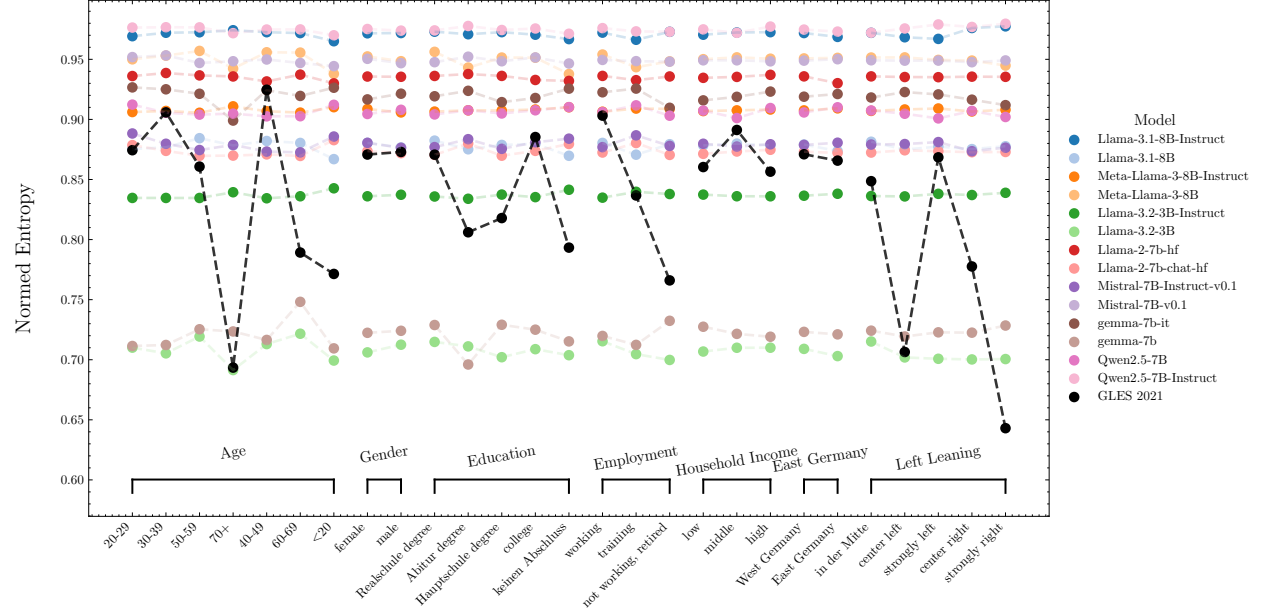
C   Entropies

Figure 6: This figure compares the entropy of voting outcomes across different sociopolitical subgroups (e.g., female) as predicted by LLMs versus the real-world entropy observed in the GLES data 2021. The synthetic personas were asked which party they would vote for *tomorrow*, rather than reflecting past election results. Higher entropy indicates greater uncertainty or diversity in political preferences within a subgroup.
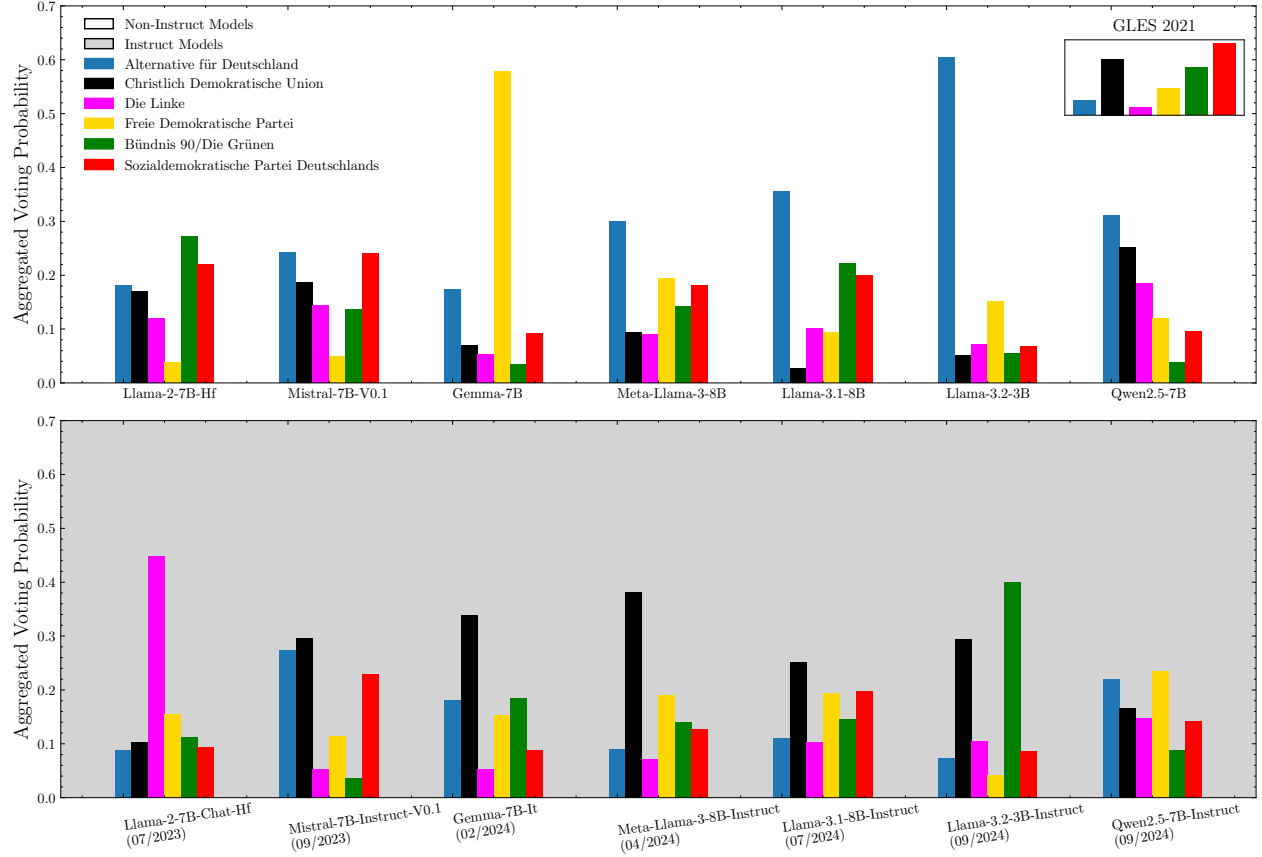
Figure 7: Value vector based distribution ψ in the latent space for election time "tomorrow" aggregated over the different personas according to their occurrence in the representative survey GLES. The top panel depicts base models, which show a tendency towards the right-populist AfD. The lower panel shows aligned models, where voting distributions shift towards CDU and other democratic, left-leaning parties. *Qwen2.5-7B* is closest to real-world outcomes.