Is there a "Hierarchy of Personas" in synthetic data generated through Large Language Models?

Ben Lasse Wolf
German Centre for Higher Education Research and Science Studies (DZHW)
Leibniz University Hannover

Joshua Claassen
German Centre for Higher Education Research and Science Studies (DZHW)
Leibniz University Hannover

Jan Karem Höhne
German Centre for Higher Education Research and Science Studies (DZHW)
Leibniz University Hannover

Abstract
Survey researchers increasingly use Large Language Models (LLMs) as synthetic respondents in so-called silicon sampling approaches. Although this is typically coupled with persona prompting as a technique to steer LLM responses, little attention has been given to the relative influence of different persona dimensions. To address this gap, we introduce the concept of a "Hierarchy of Personas" by investigating how the persona dimensions age, education, gender, and party affiliation influence LLM responses to LGBTQ-related survey items. To this end, we use the instruction-tuned open model Gemma-3 and prompt it to repeatedly respond to four LGBTQ-related survey items while role-playing 162 distinct persona profiles, resulting in a total of 3,240 responses. Our results show that the persona-conditioned response behavior is not equally shaped by all persona dimensions, but rather follows a non-uniform and item-contingent Hierarchy of Personas. Across most items, party affiliation exerts the strongest influence, education ranks second, and age and gender have a comparatively weak influence. However, the persona-conditioned responses are sensitive to item characteristics, so that the Hierarchy of Personas is not fully uniform. Our article advances the current state-of-the-art in survey research by showcasing that persona prompting is not a neutral and transparent technique for simulating the attitudes of specific social groups, but rather constitutes a bounded form of model-internal responsiveness under controlled conditions.

Keywords: Silicon sampling, Persona prompting, Large Language Models, Synthetic data, Responsible artificial intelligence

This document is a preprint and thus it may differ from the final version.

Introduction and overarching research question

Rapid developments in the field of artificial intelligence (AI) have introduced various new methodological possibilities for survey research. For example, researchers increasingly use Large Language Models (LLMs) as synthetic respondents. In such silicon sampling approaches, LLMs are used to generate responses that approximate group-specific response patterns or distributions without relying on human respondents (Argyle et al., 2023). While silicon sampling is a scalable as well as cost- and time-efficient way of generating large amounts of data (Boelaert et al., 2025; Grossmann et al., 2023; Rupprecht et al., 2025), it remains unclear whether LLM responses can actually capture the heterogeneity characterizing human survey data. This question is closely related to recent debates about the concept of algorithmic fidelity, that is, the extent to which silicon samples are able to recover not only distributions at the aggregate level but also more complex associations and interactions of human behavior (Argyle et al., 2023). Within this broader silicon sampling literature, persona prompting has emerged as the central steering technique for configuring synthetic respondents. This typically involves assigning an LLM a distinct persona profile, for example through sociodemographic attributes, such as education or gender, and instructing it to respond to survey items from that specific perspective (Lutz et al., 2025).

However, especially in the context of vulnerable and marginalized social groups, prompting an LLM to respond as if it were a group member carries the risk of reproducing stereotypes, simplifications, or flattened representations rather than recovering authentic group perspectives based on lived experience (Wang et al., 2025). Therefore, a key question remains how different persona dimensions influence LLM responses and what this can reveal about persona prompting as a source of model bias and selective responsiveness to social background information (Gupta et al., 2023; Lutz et al., 2025).

Prior research has examined the general promises, limits and risks of persona prompting, but much less attention has been paid to the relative influence of different persona dimensions when several of them are combined. In the present article, we address this gap by investigating the following overarching research question: *How do the persona dimensions age, education, gender, and party affiliation influence LLM responses to LGBTQ-related survey items?*

To this end, we created 162 distinct persona profiles, based on all possible combinations of four persona dimensions (i.e., age, education, gender, and party affiliation), and prompted the instruction-tuned open model Gemma-3 (Gemma Team, 2025) to respond to four LGBTQ-related items while role-playing each persona profile five times (N = 3,240 responses). Importantly, by introducing the concept of a Hierarchy of Personas, our article makes a theoretical contribution to the current state-of-the-art regarding silicon sampling, especially in the context of survey research.

Background

Promises and pitfalls of LLM responses in survey research

Silicon sampling approaches are attracting growing attention in survey research, since early studies suggest that LLMs can reproduce human-like judgments and may therefore be useful for simulating human responses to survey items (Dillion et al., 2023). Relatedly, previous studies have

explored the use of LLMs for other survey tasks, including the prediction of item-nonresponse and imputation of missing values (Ji et al., 2024). However, empirical results are mixed. For example, Argyle et al. (2023) find that GPT-3 can recover group-level distributional patterns when conditioned on rich persona background descriptions derived from the American National Election Studies (ANES), indicating that silicon samples partially reproduce human survey outcomes. In contrast, Ma et al. (2025), using open-ended data from the German Longitudinal Election Study (GLES), show that the algorithmic fidelity of LLM responses decreases as within-group opinion heterogeneity increases. In line with this finding, a further GLES-based study on German vote choice by von der Heyde et al. (2025) finds that GPT-3.5-based voting predictions systematically differ from survey-based estimates.

These findings have fueled the broader methodological debate about whether LLM responses can in fact supplement or even replace human respondents. A central criticism of using LLMs as synthetic respondents is that their learned response space is largely determined by their training data and alignment procedures. Specifically, training LLMs on large corpora of human-generated text may not only internalize grammatical and factual regularities, but socially conditioned patterns, including stereotypes, hegemonic viewpoints, and other potentially harmful biases (Bender et al., 2021; Weidinger et al., 2022). In addition, based on supervised fine-tuning and reinforcement learning from human feedback (RLHF), human annotators further shape which LLM responses are considered helpful, harmless, or otherwise preferable (Ouyang et al., 2022). Therefore, LLMs may reproduce historically and socially conditioned response patterns rather than contemporary public opinion (Harding et al., 2024).

Persona prompting as a steering technique

In order to create silicon samples consisting of synthetic respondents belonging to specific groups, researches typically utilize persona prompting. Tseng et al. (2024) distinguish between LLM-role-playing, in which the LLM temporarily adopts the perspective of a member of a specific group, and LLM-personalization, in which LLM responses are adapted over time to a specific user profile derived from background information or interaction history (e.g., in recommendation systems) (Tseng et al., 2024). Importantly, the present article utilizes role-playing by instructing Gemma-3 to respond to items as if it were a respondent with a given social profile. Since the focus is to systematically investigate how Gemma-3 responds to different persona dimension inputs while imitating varying social perspectives, utilizing role-playing as the persona prompting approach was considered the most effective prompting strategy. Importantly, this approach has been also employed in previous research on synthetic data and persona prompting (see Ahnert et al., 2025; Argyle et al., 2023; Qi et al., 2025; von der Heyde et al., 2025). Thus, persona prompting is understood as "a prompting technique that is used to steer the behavior of an LLM to align with that of a specified sociodemographic group or person" (Lutz et al., 2025).

Although persona prompts can meaningfully influence LLM responses, a growing body of research cautions that such prompts can also introduce systematic harm, particularly when LLMs are instructed to speak as members of specific, marginalized social groups. Rather than

reproducing lived experience, LLM responses may simply reproduce patterns learned from the training data, which itself is shaped by existing social knowledge and inequalities. Persona prompting is therefore not only a steering technique, but a potentially bias-producing intervention into the model's learned response space.

One central concern is how persona prompts represent groups. In a recent study, Wang et al. (2025) identify three recurring problems in persona-conditioning of LLMs: (1) Misportrayal refers to persona-conditioned responses that resemble out-group perceptions rather than in-group self-representations. (2) Flattening refers to significantly reduced within-group heterogeneity that does not reflect the true diversity of the corresponding group. (3) Essentialization refers to the risk that prompting with sociodemographic attributes may reduce social identities to a set of fixed characteristics (Wang et al., 2025). In sum, these dynamics suggest that persona-conditioned responses risk speaking for marginalized groups, rather than speaking from these groups. Relatedly, evidence from the Marked Personas framework shows that persona prompts can systematically produce stereotypical descriptions of social identities, particularly for non-white and non-male identities (Cheng et al., 2023). Importantly, such representational harm is not limited to explicitly negative responses, but can also appear in more subtle, positive-sounding forms (e.g., hypersexualization of Asian or Latina women) (Cheng et al., 2023).

Furthermore, previous research suggests that persona prompts do not affect LLM responses to the same extent, but can lead to uneven stereotype activation (Gupta et al., 2023; Wang et al., 2025). For example, prompts invoking non-binary or racialized personas have been reported to produce more stereotyped language and lower linguistic diversity than prompts for majority-coded personas (Li et al., 2025; Lutz et al., 2025). Studies on silicon sampling likewise indicate that LLMs do not approximate all groups equally well, pointing to uneven responsiveness across persona dimensions. This especially applies to non-majority-coded and already marginalized personas (Santurkar et al., 2023; Qi et al., 2025). Evidence from the German context further suggests that LLM responses often rely disproportionately on politically saturated information, while struggling to reproduce more complex attitudinal and sociodemographic relationships at the group level (Ma et al., 2025; von der Heyde et al., 2025). Taken together, these findings suggest that persona dimensions should not be understood as equivalent switches that can simply be turned on and off, but as conditioning mechanisms whose behavioral influence may differ systematically depending on how strongly they are internalized, activated, and stabilized within the LLM.

Hierarchy of Personas and research hypotheses

Drawing on the current state of research, our article introduces the concept of a Hierarchy of Personas, that is, the relative extent to which the persona dimensions age, education, gender, and party affiliation influence LLM responses to LGBTQ-related items. More specifically, our analysis examines four research hypotheses about persona-conditioned response behavior under a fixed persona prompting setup.

The first hypothesis is based on the expectation that persona dimensions do not function as interchangeable or uniformly effective interventions into an LLM's learned response space. Prior

work suggests uneven responsiveness across different persona dimensions, both because the algorithmic fidelity in silicon sampling studies differs substantially across different groups, and because persona prompts appear to operate as selective steering mechanisms rather than neutral reflections of group perspectives (Argyle et al., 2023; Bisbee et al., 2024; Lutz et al., 2025; Ma et al., 2025).

> ***H1:*** *Under the fixed persona prompting setup, the persona dimensions age, education, gender, and party affiliation differ systematically in their relative influence on LLM responses to LGBTQ-related survey items, implying a Hierarchy of Personas.*

The second hypothesis specifies this expected hierarchy in more concrete terms. Existing research on silicon sampling and persona prompting indicates that LLM responses are often more strongly influenced by persona dimensions that are related to ideological worldviews than by (weakly saturated) sociodemographic dimensions (Ma et al., 2025; Qi et al., 2025). Within our study design, party affiliation constitutes the most explicitly political and ideologically saturated persona dimension. It therefore appears theoretically plausible that party affiliation will influence LLM responses more strongly than the remaining persona dimensions.

> ***H2:*** *Under the fixed persona prompting setup, party affiliation exerts a stronger relative influence on LLM responses to LGBTQ-related survey items than age, education, and gender.*

The third hypothesis follows from the notion that the performance of persona prompts is context-dependent (Beck et al., 2024; Jiang et al., 2024). For example, research on prompt perturbations and response formats demonstrates that LLM responses to survey items are sensitive to wording, option ordering, response scale design, and response generation method, implying that response patterns depend strongly on the specific measurement setup (Ahnert et al., 2025; Röttger et al., 2024; Rupprecht et al., 2025). Accordingly, the relative influence of persona dimensions may not be constant, but vary across survey items that differ in their substantive content and wording.

> ***H3****: The relative influence of persona dimensions varies across survey items rather than forming a uniform pattern.*

The fourth hypothesis is deducted from prior work suggesting that non-majority-coded personas often produce more stereotyped language, lower linguistic diversity, and less robust persona effects than majority-coded personas (Cheng et al., 2023; Lutz et al., 2025; Wang et al., 2025). Relatedly, non-binary personas are generally less represented in the model's learned response space and might therefore be more unstable. Consequently, when prompted repeatedly, non-binary personas may be associated with lower response stability.

H4: *Under the fixed prompting setup, non-binary gender personas show lower run-to-run stability of LLM responses than binary-gender personas.*

Methodology

Persona profiles

Partially in contrast to prior research, the present study intentionally excludes sensitive persona dimensions, such as race and disability status, to reduce the risk of misportrayal, flattening or essentialization of these groups (Wang et al., 2025). Instead, we focus on four common persona dimensions, including age, education, gender, and party affiliation, resulting in 162 distinct persona profiles (or combinations). For the prompt design, we mainly followed the studies by Höhne et al. (2025a; 2025b; see also "LLM setup and response generation"). These dimensions are relevant in research on human LGBTQ-related attitudes, and they are commonly used in studies on persona prompting and silicon sampling (see Bisbee et al, 2024; Boelaert et al, 2025; Ma et al., 2025; Qi et al., 2025). Table 1 summarizes the four persona dimensions and their categories.

Table 1. Persona dimensions and categories

Persona dimension	Categories
Age	1. Young adults (18 to 35 years) 2. Middle-aged adults (36 to 55 years) 3. Older adults (over 55 years)
Education	1. Low education 2. Medium education 3. High education
Gender	1. Male 2. Female 3. Non-binary
Party affiliation	1. Alternative für Deutschland (AfD) 2. Bündnis 90/Die Grünen 3. Christlich-Demokratische Union (CDU) 4. Die Linke 5. Freie Demokratische Partei (FDP) 6. Sozialdemokratische Partei Deutschlands (SPD)

Note. Party affiliation distinguishes between six German parties that have been represented in the national parliament in recent legislative periods (Die Bundeswahlleiterin, 2025).

LGBTQ-related items

To conduct our research on ecologically valid survey items, we adopt four items from established large-scale survey programs that measure different aspects of LGBTQ-related attitudes. Two items are based on the Special Eurobarometer 535 – Discrimination in the European Union (European Commission, 2023) and capture normative evaluations of equal rights related to marriage, adoption, and parenting for lesbian, gay, bisexual as well as transgender people. The remaining

two items are based on the European Social Survey (ESS, 2022: Round 11) and capture affective attitudes as well as support for equal life choices towards gay men and lesbians.

1. *Lesbian, gay and bisexual people should have the same rights as heterosexual people (marriage, adoption, parental rights).*
2. *Transgender people should have the same rights as anyone else (marriage, adoption, parental rights).*
3. *If a close family member was a gay man or a lesbian, I would feel ashamed.*
4. *Gay men and lesbians should be free to live their own life as they wish.*

All survey items were translated into German for use in the prompts. To ensure clear and consistent prompting across items, we used close German translations with minor wording adaptations. Response formats follow the original item version to maximize ecological validity. The two Eurobarometer items use a four-point Likert-type scale ranging from 1 ("totally agree") to 4 ("totally disagree"). The two ESS items use a five-point Likert-type scale ranging from 1 ("agree strongly") to 5 ("disagree strongly"). Appendices 1 and 2 provide the exact German wording of the items and Likert-type scales, respectively.

LLM setup and response generation

All responses were generated with the instruction-tuned open model Gemma-3, using the 12-billion-parameter variant of Google's Gemma-3 family (Gemma3:12B; Gemma Team, 2025). This model family is frequently used in persona prompting and silicon sampling studies (see Lutz et al., 2025; Ma et al., 2024; Ma et al., 2025; Plaza-del-Arco et al., 2025). The model is run locally using the open-source framework Ollama (Ollama, 2024) and its Python library on a macOS workstation (Appendix 3 provides technical details). Response generation took place on February 12, 2026. All scripts for response generation and preprocessing are written in Python (Version 3.13), and statistical analyses are conducted in RStudio (Version 4.3.2 (2023-10-31); Posit Team, 2025).

The prompt for instructing Gemma-3 was adopted from Höhne et al. (2025a; 2025b) and subsequently refined so that it collects both the closed-ended response and a short textual reasoning. The prompt includes the instruction to behave like a survey respondent with placeholders for the respective persona profile (i.e., combination of age, education, gender, and party affiliation; see Table 1), the text of the survey item, a description of the corresponding Likert-type scale, and instructions to respond in the required format. The final German prompt is shown in Figure 1. An English translation of the used prompt is documented in Appendix 4.

Gemma-3 was queried in chat mode via Ollama with fixed decoding parameters. Sampling is enabled, with top_p set to 0.95 and top_k to 64, following the generation configuration distributed with the official Gemma-3 model release (see https://huggingface.co/google/gemma-3-12b-it/blob/main/generation_config.json). Temperature is set to the default value of 1.0 (see https://ollama.com/library/gemma3/blobs/3116c5225075). To preserve stochastic variation while ensuring reproducibility, each model call uses a fixed but run-specific random seed, which is

recorded in the dataset. The maximum output length for the reasoning text is limited to 200 tokens via Ollama's generation limit (num_predict) to avoid extensive explanations that would increase computational costs. To detect potential response generation issues, we manually checked all reasonings provided by Gemma-3. However, we did not detect (overly) inconsistent reasonings.

Finally, for each persona profile, the script iterates over all four survey items. As each persona-item combination is queried five times, this results in 162 personas × 4 items × 5 runs = 3,240 numeric responses and the same number of corresponding reasoning texts.

```
Verhalte dich wie eine deutschsprachige {gender} Person {age}, mit {education} und einer Nähe zur Partei {party}, die an einer Umfrage teilnimmt.
Beantworte aus der Sicht dieser Persona die folgende Aussage ausschließlich als Zustimmungsgrad zur AUSSAGE (nicht zur Thematik allgemein): {item}

{scale_block}

Antwortformat (wichtig):
Zeile 1: {answer_line}
Zeile 2: Reasoning: 1–2 kurze Sätze, die aus der Sicht dieser Persona erklären, warum sie diesen Zustimmungsgrad zu GENAU dieser Aussage wählt.
Keine weiteren Ausgaben.
```

Figure 1. Prompt template

Analytical strategy

Before data analysis, the raw responses were checked for parsing errors (e.g., non-numeric, out-of-range, and missing responses) but no erroneous responses were observed. In addition, to ease interpretation, the negatively worded third item was reverse-coded so that lower values consistently indicate more LGTBQ-accepting responses across all four survey items. Based on the cleaned responses, we follow a three-step analytical strategy. First, for each item, we report descriptive statistics on Gemma-3's responses by calculating the mean and standard deviation across all 162 persona profiles and runs, respectively.

Second, we estimate item-wise linear regression models at the run level (n = 810), using the Gemma-3 responses as dependent variable, to assess how strongly the four persona dimensions influence responses, and whether some dimensions have a stronger relative influence than others. For each item, five regression models are estimated, including four one-way models with the persona dimensions as single predictors, respectively, as well as one full model that includes all four persona dimensions simultaneously. All persona dimensions are dummy-coded with one reference category per persona dimension: older adults, low education, male gender, and Alternative für Deutschland (party affiliation). These categories are selected for interpretative clarity and are not intended to represent a neutral or average persona profile.

To assess H1, postulating systematic differences between persona dimensions in their relative influence on Gemma-3's responses, we compare the adjusted R^2 from the one-way models. For each item, the adjusted R^2 indicates how much of the response variation can be explained by

the persona dimension under investigation, taking model complexity into account. Furthermore, to assess H2, postulating that party affiliation exerts the strongest relative influence on Gemma-3's responses, we investigate both the adjusted R^2 of the one-way models as well as the coefficients of the full model. Additionally, we evaluate H3, postulating that the relative influence of persona dimensions varies across items, by examining the adjusted R^2 of the one-way models across items.

Finally, we focus on H4, postulating lower run-to-run response stability for non-binary gender personas than for binary gender personas. To this end, we compare the standard deviation and response range across runs between gender categories, separately for each item. Appendix 5 provides the same statistics for all persona dimensions under investigation.

Results

Descriptive statistics on Gemma-3's responses

In the first step, for each item, we examine the mean and standard deviation of Gemma-3's responses across persona profiles and runs. Table 2 shows the results. Items 1, 2, and 4 show comparatively low mean values ranging from 1.39 (item 1) to 1.71 (item 2), indicating overall high acceptance with the statements. In contrast, the reverse-coded (and negatively worded) item 3 shows a substantially higher mean value of 3.79, indicating a rather low acceptance with the statement. This suggests that the items do not only differ in scale format and polarity, but also in the average level of acceptance they elicit from the LLM. Interestingly, item 3 also has a substantially higher standard deviation (1.30) than the remaining items (ranging from 0.73 to 0.97), indicating that between-persona variation depends on the item employed. However, since items 1 and 2 come with four-point and items 3 and 4 with five-point scales, mean values and standard deviations should be interpreted and compared with caution.

Table 2. Descriptive statistics and scale format by item

Item	Polarity	Response scale	M	SD
1	Positive	1 to 4	1.39	0.73
2	Positive	1 to 4	1.71	0.95
3	Negative	1 to 5	3.79	1.30
4	Positive	1 to 5	1.51	0.97

Note. M = mean response; SD = standard deviation. Both statistics are based on persona-item means averaged across five runs. Lower values consistently indicate more accepting responses under the harmonized coding (item 3 is reverse-coded). n = 162 persona profiles per item.

Linear regression models

In the second step, to systematically assess the Hierarchy of Personas across the four items, we estimate linear regression models. In total, four one-way models are estimated per item: age, education, gender, and party affiliation. The central hierarchy metric is the adjusted R^2 of these one-way models, which indicates how much response variation is explained by a given persona dimension when considered in isolation and while accounting for model complexity. Table 3

reports the adjusted R^2 values of the item-wise one-way models together with the resulting hierarchy ranks.

Across the four items, the results reveal a Hierarchy of Persona dimensions that vary in their relative explanatory power (or influence), providing support for H1. In particular, party affiliation and education emerge as the most influential persona dimensions, whereas age has comparatively little and gender even less influence across the items. A particularly clear pattern emerges for items 1, 2, and 4. For all three items, the party affiliation model shows the highest adjusted R^2 values (item 1 = 0.65; item 2 = 0.69; item 4 = 0.79). The education model ranks second for all three items (item 1 = 0.09; item 2 = 0.13; item 4 = 0.04), while the age model has only modest explanatory power (item 1 = 0.02; item 2 = 0.03; item 4 = 0.01). The gender model shows almost no isolated explanatory power, with adjusted R^2 values close to zero. In conclusion, these results indicate that the Hierarchy of Personas is clearly structured, and, for most items, most strongly organized by party affiliation. This observed pattern is consistent with H2.

At the same time, the hierarchy is not fully uniform across the four items. Specifically, for item 3, the education model ranks first (adjusted R^2 value = 0.39), while the party affiliation model ranks second (adjusted R^2 value = 0.28). The gender and age models remain weak (adjusted R^2 values = 0.01 for both models). The reversal of the top two ranks indicates that the relative influence of persona dimensions is not constant across items, but seems to depend on item characteristics, such as content and wording. This pattern supports H3 and suggests that the Hierarchy of Personas varies across items rather than forming a uniform pattern. To complement the one-way model comparisons, item-wise full regression models are estimated at the run level, including age, education, gender, and party affiliation simultaneously. Table 4 reports the coefficients and standard errors (SE) of these models. Across the four items, the full models achieve high levels of explanatory power. Adjusted R^2 values range from 0.70 (item 3) to 0.86 (item 2), indicating that the four persona dimensions jointly account for a substantial share of variation in Gemma-3's responses.

In line with our previous results on H2, the coefficients of party affiliation are consistently associated with Gemma-3's responses across all four items. Specifically, compared to Alternative für Deutschland (AfD), all parties are associated with more LGBTQ accepting responses. This association is most pronounced for left-leaning parties (i.e., Bündnis 90/Die Grünen, Die Linke, and SPD). Similarly, education shows a clear and consistent pattern across items. Compared to low education, both medium and high education are associated with more accepting responses across all four items. The age coefficients point in a similar direction, albeit with a smaller effect size. Compared to older adults, younger and middle-aged personas are associated with more accepting responses. As before, gender plays a rather limited and item-dependent role in the full models. Although some gender coefficients reach statistical significance, their size remains modest compared to those of party affiliation and education.

Table 3. Item-wise one-way models and resulting hierarchy of persona dimensions

	Age model		Education model		Gender model		Party affiliation model	
Item	Adjusted R^2	Rank	Adjusted R^2	Rank	Adjusted R^2	Rank	Adjusted R^2	Rank
1	0.02	3	0.09	2	0.00	4	0.65	**1**
2	0.03	3	0.13	2	0.00	4	0.69	**1**
3	0.01	4	0.39	**1**	0.01	3	0.28	2
4	0.01	3	0.04	2	0.00	4	0.79	**1**

Note. Entries report adjusted R^2 values from item-wise one-way models estimated at the run level. The row-wise rank refers to the relative explanatory power of the four persona dimensions within each item, with rank 1 indicating the strongest isolated predictor.

Table 4. Item-wise full OLS models

	Item 1		Item 2		Item 3		Item 4	
	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
Age (reference: older adults)								
Young adults	-0.26***	0.03	-0.45***	0.03	-0.39***	0.06	-0.26***	0.03
Middle-aged adults	-0.16***	0.03	-0.22***	0.03	-0.23***	0.06	-0.18***	0.03
Education (reference: low education)								
Medium education	-0.33***	0.03	-0.60***	0.03	-1.44***	0.06	-0.35***	0.03
High education	-0.53***	0.03	-0.81***	0.03	-1.99***	0.06	-0.50***	0.03
Gender (reference: male)								
Female	-0.04	0.03	-0.09**	0.03	-0.20**	0.06	-0.10**	0.03
Non-binary	-0.09**	0.03	-0.17***	0.03	0.21***	0.06	-0.03	0.03
Party affiliation (reference: Alternative für Deutschland (AfD))								
Bündnis 90/Die Grünen	-1.67***	0.04	-2.22***	0.04	-2.19***	0.09	-2.40***	0.05
Christlich-Demokratische Union (CDU)	-1.26***	0.04	-1.45***	0.04	-0.90***	0.09	-1.82***	0.05
Die Linke	-1.67***	0.04	-2.24***	0.04	-1.94***	0.09	-2.40***	0.05
Freie Demokratische Partei (FDP)	-1.41***	0.04	-1.80***	0.04	-0.96***	0.09	-2.30***	0.05
Sozialdemokratische Partei Deutschlands (SPD)	-1.67***	0.04	-2.23***	0.04	-1.24***	0.09	-2.40***	0.05
Adjusted R^2	0.76		0.86		0.70		0.85	
Responses	810		810		810		810	

Note. Entries report unstandardized coefficients (Coef.) and standard errors (SE) from item-wise full OLS models estimated at the run level. Lower values indicate more LGBTQ accepting responses, whereas higher values indicate less accepting responses. Item 3 is reverse-coded.

***p < 0.001, **p < 0.01, *p < 0.05

Run-to-run response stability

In the third step, to assess Gemma-3's response stability across runs, we investigate run-to-run variation between gender categories. Specifically, for each persona-item combination, we calculate the mean and standard deviation across runs and then average them within items and persona categories. Table 5 shows the results for the gender categories. Overall, Gemma-3 shows comparatively high response stability across runs. For three of the four items, overall run-to-run variation remains low, with mean standard deviations ranging from 0.03 (item 1) to 0.06 (item 2), and mean response ranges ranging from 0.07 (item 1) to 0.13 (item 3). As before, the negatively worded item 3 constitutes an exception with the highest standard deviation (0.18) and response range (0.37), indicating that it is more sensitive to variation than the remaining items.

A comparison between gender categories reveals a mixed pattern. Regarding items 1 and 2, all three gender categories show only small differences in their standard deviation and response range. The clearest difference emerges once more for item 3, where non-binary personas display distinctly higher run-to-run variation (SD = 0.23, response range = 0.48) than both male and female personas (SD = 0.15, response range = 0.32, respectively). In contrast, item 4 shows the opposite pattern, with non-binary personas resulting in slightly more stable responses than binary personas. With regard to H4, the evidence is mixed. Non-binary personas do not show uniformly higher variation across all items, but they are less stable on negatively worded item 3, while differences for the other items remain small and inconsistent.

Furthermore, the descriptive statistics in Appendix 5 indicate that response stability also varies across age, education, and party affiliation categories, but no category appears highly unstable in a systematic manner.

Discussion and conclusion

Main results and their interpretation

This article examined how the persona dimensions age, education, gender, and party affiliation influence Gemma-3's responses to LGBTQ-related survey items under a fixed persona prompting setup. Taken together, in line with H1, the results show that these dimensions do not influence response behavior equally, but form a non-uniform, asymmetrically structured, and item-contingent Hierarchy of Personas. Specifically, in line with H2, party affiliation emerges as the most influential dimension in three out of four items. In addition, education constitutes an influential secondary dimension and exceeds party affiliation in one item, whereas age and gender have a comparatively weak influence. In line with H3, the observed *hierarchy* is not uniform across all items. Specifically, in the context of item 3, education exerts a stronger relative influence than party affiliation. Finally, our results on H4 are mixed, as only item 3 provides some indication of greater run-to-run response instability for non-binary personas.

The results on a Hierarchy of Personas are consistent with a study by Ma et al. (2025) showing that persona-conditioned LLM responses are not equally influenced by all persona dimensions, highlighting the influence of party affiliation and education compared to other dimensions, such as age. One possible explanation for our results is that party affiliation is most closely linked to ideological, cultural, and policy-relevant information, especially on normatively charged issue domains, such as LGBTQ-related attitudes. In contrast to demographic persona dimensions, such as age and gender, party affiliation is directly related to

normative orientations, issue positions, and policy preferences. Particularly, prior research with human survey data has shown that party affiliation, or political identity more generally, is a strong predictor of LGBTQ-related attitudes, with substantial differences between left-libertarian, social-democratic, conservative, and right-wing voters (Szalma & Takács, 2025). When such party affiliations are inserted into a persona prompt, they may activate ideologically saturated and already existing evaluative associations, resulting in a strong influence on LLM responses.

Table 5. Run-to-run response stability by item and gender persona category

	Overall		Male		Female		Non-binary	
Item	SD	Range	SD	Range	SD	Range	SD	Range
1	0.03	0.07	0.01	0.02	0.05	0.09	0.04	0.09
2	0.06	0.13	0.06	0.13	0.06	0.13	0.07	0.13
3	0.18	0.37	0.15	0.32	0.15	0.32	0.23	0.48
4	0.05	0.10	0.07	0.13	0.05	0.09	0.04	0.07

Note. SD = mean within-persona standard deviation across five runs. Range = mean within-persona response range, defined as the maximum response minus minimum response across five runs. Both indicators are computed on the harmonized response scale (item 3 reverse-coded). Reported values are averaged across personas within each item and gender category. n = 162 persona profiles per item overall; n = 54 personas per gender category and item.

Similarly, in existing studies on human LGBTQ-related attitudes, higher education is repeatedly associated with more supportive attitudes towards sexual minorities and LGBTQ-related rights, often because it correlates with liberal values and greater exposure to diversity (Donaldson et al., 2017; Szalma & Takács, 2025). In the context of our article, however, education should not be interpreted as functioning analogously to these real-world mechanisms. Rather, the comparatively high influence of education may suggest that Gemma-3 associates educational status with distinct discursive and normative positions. From this perspective, education appears to operate not as a simple sociodemographic attribute within the LLM's response space, but as a socially meaningful dimension linked to specific LGTBQ-related attitudes that directly influence response behavior. Interestingly, education surpasses party affiliation in its relative influence when it comes to item 3. This may suggest that education becomes more influential when the item topic shifts from more normative acceptance and evaluations of equal rights to affective LGBTQ acceptance. However, since we used only one item on affective acceptance, this implication should be interpreted cautiously and may also reflect measurement sensitivity related to the item's negative wording and format. We therefore encourage future studies to extend our investigation in this direction providing further, more robust findings.

The weak influence of age and gender may indicate that they activate broader social information that are less tightly linked to LGBTQ-related attitudes, whereas party affiliation and education are directly connected to attitude patterns in the LLM's learned response space. Another possible interpretation is that information associated with age and gender is effectively overshadowed once a highly informative and politically saturated persona dimension, such as party affiliation, is present in the prompt.

A further nuance to the interpretation of the observed hierarchy is represented by item 3 that differs systematically from the remaining items. Specifically, it shows the highest run-to-run variation, the greatest between-persona variation, and it is the only item, where education exceeds party affiliation in terms of relative influence. This suggests that persona-conditioned response behavior depends not only on the assigned persona profile, but also on the item through which the synthetic responses are generated. In other words, the Hierarchy of Personas appears to be measurement-sensitive.

Furthermore, the results on gender are striking, as the current literature suggests that non-majority-coded personas, such as non-binary gender, tend to produce not only more stereotyped language and lower linguistic diversity, but also less robust persona effects (Cheng et al., 2023; Lutz et al., 2025; Wang et al., 2025). Our results are only partially in line with this literature. While item 3 provides some indication of greater response instability for non-binary personas, this does not apply to the remaining items. Thus, non-majority-coded personas, such as non-binary gender, do not necessarily seem to translate into strong response variation in every survey-like setting. Our results also strengthen the importance of distinguishing between response formats. A persona dimension may influence linguistic style, diversity, and the activation of stereotypes in open-ended responses, without necessarily becoming a dominant influence on closed-ended responses.

Implications for survey research

Overall, our results have important implications for the methodological debate on silicon sampling, algorithmic fidelity, and machine bias, especially in the context of survey research. Most importantly, the observed Hierarchy of Personas indicates that Gemma-3 does not reproduce a "real" distribution of LGBTQ-related attitudes for specific persona profiles, since the persona-conditioned responses are unevenly responsive to different persona dimensions. Therefore, our article calls for a critical reflection on silicon sampling and persona prompting as well as their appropriateness for simulating opinions and response distributions, especially in socially and politically sensitive domains. Even if persona-conditioned responses display structured variation and the influence of an analytically recoverable Hierarchy of Personas, this should not be confused with access to lived experiences or authentic group perspectives.

We urge future researchers to approach synthetic respondents not as substitutes for human respondents but as simulations that are highly dependent on the LLM employed and whose responses reflect training data, alignment procedures, and prompt design (Argyle et al., 2023; Boelaert et al., 2025; Ma et al., 2025). As LLMs do not respond equally to all persona dimensions, they do not simulate the entire social profile of a synthetic respondent, but rather selected components of it. Any observable group variation in a silicon sample produced through persona prompting may therefore reflect differential responsiveness to persona dimensions rather than actual group differences. In this sense, the Hierarchy of Personas should be treated as a relevant methodological consideration when evaluating silicon sampling based on persona prompting approaches.

A further implication concerns the observation that persona prompting cannot be assessed independently of persona and item wording, response format, and simulation setup (Ahnert et al., 2025; Rupprecht et al., 2025). The deviation of item 3 shows that the influence of persona dimensions is not stable across items, even when the LLM, prompt, and persona dimensions

are held constant. For future LLM-based survey research, this implies that the overall study design should be rigorously documented to enhance transparency and reproducibility. Relatedly, studies using synthetic respondents should conduct robustness checks to account for alternative item wordings, persona formulations, scale formats, and prompting conditions.

Limitations and contributions

Although our article provides novel insights on persona prompting in the context of survey research, it has several limitations, opening avenues for future research. First, we examined only one LLM, namely Gemma-3, in the instruction-tuned twelve-billion-parameter version. The observed Hierarchy of Personas may not generalize to other LLMs, model sizes, or alignment procedures (e.g., RLHF-aligned versus non-RLHF-aligned models). Thus, we encourage future studies to examine the influence of persona dimensions between different LLMs, including varying parameter settings. Second, and related to the previous point, our results may depend on the persona dimensions, prompt design, and response format employed. It would be worthwhile replicating our analyses with different persona dimensions (e.g., personality traits), prompt formulations (e.g., demographic priming with names or titles), and alternative response formats (e.g., open-ended questions) to examine potential changes to the Hierarchy of Personas. Third, we used four items from well-established social surveys dealing with LGBTQ-related attitudes. While this design decision highlights the ecological validity of the set of survey items used in this study, it also partially limits generalizability of our findings to other issue domains (or topics). Future studies thus may examine the Hierarchy of Personas in the context of other topics, such as xenophobia. Fourth, the influence of item wording and content on Gemma-3's responses cannot be disentangled in the present study design. As elaborated previously, the deviating response behavior with respect to item 3 could reflect the item's negative wording, its topic (i.e., affective acceptance), or the interaction of both features. Thus, further research is needed to better disentangle the influence of item features by, for example, examining alternative item wordings.

Another point is that we do not benchmark model responses against human survey data, as it is frequently done in other studies (see Argyle et al., 2023; Bisbee et al., 2024; Ma et al., 2025; Qi et al., 2025). However, the aim of this study is not to evaluate the "empirical accuracy" of LLM-generated responses to LGBTQ-related survey items, but rather to examine the model's internal responsiveness to different persona dimensions within a controlled full-factorial design. Because comparable human benchmark survey data covering all persona combinations are not available to us, such a comparison would be difficult to implement and would address a completely different research question.

Taken together, the key contribution of our article lies in demonstrating that persona prompting reveals structured and interpretable variation in LLM response behavior, and that this variation follows a Hierarchy of Personas. In the context of synthetic respondents or silicon sampling, the key issue is thus not simply whether persona prompting matters, but which persona dimensions matter most and under which conditions. Based on this reasoning, our article highlights that survey research methods can be utilized to shed light on how LLMs generate responses to sensitive topics. Although silicon samples and persona prompting are unlikely to replace human respondents in survey research, they can contribute to important societal debates about model bias, prompt sensitivity, and response behavior of LLMs.

References

Ahnert, G., Haensch, A.-C., Plank, B., & Strohmaier, M. (2025). *Survey response generation: Generating closed-ended survey responses in-silico with large language models.* arXiv. https://doi.org/10.48550/arXiv.2510.11586

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337-351. https://doi.org/10.1017/pan.2023.2

Boelaert, J., Coavoux, S., Ollion, É., Petev, I., & Präg, P. (2025). Machine bias. How do generative language models answer opinion polls? *Sociological Methods & Research, 54*(3), 1156-1196. https://doi.org/10.1177/00491241251330582

Beck, T., Schuff, H., Lauscher, A., & Gurevych, I. (2024). Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In Y. Graham, & M. Purver (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2589-2615). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-long.159

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FaccT '21)* (pp. 610-623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis, 32*(4), 401-416. https://doi.org/10.1017/pan.2024.5

Die Bundeswahlleiterin (2025). Ergebnisse früherer Bundestagswahlen. https://www.bundeswahlleiterin.de/dam/jcr/397735e3-0585-46f6-a0b5-2c60c5b83de6/btw_ab49_gesamt.pdf

Donaldson, C. D., Handren, L. M., & Lac, A. (2017). Applying multilevel modeling to understand individual and cross-cultural variations in attitudes toward homosexual people across 28 European countries. *Journal of Cross-Cultural Psychology*, *48*(1), 93-112. https://doi.org/10.1177/0022022116672488

Cheng, M., Durmus, E., & Jurafsky, D. (2023). Marked personas: Using natural language prompts to measure stereotypes in language models. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1504-1532). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.84

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants?. *Trends in Cognitive Sciences, 27*(7), 597-600. https://doi.org/10.1016/j.tics.2023.04.008

European Commission: Directorate-General for Justice and Consumers & Kantar. (2023). *Discrimination in the European Union: Report*. European Commission. https://data.europa.eu/doi/10.2838/936462.

European Social Survey (ESS) (2022). *ESS Round 11 Source Questionnaire*. London: ESS ERIC Headquarters c/o City, University of London.

Gemma Team (2025). *Gemma 3 Technical Report.* arXiv. https://doi.org/10.48550/arXiv.2503.19786

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. *Science*, *380*, 1108-1109. https://doi.org/10.1126/science.adi1778

Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., & Khot, T. (2023). *Bias runs deep: Implicit reasoning biases in persona-assigned LLMs.* arXiv. https://doi.org/10.48550/arXiv.2311.04892

Harding, J., D'Alessandro, W., Laskowski, N. G., & Long, R. (2024). AI language models cannot replace human research participants. *AI & Society, 39*(5), 2603-2605. https://doi.org/10.1007/s00146-023-01725-x

Höhne, J. K., Claassen, J., Shahania, S., & Broneske, D. (2025a). Bots in web survey interviews: A showcase. *International Journal of Market Research*, *67*(1), 3-12. https://journals.sagepub.com/doi/10.1177/14707853241297009

Höhne, J. K., Claassen, J., & Wolf, B. L. (2025b). LLM-driven bot infiltration: protecting web surveys through prompt injections. *International Journal of Social Research Methodology*. https://doi.org/10.1080/13645579.2025.2598606

Ji, J., Kim, J., & Kim, Y. (2024). Predicting missing values in survey data using prompt engineering for addressing item non-response. *Future Internet*, *16*(10), 351. https://doi.org/10.3390/fi16100351

Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., & Kabbara, J. (2024). PersonaLLM: Investigating the ability of large language models to express personality traits. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 3605-3627). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.229

Li, A., Chen, H., Namkoong, H., & Peng, T. (2025). *LLM generated persona is a promise with a catch.* arXiv. https://doi.org/10.48550/arXiv.2503.16527

Lutz, M., Sen, I., Ahnert, G., Rogers, E., & Strohmaier, M. (2025). The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 23212-23237). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-emnlp.1261

Ma, B., Wang, X., Hu, T., Haensch, A. C., Hedderich, M. A., Plank, B., & Kreuter, F. (2024). The potential and challenges of evaluating attitudes, opinions, and values in large language models. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 8783-8805). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.513

Ma, B., Yoztyurk, B., Haensch, A. C., Wang, X., Herklotz, M., Kreuter, F., Plank, B., & Assenmacher, M. (2025). Algorithmic fidelity of large language models in generating synthetic German public opinions: A case study. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (pp. 1785-1809). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.90

Ollama. (2024). *Ollama python library*. GitHub. https://github.com/ollama/ollama-python

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. & Lowe, R. (2022). *Training language models to follow instructions with human feedback*. arXiv. https://doi.org/10.48550/arXiv.2203.02155

Plaza-del-Arco, F. M., Röttger, P., Scherrer, N., Borgonovo, E., Plischke, E., & Hovy, D. (2025). No for some, yes for others: Persona prompts and other sources of false refusal in language models. In C. Zhang, E. Allaway, H. Shen, L. Miculicich, Y. Li, M. M'hamdi, P. Limkonchotiwat, R. H. Bai, S. T.y.s.s., S. S. Han, S. Thapa, & W. B. Rim (Eds.), *Proceedings of the 9th Widening NLP Workshop* (pp. 268-282). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.winlp-main.39

Posit Team (2025). RStudio: Integrated development environment for R. Posit Software, Boston, MA. http://posit.co/

Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H., Schuetze, H., & Hovy, D. (2024). Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15295-15311). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.816

Rupprecht, J., Ahnert, G., & Strohmaier, M. (2025). *Prompt perturbations reveal human-like biases in Large Language Model survey responses*. arXiv. https://doi.org/10.48550/arXiv.2507.07188

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect?. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (pp. 29971-30004). JMLR.org https://dl.acm.org/doi/abs/10.5555/3618408.3619652

Szalma, I., & Takács, J. (2025). The impact of political-demographic considerations on European attitudes towards parenting and adoption by same-sex couples. *European Journal of Politics and Gender*, 1-27. https://doi.org/10.1332/25151088Y2024D000000072

Tseng, Y. M., Huang, Y. C., Hsiao, T. Y., Chen, W. L., Huang, C. W., Meng, Y., & Chen, Y. N. (2024). Two tales of persona in LLMs: A survey of role-playing and personalization. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 16612-16631). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.969

Qi, W., Lyu, H., & Luo, J. (2025). Representation bias in political sample simulations with large language models. In *Companion Proceedings of the ACM on Web Conference 2025 (WWW '25)* (pp. 1264-1267). Association for Computing Machinery. https://doi.org/10.1145/3701716.3715591

Von Der Heyde, L., Haensch, A. C., & Wenz, A. (2025). Vox populi, vox AI? Using large language models to estimate German vote choice. *Social Science Computer Review*. https://doi.org/10.1177/08944393251337014

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, *7*(3), 400-411. https://doi.org/10.1038/s42256-025-00986-z

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., & Gabriel, I. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on Fairness, Accountability, and Transparency* (FAccT '22) (pp. 214-229). Association for Computing Machinery. https://doi.org/10.1145/3531146.3533088

Appendix 1

Close German translations/prompt wordings of ESS/Eurobarometer items

Item	Original wording	German translation/Prompt wordings
EB1	Lesbian, gay and bisexual people should have the same rights as heterosexual people (marriage, adoption, parental rights).	Lesbische, schwule und bisexuelle Personen sollten die gleichen Rechte haben wie heterosexuelle Personen (z.B. Eheschließung, Adoption, Elternrechte).
EB2	Transgender people should have the same rights as anyone else (marriage, adoption, parental rights).	Transgender-Personen sollten die gleichen Rechte haben wie heterosexuelle Personen (z.B. Eheschließung, Adoption, Elternrechte).
ESS1	If a close family member was a gay man or a lesbian, I would feel ashamed.	Wenn ein nahes Familienmitglied ein schwuler Mann oder eine lesbische Frau wäre, würde ich mich schämen.
ESS2	Gay men and lesbians should be free to live their own life as they wish.	Schwule Männer und lesbische Frauen sollten frei sein, ihr Leben so zu leben, wie sie es wünschen.

Note. The German prompt wording of item 2 was slightly adapted for comparability with item 1. While the original source item refers to equal rights "as anyone else" the prompt wording used here refers to equal rights "wie heterosexuelle Personen", closely mirroring the wording of item 1. This was intended to keep the two equal-rights items more parallel in wording so that they differ primarily in the addressed target group. Since the thesis does not aim to replicate any human survey distribution but to systematically investigate persona-conditioned response behavior, minor wording adaptations were treated as a pragmatic way to maximize comparability within the fixed prompting setup.

Appendix 2

Prompt wordings and English translations of *{scale_blocks}* and *{answer_lines}* used in prompt template for data collection

	Prompt wordings		English translations	
Source survey/scales	{scale_block}	{answer_line}	{scale_block}	{answer_line}
European Social Survey/ESS5	"Die Antwort erfolgt auf einer fünfstufigen Likert-Skala: 1 = stimme voll und ganz zu 2 = stimme eher zu 3 = teils/teils 4 = stimme eher nicht zu 5 = stimme überhaupt nicht zu"	"Zeile 1: NUR die Ziffer 1, 2, 3, 4 oder 5."	„Responses are given on a five-point Likert-type scale: 1 = strongly agree 2= somewhat agree 3 = neither agree nor disagree 4 = somewhat disagree 5 = strongly disagree"	"Line 1: ONLY the digit 1, 2, 3, 4, or 5."
Eurobarometer/EB4	„Die Antwort erfolgt auf einer vierstufigen Likert-Skala: 1 = stimme voll und ganz zu 2 = stimme eher zu 3 = stimme eher nicht zu 4 = stimme überhaupt nicht zu"	"Zeile 1: NUR die Ziffer 1, 2, 3 oder 4."	„Responses are given on a four-point Likert-type scale: 1 = strongly agree 2= somewhat agree 3 = somewhat disagree 4 = strongly disagree"	„Line 1: ONLY the digit 1, 2, 3, or 4."

Appendix 3

Technical details of the macOS workstation used for data collection:

- Model: Apple Mac mini (M1, 2020)
- Chip: Apple M1 (8-core CPU, 8-core GPU)
- Memory: 16 GB unified memory
- Storage: 256 GB SSD
- Operating System: macOS Sequoia [Version 15.6.1]

Appendix 4

English translation of the prompt

Act as a German-speaking {gender} person {age}, with {education}, and an affinity for the {party} who is participating in a survey. Respond from this persona's perspective to the following statement by indicating your level of agreement with the STATEMENT itself (not with the topic in general): {item} {scale_block} Response format (important): Line 1: {answer_line} Line 2: Reasoning: 1–2 short sentences explaining, from this persona's perspective, why this level of agreement was chosen for THIS specific statement. No further output.

Appendix 5

Run-to-run stability by item and all persona categories

	Item 1		Item 2		Item 3		Item 4	
	SD	Range	SD	Range	SD	Range	SD	Range
Age (n = 54 per category)								
Young adults	0.06	0.11	0.04	0.07	0.13	0.26	0.04	0.07
Middle-aged adults	0.03	0.06	0.07	0.15	0.22	0.46	0.06	0.11
Older adults	0.02	0.04	0.08	0.17	0.18	0.39	0.06	0.11
Gender (n = 54 per category)								
Male	0.01	0.02	0.06	0.13	0.15	0.32	0.07	0.13
Female	0.05	0.09	0.06	0.13	0.15	0.32	0.05	0.09
Non-binary	0.04	0.09	0.07	0.13	0.23	0.48	0.04	0.07
Education (n = 54 per category)								
Low education	0.07	0.13	0.07	0.13	0.06	0.11	0.08	0.17
Medium education	0.02	0.06	0.05	0.09	0.14	0.30	0.03	0.06
High education	0.01	0.02	0.08	0.17	0.33	0.70	0.04	0.07
Party affiliation (n = 27 per category)								
Alternative für Deutschland (AfD)	0.18	0.37	0.18	0.37	0.00	0.00	0.21	0.41
Bündnis 90/Die Grünen	0.00	0.00	0.00	0.00	0.34	0.70	0.00	0.00
Christlich-Demokratische Union (CDU)	0.02	0.04	0.09	0.19	0.18	0.37	0.08	0.15
Die Linke	0.00	0.00	0.02	0.04	0.22	0.52	0.00	0.00
Freie Demokratische Partei (FDP)	0.00	0.00	0.08	0.15	0.14	0.30	0.02	0.04
Sozialdemokratische Partei Deutschlands (SPD)	0.00	0.00	0.02	0.04	0.16	0.33	0.00	0.00

Note. SD = mean within-persona standard deviation across five runs. Range = mean within-persona response range, defined as the maximum response minus minimum response across five runs. Both indicators are computed on the harmonized response scale (item 3 reverse-coded). Reported values are averaged across personas within each item and persona category.