German General Social Survey Personas: A Survey-Derived Persona Prompt Collection for Population-Aligned LLM Studies

Jens Rupprecht[1,*], Leon Fröhling[2,*], Claudia Wagner[2,3,4], Markus Strohmaier[1,2,4]
[1]University of Mannheim, [2]GESIS – Leibniz Institute for the Social Sciences,
[3]RWTH Aachen University, [4]Complexity Science Hub

Abstract

The use of Large Language Models (LLMs) for simulating human perspectives via persona prompting is gaining traction in computational social science. However, well-curated, empirically grounded persona collections remain scarce, limiting the accuracy and representativeness of such simulations. Here, we introduce the *German General Social Survey Personas* (*GGSS Personas*) collection, a *comprehensive* and *representative* persona prompt collection built from the German General Social Survey (ALLBUS). The *GGSS Personas* and their persona prompts are designed to be easily plugged into prompts for all types of LLMs and tasks, steering models to generate responses aligned with the underlying German population. We evaluate *GGSS Personas* by prompting various LLMs to simulate survey response distributions across diverse topics, demonstrating that *GGSS Personas*-guided LLMs outperform state-of-the-art classifiers, particularly under data scarcity. Furthermore, we analyze how the representativity and attribute selection within persona prompts affect alignment with population responses. Our findings suggest that *GGSS Personas* provide a potentially valuable resource for research on LLM-based social simulations that enables more systematic explorations of population-aligned persona prompting in NLP and social science research.

Keywords: Corpus (Creation, Annotation, etc.), Language Representation Models, Profiling, Bias, Safety

1. Introduction

Simulating human behavior represents a complex challenge given the vast diversity of human experiences and perspectives. Traditional research methods, such as large-scale question-oriented surveys, are crucial for informing social science research and policy decisions, but they are often costly and time-consuming to administer. As a result, researchers are increasingly turning to innovative approaches to capture the nuances of human behavior. One recent avenue is the use of large language models (LLMs) and so-called persona prompting (Chen et al., 2024; Tseng et al., 2024; Lutz et al., 2025; Zhang et al., 2025), i.e., the use of personas descriptions to induce role-playing in LLMs, as an accessible option to steer and control simulated human behavior in surveys (Ma et al., 2025; Miranda and Balbi, 2025).

Persona prompting refers to a broad range of prompting techniques that aim to harness the general world knowledge captured by LLMs and steer the generation of simulated survey responses toward the perspective of the persona introduced in the prompt. Compared to alternative approaches for the alignment of LLMs, such as training dedicated models (Feng et al., 2023) or fine-tuning existing ones (Orlikowski et al., 2025; Suh et al., 2025), prompt-based approaches offer a number of crucial advantages: they are *modular* (i.e., they can be used with different models and for different tasks), *upgradable* (i.e., they can be modified and can be expected to improve in performance with the release of improved models), and *shareable* (i.e., the prompts can easily be made available to other potential uses and users), among other advantages.

German General Social Survey Personas are a prompt collection that consists of $5,246$ personas constructed from a representative sample of the German population. By grounding the *German General Social Survey Personas* collection in the information available in the ALLBUS (GESIS, 2025a)—the German General Social Survey and thus the premier source of information on the attributes and attitudes of the German population—we demonstrate that more systematic and principled approaches to the construction of persona prompt collections are possible [1].

GGSS Personas offers various potential advantages: First, personas introduce relevant contextual information to the model, providing it with the necessary information to meaningfully anchor predictions for a task or target variable in empirically observed associations and connections. Second, the ALLBUS is a probability-based survey and the personas derived from it aim to represent the German population. While there has been a focus on the biased representation of populations that LLMs default towards in their generation of survey responses (Santurkar et al., 2023; Durmus et al., 2024; Sen et al., 2025), the *GGSS Personas* may potentially help to align LLMs better with the German population.

* Equal contribution.

[1]German General Social Survey Personas are available after registration: https://doi.org/10.4232/1.14707

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1761–1780
11–16 May 2026.

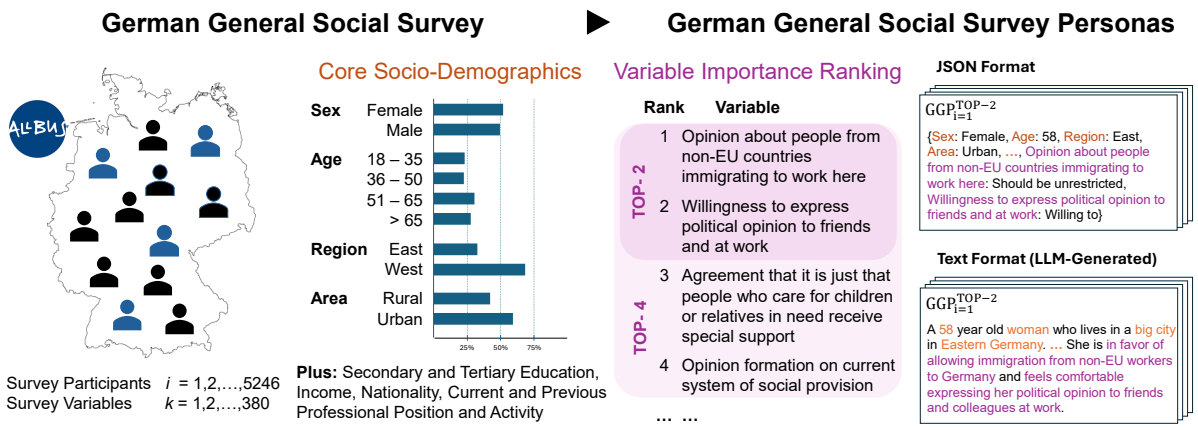

Figure 1: **Grounding the *German General Social Survey Personas* (*GGSS Personas*) in the ALLBUS survey.** We construct individual persona prompts for each ALLBUS participant, varying size and composition via available attributes. A global variable importance ranking informs the selection of the k most important attributes (TOP-k). Personas comprise a fixed block of core socio-demographics and a more extensible block of TOP-k attributes that allow varying information content. The *GGSS Personas* is available in JSON- and full-text formats. The originally German survey items and personas are presented in English for illustration.

As a main contribution, this paper introduces *German General Social Survey Personas*—a novel textual resource for NLP and Computational Social Science (CSS). We describe the creation process of *GGSS Personas* and offer a preliminary analysis of its potential and utility through two exemplary demonstrations.

2. Related Work

Persona prompting is a **versatile and accessible method for social simulations** and pluralistic alignment with LLMs (Anthis et al., 2025; Sorensen et al., 2024). There is already a wide range of applications and promising demonstrations across disciplines and tasks, including (but not limited to): vote choice prediction (Argyle et al., 2023; von der Heyde et al., 2025), generation of synthetic public opinions (Hwang et al., 2023; Ma et al., 2025; Miranda and Balbi, 2025), annotation of (subjective) constructs (Hu and Collier, 2024; Beck et al., 2024; Fröhling et al., 2025), simulations of participants in economic and social science experiments (Aher et al., 2023; Hewitt et al., 2024), including for underrepresented and otherwise hard to reach populations (Gonzalez-Bonorino et al., 2025), simulation of pluralistic debates (Ashkinaze et al., 2025), behavioral simulation of users (Chen et al., 2025b; Salem et al., 2025), and even productive real world tasks such as red teaming (Deng et al., 2025). Systematic reviews of the literature on the use of LLMs for role-playing and persona prompting provide an overview of this dynamic area of research (Lutz et al., 2025; Zhang et al., 2025; Chen et al., 2024; Tseng et al., 2024).

Beyond purely prompt-based approaches, Kim and Lee (2023) propose a framework that fine-tunes LLMs on survey data for predicting responses at an individual level, and Cao et al. (2025) and Suh et al. (2025) use group-level information to fine-tune LLMs for the simulation of response distributions. Orlikowski et al. (2025) compare the use of personas for (zero-shot) persona prompting with fine-tuning LLMs on pairs of annotators' persona descriptions and their annotations, finding that fine-tuning strictly outperforms zero-shot in simulating individual-level annotations. Apart from fine-tuning, Chen et al. (2025a) introduce the idea of persona vectors, extracted via persona descriptions and used to monitor and steer LLM generations. Sun et al. (2025) propose an approach that leverages personas via retrieval-augmentation to personalize LLM responses.

Conceptual criticism of the use of persona prompting include the work by Cheng et al. (2023a,b), who show that simulations of political and marginalized groups tend towards caricature, and that LLM-generated personas of non-white, non-male demographics exhibit patterns of othering and exoticization. Similarly, Wang et al. (2025) show how LLMs are likely to misportray and flatten the representation of demographic groups. Kim et al. (2024) investigate how the use of personas might degrade performance by distracting the model from the task at hand. Kirk et al. (2024) discuss the broader context and societal implications of aligning LLMs with individuals.

In the literature, there are some noteworthy demonstrations of how **well-curated, empirically-grounded persona collections** lead to improved

performance in their applications. Park et al. (2024) show how the use of transcripts of hour-long interviews with participants, i.e., extremely information-rich persona context, leads to impressive LLM performance in predicting individuals' survey responses. Similarly, Toubia et al. (2025) survey a representative sample of US-based crowdworkers to collect extensive information on their demographic-, psychological-, economic-, personality-, and cognitive measures, and Peng et al. (2025) show that the use of this detailed personal information improves the correlations between human and LLM-simulated responses.

However, this quality comes at the price of up to $\$100$ paid to $1,052$ participants for hour-long interviews and $\$37$ paid to $2,058$ crowdworkers for answering 500 survey questions, respectively. Using **freely and publicly available survey data** offers an accessible alternative for grounding persona collections in high-quality empirical information. Most similar to our work, Castricato et al. (2025) introduce PERSONA, a collection of $1,586$ synthetic personas created from US census data. In contrast to their approach, our *GGSS Personas* establishes 1:1 correspondence between survey respondents and persona prompts, and automatically identifies important persona attributes from the full range of survey variables instead of limiting selection to a set of 31 variables.

With the introduction of *GGSS Personas*, we aim to contribute a well-curated, empirically-grounded persona collection as a crucial resource for meaningful advances in research on persona-based LLM personalization.

3. *German General Social Survey Personas*

German General Social Survey Personas is derived from the ALLBUS, the German General Social Survey (Figure 1). While the *GGSS Personas* is necessarily tied to the German population at a specific moment in time, we propose a creation process that can be applied to different (population) surveys in order to create survey-derived persona collections for different contexts and purposes.

Grounding *German General Social Survey Personas* Ideally, a population survey panel used for the construction of persona prompt collections should cover a set of core demographics and additional attributes for a representative sample drawn from the population of interest. Following the discussion on the use of the terms *representative sample* and *representativity* in the scientific literature by Kruskal and Mosteller (1979), we believe that population-level general social surveys such as the ALLBUS in Germany, the General Social Survey (GSS) in the US, the British Social Attitudes Survey (BSA) in the UK, or the European Social Survey (ESS) covering 31 European Countries in its latest round, could represent plausible empirical foundations for representative persona prompt collections.

TOP-k	Average Attributes	Full Personas
2	1.34	3374
4	2.68	1848
8	5.09	821
16	10.52	0
32	21.28	0
64	38.80	0
128	77.75	0
256	157.54	0
380	242.27	0

Table 1: **Completeness of persona descriptions.** While the TOP-k value in each row represents the maximum possible number of attributes featured in the collection's personas, the actually included average number is always lower due to some variables being unavailable for each survey respondent.

To create this first version of the *GGSS Personas*, we use the latest release of the biannual ALLBUS (GESIS, 2025a), released in January 2025 and including responses collected between April and September 2023. It features $5,246$ participants and a total of 605 variables. We work with the ALLBUScompact [2] (GESIS, 2025b), the openly available version of the ALLBUS, featuring 579 variables. For data protection reasons, the ALLBUScompact contains information in reduced detail on the sociodemographic characteristics age, origin, occupation and income, and omits all fine-grained geographic variables.

In our preliminary analysis, we explore the effects of using a representative persona collection such as the *GGSS Personas*—representative of the German population by virtue of mirroring all ALLBUS participants—as well as unrepresentative variants of it, created by systematically oversampling on selected attributes (Section 4).

Attribute Selection Besides the choice for the population survey that serves as the fundament of a representative persona prompt collection, the second relevant selection concerns the set of survey variables that are included in the persona descriptions. These persona attributes should be chosen in a way that they provide the (most) relevant context to the LLM for it to make an informed prediction of the individual's attitude or behavior for a given task.

While it is theoretically feasible to use all 579 variables featured in the latest release of the ALL-

[2] https://search.gesis.org/research_data/ZA8832

BUScompact in the persona prompts, this would unavoidably come at the price of increased context window requirements, computational resources, as well as potentially degrading performance due to the dilution of relevant information or the failure of the model to put equal attention to all variables in a lengthy persona description. For example, Shi et al. (2023) have shown that LLMs can easily be distracted by irrelevant information included in the context window, and de Araujo et al. (2025) have observed performance drops after including irrelevant persona details in persona prompts.

One of the core advantages of *general* population surveys such as the ALLBUS is the coverage of a variety of different topics, ranging from core demographic variables such as age and gender to very specific, thematic variables covering, e.g., the participant's political views or lifestyle choices. Therefore, we design a data-driven method that allows us to identify and select the statistically most important variables, instead of having to (artificially) restrict the *GGSS Personas* to cover only task-specific areas such as socio-demographics, psychological inventories, or political beliefs.

From the 579 available survey variables, we exclude those that are purely technical interview paradata (e.g., information on the interviewer), those that we identify as core socio-demographic variables and include in all persona descriptions (see Figure 1), and those that we randomly select as outcome variables (see Section 4). For each of the 406 remaining variables, we fit a Random Forest Classifier, using all other 405 variables as predictors. For each fitted model, we identify the ten most important predictors using their feature importance scores, and aggregate those scores for each predictor to create a variable importance ranking. This ranking features 380 variables after dropping all variables that never appeared among the top ten most important features for any fitted model.

We expect that variables with higher aggregate feature importance score explain more variance of the other variables in the survey. This ranking thus allows us to create versions of the *GGSS Personas* collection for different computational budgets and needs by selecting only the k most important attributes (*TOP-k*). In situations with restricted prompt length or inference time, versions with less attributes may be preferable, while in unrestricted settings, the number of included attributes may be increased, up to the inclusion of all 380 survey variables. In our preliminary analysis, we explore the effects of increasing the number of persona attributes on the ability of LLMs to accurately simulate survey response distributions (Section 4).

An important limitation of our approach is its inability to deal with missing values. Because large surveys such as the ALLBUS work with questionnaire splits to manage the number of questions each participant is asked to answer, no participant will have answered all survey questions. Thus, some information is just not available for some of the participants, meaning that for any given variable some persona descriptions will always be incomplete, missing individual attributes. Table 1 shows the extent to which the persona descriptions across the different TOP-k collections are complete. The k thus represents the maximum number of attributes possibly featured in a persona of a given TOP-k collection, and the reported average is the actual number of attributes included. We discuss the implications for the *GGSS Personas* and our preliminary analysis in Section 6.

Persona Generation In the final step of our pipeline, we turn the persona attributes into a format that can easily be plugged into any LLM prompt to explicitly set the persona perspective that the model is supposed to take on during response generation.

There is a broad range of different prompting techniques in the literature (Lutz et al., 2025), ranging from the use of key-value pairs in a dictionary or **JSON-format** (e.g., Castricato et al. (2025)) to the use of **full-text** persona descriptions (e.g., Ge et al. (2024)), oftentimes generated using LLMs. While personas in the JSON-format are easier to construct and more convenient to work with, the full-text persona descriptions resemble the natural language interactions that especially instruction-tuned LLMs have been optimized for more closely.

We evaluate the impact of using different persona formats for predicting survey response distributions. Since we do not observe a large impact of the persona format on the experimental results, (see Appendix A Figure 7) we use the JSON-format for all following experiments, expecting the results to generalize to other persona formats. However, we make both the JSON as well as the full-text version of the *GGSS Personas* publicly available. We validate the full-text personas that are LLM-generated from the JSON-personas as input. Two co-authors of this paper independently annotated a sample of 80 personas (20 randomly-sampled personas per TOP-2, -4, -8 and -16 collection) for hallucinations, misrepresentation or omission of individual attributes. We find that 69 of 80 (or 86.25%) full-text personas are accurate depictions of the JSON-personas, with errors occurring rarely through misrepresentation, e.g., of double negations or complex temporal orders, or through the omission of single attributes. We further describe the human validation process in Appendix C.1.

4. Preliminary Analysis

4.1. Experimental Setup

Possible use cases and tasks of representative persona collections are plentiful. In this paper we focus on evaluating the utility of *GGSS Personas* for simulating the response distribution of survey participants across a diverse set of outcome variables.

Outcome Variables We randomly sample 27 survey variables across a range of different topics—three variables from each of the nine different topic areas covered in the ALLBUS—serving as outcome variables for which we simulate response distributions. We use the persona prompts as input and instruct the model to predict the response of the corresponding individual for the outcome variable. As described above, the outcome variables have been excluded from the set of persona attributes used to construct the persona prompts.

For the 2023 version of the ALLBUS, the *key topics* defined in the documentation are: *Lifestyle*, *Social Inequality*, *Religion*, *Ethnocentrism*, and *Political Tendency*. We construct four additional topics from the large set of remaining variables and their categorization: *Values & Life Goals*, *Economic Situation*, *Social Capital*, and *Morality*. Appendix A Table 4 provides details on the 27 sampled outcome variables.

Model Selection For generalizability, we selected five open weights, instruction-tuned LLMs, varying in size, developer, and origin. For some results, we average the performance of Mistral-7B, Llama-3.1-8B, and Qwen3-8B, referring to them as the *7/8B LLMs* models. In comparison, we consider Gemma-3-12B-it and Llama-3.3-70B to be medium and large models, respectively.

We document the release and knowledge cutoff dates of the selected LLMs (see Appendix B.1 Table 2) showing that there is no issue with potential data leakage or contamination of the training data. All known knowledge cutoff dates fall before the first release of the ALLBUScompact (10/01/2025). The only model with an unknown knowledge cut-off date released after the release of the ALLBUScompact is Qwen3-8B, released 27/04/2025. However, given the usual gap of multiple months between the beginning of its training process (the relevant point in time for the knowledge cutoff) and the release of a model, we can be (almost) certain that none of the models we are using in this work could have had access to any of the ALLBUS data during training.

Baseline Method As baselines for the survey response distribution prediction task, we fit random forest classifiers on training datasets of varying size and with varying numbers of persona attributes used as input features. We use the predictions of the random forest on an individual level and aggregate them across all personas to generate a simulated survey response distribution.

This choice of baseline puts the persona-based zero-shot LLM approach at a disadvantage. While our approach always acts on a single persona description as input, the random forest classifiers are fitted on training samples representing persona descriptions and their corresponding responses for the respective prediction tasks, with training dataset sizes systematically ranging from $n = 2$ to $n = 2,048$. We reserve 20% of respondents as fixed test set, and randomly select n of the remaining respondents for training the baseline classifiers. The setting of classifier hyperparameters follows Miranda and Balbi (2025) and is reported in Appendix C. We compare the classifier predictions with the groundtruth response distribution of the test set respondents. The LLMs do not have access to any training samples.

Similarly, we vary the number of persona attributes k used as input features, systematically ranging from $k = 2$ to $k = 380$. For every k, we construct the set of persona attributes by selecting the k most important attributes according to the variable importance ranking presented in Section 3.

Evaluation We measure the performance of our approach as well as of the different baselines by calculating the Jensen-Shannon Distance (JSD) between the predicted and the actual response distribution of each sampled outcome variable. The predicted response distribution is created by aggregating the model predictions for a given variable across all participants, and the actual response distribution is the aggregate of the participants' responses found in the ALLBUS. This procedure is in line with the suggestion offered by Sorensen et al. (2024) for the evaluation of distributionally pluralistic models.

Thus, the JSD measures how well the predicted responses approximate the actual responses across the entire range of different survey participants. The JSD is normalized to the range from 0 to 1, with lower values indicating higher similarity between the two distributions and thus better performance of the model in predicting the response distribution. We report the JSD for either the individual variables selected as prediction tasks or averaged across the full set of 27 prediction task variables.

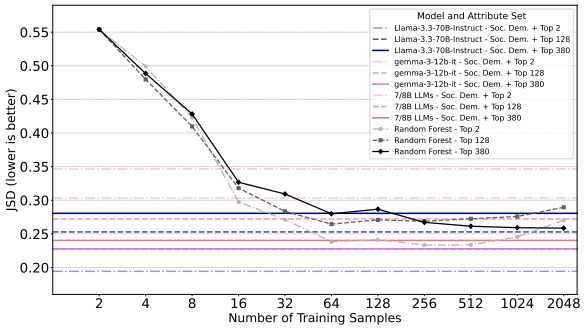

Figure 2: **Change in alignment with increased number of training samples.** We show the JSD between the survey response distribution and response distributions generated using persona-prompting with different LLMs as well as response distributions produced using random forest classifiers with increasing training set sizes. The alignment (averaged across 27 outcome variables) is better when using persona-prompting, particularly for small n—LLMs are already well-aligned, even without training samples.

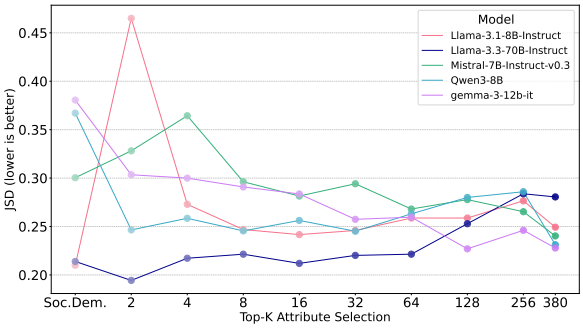

Figure 3: **Change in alignment with increased number of persona attributes.** We show the JSD between the survey response distribution and response distributions generated using increasingly large sets of TOP-k persona attributes for different LLMs. The largest LLAMA model outperforms others up until the TOP-64 attributes are used, showing best alignment when using only the TOP-2 attributes. Across models, adding persona attributes does not monotonously lead to better alignment.

Population Selection As discussed above, the *GGSS Personas* is representative of the German population because it mirrors the randomly sampled participants of the ALLBUS.

In the experiments in which we are interested in the effects of representativity on the ability of persona-prompted LLMs to replicate the population's response distribution, we systematically create persona collections that are by design *un*representative, e.g., oversampled on specific attributes. We sample 500 participants from the highest income class to oversample on *income*, 500 participants that lean conservative (based on their responses to the TOP-2 attributes, see Figure 1) to oversample on ideological leaning, and 500 students to mimic a typical convenience sample population in social science studies (Sears, 1986).

In addition, we run the same survey response tasks with the popular, but empirically not grounded *PersonaHub* collection (Ge et al., 2024), as well as using *no personas* at all. To ensure comparability of the results, we randomly sampled a subset of 500 personas from the representative *GGSS Personas*. By including multiple baselines we try to identify whether *GGSS Personas*'s representativeness offers any advantages over non-representative alternatives.

Response Generation We generate the synthetic survey response distribution for each of the selected tasks by prompting the LLMs with the persona descriptions featuring the socio-demographic characteristics and the TOP-k attributes of each of the $5,246$ ALLBUS participants. The models are restricted to generate only the first token of the available responses, thus indicating the corresponding response label. We use *vLLM* without varying any of the default response generation parameters. An exemplary prompt is given in Appendix B.1.

To generate responses for the *PersonaHub* collection, we include the full-text persona description in the same manner. We only prompt the model with the survey question of the prediction task to generate responses with *no persona*.

We only observe minor difficulties with models refusing to generate valid responses, with shares of invalid responses ranging from 1.44% for LLAMA-3.1-8B to 9.91% for QWEN3-8B (see Appendix B.1 Table 3). Interestingly, the share of invalid responses increases when the *PersonaHub* or the oversampled populations are used.

4.2. Results

We show that persona-prompting with the *GGSS Personas* offers advantages over traditional statistical methods for survey response predictions in situations marked by data sparsity with few survey participants and little information about them available (Figure 2), and that increasing the level of information reflected in persona descriptions does not necessarily lead to improved performance (Figure 3).

Additionally, we show that using the *GGSS Personas* leads to predicted response distributions that are better aligned with the survey response distribution than those produced using established baselines in 13 out of 27 tasks and across five out of nine topics (Figure 4). Finally, we show that the representativity of the *GGSS Personas* seems to have only little influence on the level of alignment (Figure 5).

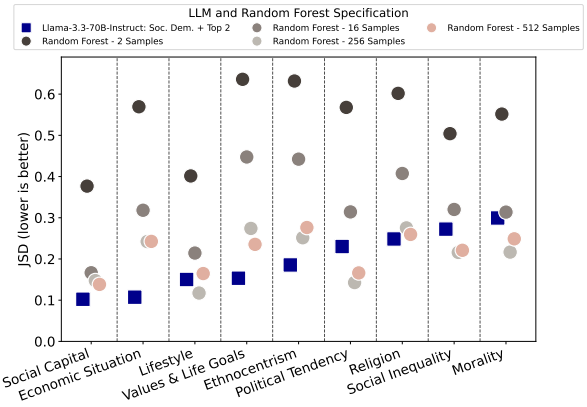

Figure 4: **Alignment comparison across different topics.** We compare the Jensen-Shannon Distance (JSD) between the survey response distribution and the response distribution generated using the best *GGSS Personas* configuration as well as the response distributions from random forest baselines with different training set sizes. Using the *GGSS Personas* and thus no input other than a single persona description produces response distributions that are better aligned with survey responses than the best random forest baselines across five out of nine different topics.

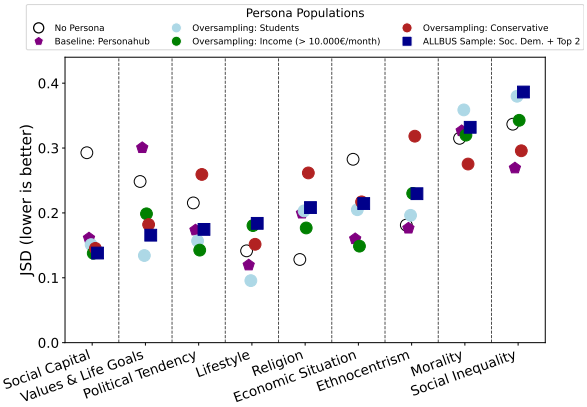

Figure 5: **Alignment comparison with unrepresentative persona collections.** We compare the average Jensen-Shannon Distance of response distributions generated with representative (*GGSS Personas*), unrepresentative (baselines), and *no persona* collections to the survey response distribution. All synthetic responses are generated using LLAMA-3.3-70B-INSTRUCT. The unrepresentative baselines and the *GGSS Personas* cluster closely together in terms of their JSD scores, indicating that representativity only has a small influence on the alignment.

Alignment for varying data availability. A main advantage of the persona-prompting approach is its ability to tap the world-knowledge of LLMs, allowing them to adapt without much input data to diverse new contexts. Thus, we are interested in evaluating how our approach fares in situations of data scarcity. We simulate these situations by varying the number of training samples n made available to the random forest classifiers serving as baselines, and by considering different sets of persona attributes k included in the persona prompts and passed as input to the baseline classifiers.

Figure 2 shows that LLM predictions of the survey response distribution are generally well-aligned, even though the LLM does not have access to any training data. Compared with the random forest classifiers trained on increasingly large training sets, the best performing LLMs outperform these baselines across all constellations. Generally, the persona-prompted LLMs offer the greatest advantage over the established alternative for survey response prediction when training data is scarce, i.e., when the random forest classifiers are trained on only small numbers of training samples. Figure 2 already indicates that using more persona attributes in persona prompts does not necessarily lead to better alignment with the survey response distribution, as shown by the fact that the best alignment results when using the largest model LLAMA-3.3-70B-INSTRUCT, but only the TOP-2 attributes for constructing the personas.

We examine this observation more closely in Figure 3, which shows that there is no monotonous relationship between the number of TOP-k attributes used for constructing the personas and the resulting alignment between the generated and the actual response distributions. Surprisingly, it is again the version of the *GGSS Personas* that uses only the TOP-2 persona attributes that leads to the best alignment across all constellations we tested. For the best-performing LLM, we even see a clear tendency for the alignment to degrade with higher numbers of included persona attributes, highlighting a potential signal-noise trade-off when including more and increasingly unimportant persona attributes. Figure 3 also reveals the general tendency that the best possible alignment improves with the size of the persona-prompted LLM used for generating the synthetic response distribution. Only the unlikely strong alignment of the relatively small LLAMA-3.1-8B-INSTRUCT prompted with only socio-demographic information breaks this trend, warranting closer investigation.

Alignment across topics and survey questions. In Figure 4, we compare the performance of persona-prompting LLMs using the *GGSS Personas* against random forest classifiers fitted on increasing numbers of training samples, now showing the disaggregated Jensen-Shannon Distances measured across nine different topics. Based on Figure 3, we use the best performing constellation

for predicting response distributions, i.e., LLAMA-3.3-70B-INSTRUCT prompted with TOP-2 personas, and increase the number of training samples for the baseline classifiers up to $n = 512$, after which we do not observe any more performance improvements. We report the average JSD between the predicted and actual survey response distributions across the three outcome variables randomly sampled for each topic.

Our persona-prompted LLM, receiving only the TOP-2 persona attributes in the prompt, outperforms even the best informed baseline with access to up to $n = 512$ of training samples and the same set of attributes in predicting the response distribution for 13 out of 27 different outcome variables. On the topic-level, the LLM-based approach turns out to be particularly strong in predicting response distributions in the categories *Social Capital*, *Economic Situation*, *Religion*, *Value & Life Goals*, and *Ethnocentrism*, where it achieves lower average JSD to the actual response distribution than even the best random forest classifiers. The absolute alignment using the *GGSS Personas* to persona-prompt LLMs is best for the topics *Social Capital* and *Economic Situation*, and worst for the topics *Social Inequality* and *Morality*.

Generally, the strong performance of the persona-prompt approach for at least some topics should be considered in the context of the fact that the baseline classifiers were able to learn persona-response-patterns from the training data, while the persona-based approach is purely zero-shot. The LLM has to rely solely on its *world knowledge* and the associations triggered by the single persona description.

Alignment compared to baseline persona populations. We run the same experiments using unrepresentative—oversampled on specific attributes and not empirically grounded—persona collections as baselines, as described in Section 4.1. In this comparison, there is no clear pattern showing that response distributions predicted using the representative *GGSS Personas* are consistently better aligned than the baselines they are compared against. Instead, Figure 5 shows that the different persona collections are generally scattered closely together around similar alignment scores across different topics, indicating that the predictive performance is highly topic dependent.

For the topics *Social Capital*, *Values & Life Goals*, *Political Tendency* and *Economic Situation*, the persona-prompted response distributions are closer aligned with the survey response distribution than the response distributions generated using *no personas*, indicating that the TOP-2 attributes used in the persona descriptions might have been helpful in correctly predicting the responses in these areas. The baseline oversampled on conservatives (as derived from the TOP-2 attributes) deviates most strongly from the *GGSS Personas*, strengthening our assumption that these attributes have the highest impact on the prediction of other ALLBUS variables. The fact that the empirically ungrounded *PersonaHub* collection aligns so closely with the different persona collections based on the ALLBUS is surprising.

5. Discussion

This work introduces *German General Social Survey Personas* - a novel prompt collection designed to evaluate and enhance the use of persona-based prompting for modeling response distributions of populations. Our preliminary experiments demonstrate the current ability of persona-based approaches to outperform traditional response prediction and imputation methods, particularly in situations characterized by limited data availability. These findings suggest that persona prompting can serve as a viable and flexible alternative in low-resource contexts where conventional supervised methods may struggle.

A key consideration in developing *German General Social Survey Personas* is achieving a balance between generalizability—ensuring that the collection remains applicable across a wide range of downstream tasks—and specificity, which is essential for strong performance in domain-specific applications. In this work, we propose leveraging general population surveys to construct empirically informed persona collections. Interestingly, our results indicate that personas based on a small subset of key variables outperform those that incorporate all available survey variables. This suggests that including only the most informative attributes may enhance both efficiency and predictive accuracy. In our preliminary analysis we only considered the average importance of variables for all other outcome variable. An important direction for future research is to explore how persona descriptions can be systematically tuned to optimally align with observed response patterns.

Recent work has begun fine-tuning large language models (LLMs) not only to generate single "correct" answers, but to minimize the divergence between the model's predicted response distributions and empirical human survey data from sources such as the World Values Survey and Pew Global Attitudes Survey (Cao et al., 2025; Suh et al., 2025). However, these supervised distribution-fitting methods require substantial computational resources and access to proprietary model weights, making them less applicable in low-resource or restricted-access settings. In contrast, persona prompting offers a more lightweight and adaptable

alternative that can be applied even to publicly available models. The performance of persona-based methods can potentially further be improved through few-shot approaches and with the release of increasingly capable language models. This highlights a promising avenue for future work, combining the interpretability and flexibility of prompting with the precision gains typically associated with fine-tuned models.

German General Social Survey Personas can act as a valuable resource that supports systematic comparisons of population-alignment of LLMs within different tasks (e.g., survey response generation, discussion simulations or behavior predictions).

6. Limitations

The idea of turning general social surveys such as the ALLBUS into empirically grounded persona collection comes with a number of limitations. First, surveys only offer a **snapshot of a population that dynamically changes over time**, both in its composition (e.g., due to migration or demographic change) as well as in the attitudes and beliefs held by individuals. To explicitly mark this temporal dependency, we date the released *GGSS Personas* through versioning. A partial alleviation of this limitation could also be in the use of reweighting techniques, which we leave for future work to explore.

Second, while we rule out the possibility that any of our tested LLMs has had access to the ALLBUS data we use, another (indirect) form of **data leakage** could occur across different releases of the ALLBUS, given that such surveys to a large degree run the same questions in every iteration. Future work could thus investigate performance differences between *GGSS Personas* versions created from and evaluated on current and past releases of the ALLBUS.

Third, missing information at an individual level introduces a **trade-off between the representativity of the persona collection and the completeness of the individual persona descriptions**. In our experiments, we prioritize representativity by including all personas instead of selecting only those that are complete, i.e., have information on all attributes available. By releasing the *GGSS Personas* with all personas, including those that are incomplete, we allow users of the collection to resolve this trade-off as they see fit—they could prioritize representativity by working with the collection as is, or they could prioritize persona completeness by selecting only personas with full information. Future work could explore how imputation procedures, possibly LLM-based (Castricato et al., 2025), might help to resolve this trade-off.

Fourth, our **attribute selection procedure does not account for correlations** and possible multicollinearity between predictors. Future work could test methods that evaluate all possible subsets of k attributes to improve upon the sets of TOP-k attributes currently used for persona creation.

In addition, LLM-generated closed-ended survey responses are susceptible to minimal perturbations, such as the positioning of the answer option (Rupprecht et al., 2025; Tjuatja et al., 2024). Special attention should be put on the robustness of synthetic survey responses by applying prompt perturbations to ensure that the results are not only artifacts of the prompting style.

Lastly, we are **relying exclusively on survey data**, both for the creation of persona descriptions and as a source of human groundtruth to evaluate our approach against. However, there is some evidence that survey data itself might be limited in accurately capturing human attitudes and behaviors. Future work could on the one hand explore the inclusion of behavioral data in persona prompts as a more immediate proxy of human behavior, and on the other hand search additional sources of groundtruth (such as election outcomes) to evaluate against.

Ethical Considerations

Generating artificial personas and representative collections might be relevant in various domains and applied to different use cases. However, surveying an artificial persona collection instead of the real, underlying population can be senseless or even dangerous depending on the downstream task at hand. For example, artificial persona collections can be exploited to pre-test and optimize targeted political or manipulative messaging. Such use risks amplifying disinformation campaigns and undermining democratic processes. Further, over-reliance on persona collections and its results is risky when there is no ground truth data of the real target population available as the alignment of the artificial responses cannot be evaluated. Frequent reliance on artificial responses may normalize their use where human perspectives are irreplaceable (e.g. in policymaking or clinical trials). This risks sidelining real human voices in domains directly impacting human lives. Researchers should also consider ethical evasion as one possible issue with synthetic persona collections and responses. Synthetic respondents might be viewed like a way to bypass obligatory ethical review processes since no real human participants are involved. This might encourage under-regulated research practices and in the long run weaken ethical safeguards.

Acknowledgement

We thank all anonymous reviewers for their thoughtful feedback and constructive suggestions, which greatly improved the quality of this work. The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

7. References

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S. Bernstein. 2025. Position: LLM Social Simulations Are a Promising Research Method. In *ICML 2025 Position Paper Track, 42nd International Conference on Machine Learning*.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351.

Joshua Ashkinaze, Emily Fry, Narendra Edara, Eric Gilbert, and Ceren Budak. 2025. Plurals: A System for Guiding LLMs via Simulated Social Ensembles. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. 2025. Specializing Large Language Models to Simulate Survey Response Distributions for Global Populations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3141–3154.

Louis Castricato, Nathan Lile, Rafael Rafailov, Jan-Philipp Fränken, and Chelsea Finn. 2025. PERSONA: A Reproducible Testbed for Pluralistic Alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11348–11368.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From Persona to Personalization: A Survey on Role-Playing Language Agents. *Transactions on Machine Learning Research*. Survey Certification.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025a. Persona Vectors: Monitoring and Controlling Character Traits in Language Models. *arXiv preprint arXiv:2507.21509*.

Sihan Chen, John P. Lalor, Yi Yang, and Ahmed Abbasi. 2025b. PersonaTwin: A Multi-Tier Prompt Conditioning Framework for Generating and Evaluating Personalized Digital Twins. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 774–788.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023a. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023b. CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875.

Pedro Henrique Luz de Araujo, Paul Röttger, Dirk Hovy, and Benjamin Roth. 2025. Principled Personas: Defining and Measuring the Intended Effects of Persona Prompting on Task Performance. *arXiv preprint arXiv:2508.19764*.

Wesley Hanwen Deng, Sunnie S. Y. Kim, Akshita Jha, Ken Holstein, Motahhare Eslami, Lauren Wilcox, and Leon A. Gatys. 2025. Personateaming: Exploring How Introducing Personas Can Improve Automated AI Red-Teaming. *arXiv preprint arXiv:2509.03728*.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny

Hernandez, Nicholas Joseph, et al. 2024. Towards Measuring the Representation of Subjective Global Opinions in Language Models. In *First Conference on Language Modeling*.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762.

Leon Fröhling, Gianluca Demartini, and Dennis Assenmacher. 2025. Personas With Attitudes: Controlling LLMs for Diverse Data Annotation. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*.

GESIS Leibniz-Institut für Sozialwissenschaften. 2025a. German General Social Survey ALLBUS 2023. (ZA8830; Version 1.2.0) [Data set]. GESIS, Cologne. https://doi.org/10.4232/1.14544.

GESIS Leibniz-Institut für Sozialwissenschaften. 2025b. German General Social Survey ALLBUS-compact 2023. (ZA8831; Version 1.3.0) [Data set]. GESIS, Cologne. https://doi.org/10.4232/1.14545.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling Synthetic Data Creation With 1,000,000,000 Personas. *arXiv preprint arXiv:2406.20094*.

Augusto Gonzalez-Bonorino, Monica Capra, and Emilio Pantoja. 2025. LLMs Model Non-WEIRD Populations: Experiments With Synthetic Cultural Agents. *arXiv preprint arXiv:2501.06834*.

Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. Predicting Results of Social Science Experiments Using Large Language Models. *Preprint*.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the Persona Effect in LLM Simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307.

Eunjeong Hwang, Bodhisattwa Majumder, and Niket Tandon. 2023. Aligning Language Models to User Opinions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919.

Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2024. Persona Is a Double-Edged Sword: Mitigating the Negative Impact of Role-Playing Prompts in Zero-Shot Reasoning Tasks. *arXiv preprint arXiv:2408.08631*.

Junsol Kim and Byungkyu Lee. 2023. AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction. *arXiv preprint arXiv:2305.09620*.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2024. The Benefits, Risks and Bounds of Personalizing the Alignment of Large Language Models to Individuals. *Nature Machine Intelligence*, 6(4):383–392.

William Kruskal and Frederick Mosteller. 1979. Representative Sampling, II: Scientific Literature, Excluding Statistics. *International Statistical Review/Revue Internationale de Statistique*, pages 111–127.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving With PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. 2025. The Prompt Makes the Person(a): A Systematic Evaluation of Sociodemographic Persona Prompting for Large Language Models. *arXiv preprint arXiv:2507.16076*.

Bolei Ma, Berk Yoztyurk, Anna-Carolina Haensch, Xinpeng Wang, Markus Herklotz, Frauke Kreuter, Barbara Plank, and Matthias Aßenmacher. 2025. Algorithmic Fidelity of Large Language Models in Generating Synthetic German Public Opinions: A Case Study. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1785–1809.

Fernando Miranda and Pedro Paulo Balbi. 2025. Simulating Public Opinion: Comparing Distributional and Individual-Level Predictions from LLMs and Random Forests. *Entropy*, 27(9):923.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond Demographics: Fine-Tuning Large Language Models to Predict Individuals' Subjective Text Perceptions. *arXiv preprint arXiv:2502.20897*.

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

Tiany Peng, George Gui, Daniel J. Merlau, Grace Jiarui Fan, Malek Ben Sliman, Melanie Brucks, Eric J. Johnson, Vicki Morwitz, Abdullah Althenayyan, Silvia Bellezza, et al. 2025. A Mega-Study of Digital Twins Reveals Strengths, Weaknesses and Opportunities for Further Improvement. *arXiv preprint arXiv:2509.19088*.

Jens Rupprecht, Georg Ahnert, and Markus Strohmaier. 2025. Prompt perturbations reveal human-like biases in large language model survey responses. *arXiv preprint arXiv:2507.07188*.

Paulo Salem, Robert Sim, Christopher Olsen, Prerit Saxena, Rafael Barcelos, and Yi Ding. 2025. TinyTroupe: An LLM-Powered Multiagent Persona Simulation Toolkit. *arXiv preprint arXiv:2507.09788*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

David O. Sears. 1986. College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature. *Journal of Personality and Social Psychology*, 51(3):515.

Indira Sen, Marlene Lutz, Elisa Rogers, David Garcia, and Markus Strohmaier. 2025. Missing the Margins: A Systematic Literature Review on the Demographic Representativeness of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24263–24289. Association for Computational Linguistics.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large Language Models Can Be Easily Distracted by Irrelevant Context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. 2024. Position: A Roadmap to Pluralistic Alignment. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46280–46302.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions. In *80th Annual AAPOR Conference*. AAPOR.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi Fung, Hou Pong Chan, Kevin Small, ChengXiang Zhai, and Heng Ji. 2025. Persona-DB: Efficient Large Language Model Personalization for Response Prediction With Collaborative Data Refinement. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 281–296.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Olivier Toubia, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen. 2025. Database Report: Twin-2K-500: A Data Set for Building Digital Twins of Over 2,000 People Based on Their Answers to Over 500 Questions. *Marketing Science*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631.

Leah von der Heyde, Anna-Carolina Haensch, and Alexander Wenz. 2025. Vox Populi, Vox AI? Using Large Language Models to Estimate German Vote Choice. *Social Science Computer Review*.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2025. Large Language Models That Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups. *Nature Machine Intelligence*, pages 1–12.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, Nesreen K. Ahmed, and Yu Wang. 2025. Personalization of Large Language Models: A Survey. *Transactions on Machine Learning Research*. Survey Certification.

A. *German General Social Survey Personas*

Figure 6 presents the original survey items and persona descriptions that have been translated to English for Figure 1.

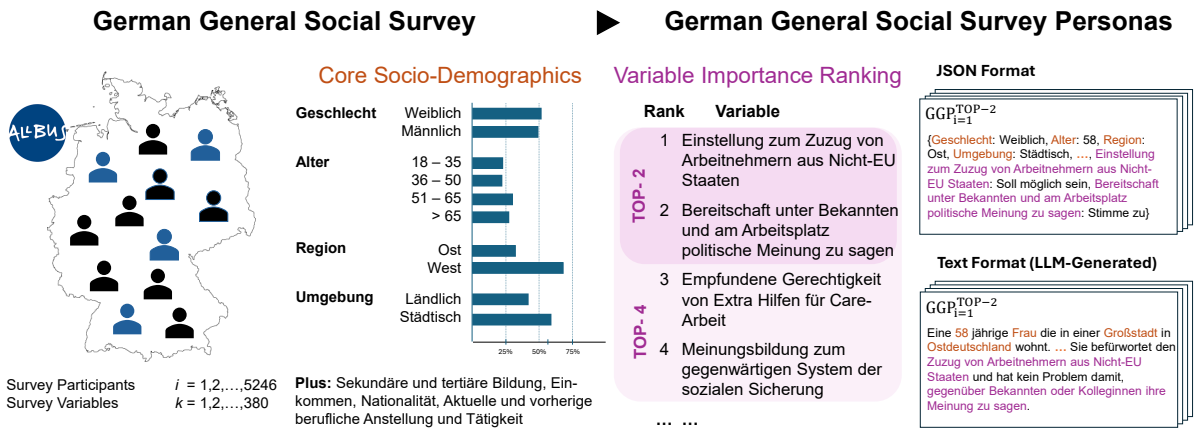

Figure 6: **German Version of Figure 1:** Grounding the German General Social Survey Personas (GGSS Personas) in the ALLBUS survey.

Persona Generation We do not observe large differences in response distribution alignment between using different persona formats. Since there is no format that consistently outperforms the others, we decided to conduct our experiments with the personas in **JSON**-format. Figure 7 underlines this decision exemplarily for `Llama-3.1-8B-Instruct`.

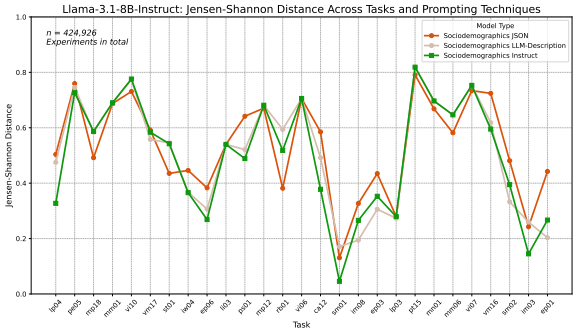

Figure 7: **Alignment Comparison for different Persona Prompting Techniques.** Except for some tasks, we do not observe consistent and large differences in alignment depending on the exact prompting strategy. In our test, it did make no to very little difference whether the persona attributes are introduced with a a json format, as a description of a person, such as "*Take the view of the following person that ...*". or as a direct instruction, such as "*You are a 50 year old male from western Germany*".

B. Preliminary Analysis

B.1. Experimental Setup

Model Selection Table 2 shows the dates of knowledge-cutoff and release for the LLMs used in our preliminary analysis.

Model	K-Cutoff	R-Date
Llama-3.1-8B Instruct	12/2023	23/07/2024
Llama-3.3-70B-Instruct	12/2023	06/12/2024
Mistral-7B-Instruct-v0.3	unknown	22/05/2024
Qwen3-8B	unknown	27/04/2025
gemma-3-12B-it	08/2024	10/03/2025

Table 2: **LLM release dates for empirical evaluation.** All models with a known knowledge cutoff (*K-Cutoff*) were released after the ALLBUScompact date, ensuring no access to training data. For others, the small gap between ALLBUS and model release (*R-Date*) makes data access highly unlikely.

Response Generation Below, we provide an example of a German persona prompt, featuring the core set of socio-demographic attributes as well as the two attributes identified to be most important (TOP-2) according to our variable importance ranking.

> "Nehme die Perspektive der folgenden Person ein: 'Wie alt sind Sie?': '50', 'Wohnen Sie in Ost- oder Westdeutschland?': 'Westdeutschland', 'Was ist Ihr höchster Schulabschluss?': '(3) Mittlere Reife', 'In welcher Wohnumgebung wohnen Sie?': 'Großstadt', 'Wie hoch ist Ihr monatliches Nettoeinkommen?': '(11) 1250 - 1374 Euro', 'Wie beschreiben Sie Ihre aktuelle berufliche Tätigkeit?': 'Montageberufe', 'Wie beschreiben Sie Ihre letzte berufliche Tätigkeit?': 'Nicht zutreffend', 'Welches Geschlecht haben Sie?': 'Männlich', 'Welchen Universitätsabschluss besitzen Sie?': 'Kein Universitätsabschluss', 'Wie ist Ihre Einstellung zum Zuzug von Arbeitnehmern aus Nicht-EU-Staaten?': '(2) Zuzug Begrenzen', 'Käme es für Sie in Frage, Ihre

politische Meinung im Bekanntenkreis und am Arbeitsplatz zu sagen, um Einfluss zu nehmen?': '(1) Ja' Welche der Antwortmöglichkeiten ist die Reaktion der Person auf folgende Frage: Meinung zur Verantwortung, Kinder zu bekommen Antwortmöglichkeiten: ['1: BIN DERS.MEINUNG', '2: BIN ANDERER MEINUNG'] Antwort: ("

This prompt structure instructs the LLM to take on the perspective of the persona described next, provides the survey question the model is supposed to answer as well as the available response options, and sets the model to up to generate its response in the required format.

In this example, the persona is a 50-year-old male from West Germany with an intermediate school leaving certificate, living in a large city and working in assembly with a net monthly income of 1250 - 1374 Euros, supporting limiting immigration from non-EU countries, and willing to voice his political opinions among friends and at work. The prompt then asks for this persona's reaction to a question about the responsibility of having children (variable name: lp04), with the options being "agree" or "disagree".

The English translation of the provided prompt is the following:

"Assume the perspective of the following person: 'How old are you?': '50', 'Do you live in East or West Germany?': 'West Germany', 'What is your highest school leaving certificate?': '(3) Intermediate school leaving certificate', 'In which residential environment do you live?': 'Large city', 'What is your monthly net income?': '(11) 1250 - 1374 Euro', 'How do you describe your current occupation?': 'Assembly jobs', 'How do you describe your last occupation?': 'Not applicable', 'What is your gender?': 'Male', 'What university degree do you have?': 'No university degree', 'What is your attitude towards the immigration of workers from non-EU countries?': '(2) Limit immigration', 'Would you consider expressing your political opinion to friends and at work to exert influence?': '(1) Yes' Which of the answer options is the person's reaction to the following question: Opinion on the responsibility to have children Answer options: ['1: I AGREE', '2: I DISAGREE'] Answer: (".

Decoding Constraints and Invalid Responses We generated responses from the whole token distribution. We did not apply any decoding constraints such as the Top-P, Top-K sampling or temperature setting. We do acknowledge that these might impact the response generation.

A response is valid when the language model generates only one token that reflects one response option on the answer option scale. Table 3 shows the rates with which the different LLMs used for generating synthetic response distributions failed to generate valid responses.

Model	Invalid (%)
Llama-3.1-8B-Instruct	1.44
Llama-3.3-70B-Instruct	8.75
Mistral-7B-Instruct-v0.3	3.73
Qwen3-8B	9.91
gemma-3-12b-it	1.98

Table 3: **LLMs Invalid Responses by Model for TOP-K Personas Only.** `Llama-3.1-8B-Instruct` and `gemma-3-12b-it` have the lowest rates of answering with invalid responses (e.g. refusal, wrong response label, no answer). `Qwen3-8B` is the most unreliable respondent. Invalid responses increase as arbitrary or oversampled populations are used instead of the representative population.

Infrastructure We carried out the experiments of predicting attributes with GGSS Personas's synthetic responses on a high-performance computing cluster and a local server equipped with NVIDIA H100 (80GB) GPUs and Blackwell GPUs. To accommodate larger models on available hardware, we applied 8-bit quantization to `Llama-3.3-70B-Instruct`. Smaller models were run without quantization.

Further, we use *vLLM* for more efficient memory management and higher throughput as this package enables sufficient batching of many requests at a time (Kwon et al., 2023).

C. Evaluation

C.1. Validation of Full-Text Persona Generation

We generate full-text personas based on the JSON-format persona descriptions with `Gemini-2.5-flash`. Although new, state-of-the-art LLMs are powerful and capable of dealing with large context windows, the generated full-text personas must be validated. LLMs are prone to hallucinations or misinterpreting the meaning of a response to a survey item when generating a full-text persona description. Although the maximum input token size of `Gemini-2.5-flash` exceeds 1 million tokens [3], more content in the context window might increase

[3] see Gemini-2.5-flash Documentation

the likelihood of errors, hallucinations, or misinterpretation.

Apart from the minor misrepresentations and omissions found in the manually validated persona descriptions with up to 16 included attributes (TOP-2,-4,-8 and -16), we anecdotally observed that full-text personas with the largest sets of attributes (TOP-380) show major flaws. For multiple generated persona descriptions we observe generation loops, i.e., situations in which individual sentences in the persona description are repeated until the maximum token limit is reached. Therefore, we recommend using full-text personas with fewer attributes to avoid the use of these types of faulty personas.

C.2. Hyperparameters for Random Forest Classifier

The Random Forest models utilized for classification tasks throughout all experiments were instantiated from the `scikit-learn` library's `RandomForestClassifier` class. We selected the hyperparameters according to the findings of Miranda and Balbi (2025) who used the classifier in a similar task. To ensure consistency and comparability across experiments, the following hyperparameters were fixed:

- `n_estimators = 100`: The number of trees in the forest.
- `max_depth = 10`: The maximum depth of each tree.
- `min_samples_split = 2`: The minimum number of samples required to split an internal node.
- `min_samples_leaf = 2`: The minimum number of samples required to be at a leaf node.
- `random_state`: A specific integer value was used for the `random_state` parameter in each experimental run to ensure reproducibility of the results.

These parameters were selected based on preliminary experimentation to provide a reasonable balance between model complexity and generalization. During these initial tests, increasing the number of estimators beyond 100 yielded no measurable gains in performance. Conversely, increasing `max_depth` indicated potential overfitting. The goal of this configuration was to define a stable, conservative baseline classifier for comparison with the LLMs, avoiding overfitting the Random Forest for each specific topic.

D. Survey Data

The following table lists the outcome variables used to evaluate the *GGSS Personas*. These variables were left out when generating the persona descriptions to avoid introducing the relevant information of the individual response directly into the persona prompt.

Cat.	Var.	German Question	English Question	German Response	English Response
econ.	ep01	Wie beurteilen Sie ganz allgemein die heutige wirtschaftliche Lage in Deutschland?	How do you generally assess the current economic situation in Germany?	[Sehr gut; Gut; Teils gut / teils schlecht; Schlecht; Sehr schlecht]	[Very good; Good; Partly good / partly bad; Bad; Very bad]
econ.	ep03	Wie beurteilen Sie Ihre eigene wirtschaftliche Lage heute?	How do you assess your own economic situation today?	[Sehr gut; Gut; Teils gut / teils schlecht; Schlecht; Sehr schlecht]	[Very good; Good; Partly good / partly bad; Bad; Very bad]
econ.	ep06	Was glauben Sie, wie wird Ihre eigene wirtschaftliche Lage in einem Jahr sein?	What do you think your own economic situation will be like in a year?	[Wesentlich besser als heute; Etwas besser als heute; Gleichbleibend; Etwas schlechter als heute; Wesentlich schlechter als heute]	[Significantly better than today; Somewhat better than today; About the same; Somewhat worse than today; Significantly worse than today]
ethno.	mp12	Auf einer Skala von 1 (stimme überhaupt nicht zu) bis 7 (stimme voll und ganz zu), inwieweit stimmen Sie folgender Aussage zu: "Die Ausländer in Deutschland tragen dazu bei, den Fachkräftemangel zu beheben."	On a scale from 1 (strongly disagree) to 7 (strongly agree), to what extent do you agree with the following statement: “Foreigners in Germany help to alleviate the shortage of skilled workers.”	7-Punkte Skala (1 := Stimme überhaupt nicht zu — 7 := Stimme voll und ganz zu)	7-point scale (1 = Strongly disagree — 7 = Strongly agree)
ethno.	mn01	Auf einer Skala von 1 (überhaupt nicht wichtig) bis 7 (sehr wichtig), wie wichtig sollte Ihrer Meinung nach folgender Umstand bei der Entscheidung über die Vergabe der deutschen Staatsbürgerschaft sein: "Ob die Person in Deutschland geboren ist."	On a scale from 1 (not important at all) to 7 (very important), how important do you think the following factor should be in the decision on granting German citizenship: “Whether the person was born in Germany.”	7-Punkte Skala (1 := Überhaupt nicht wichtig — 7 := Sehr wichtig)	7-point scale (1 = Not important at all — 7 = Very important)

ethno.	mp18	Wegen der in den letzten Jahren nach Deutschland gekommenen Flüchtlinge - denken Sie, es ergeben sich mehr Chancen, mehr Risiken, oder weder noch in Bezug auf das Zusammenleben in unserer Gesellschaft?	Considering the refugees who have come to Germany in recent years – do you think this has led to more opportunities, more risks, or neither with regard to living together in our society?	[Deutlich mehr Risiken; Eher mehr Riskien; Weder noch; Eher mehr Chancen; Deutlich mehr Chancen]	[Significantly more risks; Somewhat more risks; Neither; Somewhat more opportunities; Significantly more opportunities]
lifesty.	li03	Auf einer Skala von 1 (unwichtig) bis 7 (sehr wichtig), wie wichtig ist Ihnen der Lebensbereich "Freizeit und Erholung"?	On a scale from 1 (unimportant) to 7 (very important), how important is the area of life “Leisure and recreation” to you?	7-Punkte Skala (1 := Unwichtig — 7 := Sehr wichtig)	7-point scale (1 = Unimportant — 7 = Very important)
lifesty.	lp03	Sind Sie derselben oder anderer Meinung mit der Aussage "Egal was manche Leute sagen: Die Situation der einfachen Leute wird nicht besser, sondern schlechter"?	Do you agree or disagree with the statement: “No matter what some people say, the situation of ordinary people is not getting better but worse”?	[Bin derselben Meinung; Bin anderer Meingung]	[Agree; Disagree]
lifesty.	lp04	Sind Sie derselben oder anderer Meinung mit der Aussage "So wie die Zukunft aussieht, kann man es kaum noch verantworten, Kinder auf die Welt zu bringen"?	Do you agree or disagree with the statement: “Given how the future looks, it is hardly justifiable to bring children into the world”?	[Bin derselben Meinung; Bin anderer Meingung]	[Agree; Disagree]
moral.	ca12	Was halten Sie von der folgenden Verhaltensweise: "Jemand raucht mehrmals in der Woche Haschisch"?	What is your opinion of the following behavior: “Someone smokes hashish several times a week”?	4-Punkte Skala (1 := Sehr schlimm — 4 = Überhaupt nicht shlimm)	4-point scale (1 = Very bad — 4 = Not bad at all)
moral	vm16	Für Paare, die sich ein Kind wüschen, aber auf natürlichem Wege keines bekommen können - wie beruteilen Sie die folgende Alternative: "Ein Paar verwendet eigene Ei- oder Samenzellen, um mit medizinischer Hilfe ein Kind zu bekommen."	For couples who wish to have a child but cannot conceive naturally – how do you evaluate the following alternative: “A couple uses their own egg or sperm cells to have a child with medical assistance.”	7-Punkte Skala (-3 := Sehr falsch — 3 := Sehr richtig)	7-point scale (-3 = Very wrong — 3 = Very right)

moral	vm17	Für Paare, die sich ein Kind wüschen, aber auf natürlichem Wege keines bekommen können - wie beruteilen Sie die folgende Alternative: "Ein Paar verwendet anonym gespendete Ei- oder Samenzellen, um mit medizinischer Hilfe ein Kind zu bekommen."	For couples who wish to have a child but cannot conceive naturally – how do you evaluate the following alternative: “A couple uses anonymously donated egg or sperm cells to have a child with medical assistance.”	7-Punkte Skala (-3 := Sehr falsch — 3 := Sehr richtig)	7-point scale (-3 = Very wrong — 3 = Very right)
pol.tend.	pt15	Auf einer Skala von 1 (überhaupt kein Vertrauen) bis 7 (sehr großes Vertrauen), wie groß ist das Vertrauen, das Sie "politischen Parteien" entgegenbringen?	On a scale from 1 (no trust at all) to 7 (a great deal of trust), how much trust do you have in “political parties”?	7-Punkte Skala (1 := Überhaupt kein Vertrauen — 7 := Sehr großes Vertrauen)	7-point scale (1 = No trust at all — 7 = A great deal of trust)
pol.tend.	pe05	Inwieweit stimmen Sie folgender Meinung zu: "Die Politiker bemühen sich im Allgemeinen darum, die Interessen der Bevölkerung zu vertreten."	To what extent do you agree with the following statement: “Politicians generally try to represent the interests of the population.”	4-Punkte Skala (1 := Stimme voll und ganz zu — 4 := Stimme überhaupt nicht zu)	4-point scale (1 = Strongly agree — 4 = Strongly disagree)
pol.tend.	ps01	Auf einer Skala von 1 (sehr zufrieden) bis 6 (sehr unzufrieden), wie zufrieden sind Sie - insgesamt betrachtet - mit den gegenwärtigen Leistungen der Bundesregierung?	On a scale from 1 (very satisfied) to 6 (very dissatisfied), overall, how satisfied are you with the current performance of the federal government?	6-Punkte Skala (1 := Sehr zufrieden — 6 := Sehr unzufrieden)	6-point scale (1 = Very satisfied — 6 = Very dissatisfied)
relig.	rb01	Inwieweit stimmen Sie folgender Aussage zu: "Es gibt einen Gott, der sich mit jedem Menschen persönlich befasst."	To what extent do you agree with the following statement: “There is a God who concerns Himself personally with every human being.”	5-Punkte Skala (1 := Stimme voll und ganz zu — 5 := Stimme überhaupt nicht zu)	5-point scale (1 = Strongly agree — 5 = Strongly disagree)
relig.	mm01	Auf einer Skala von 1 (stimme überhaupt nichts zu) bis 7 (stimme voll und ganz zu), inwieweit stimmen Sie folgender Aussage zu: "Die Ausübung des islamischen Glaubens in Deutschland sollte eingeschränkt werden."	On a scale from 1 (strongly disagree) to 7 (strongly agree), to what extent do you agree with the following statement: “The practice of the Islamic faith in Germany should be restricted.”	7-Punkte Skala (1 := Stimme überhaupt nicht zu — 7 := Stimme voll und ganz zu)	7-point scale (1 = Strongly disagree — 7 = Strongly agree)

relig. mm06	Auf einer Skala von 1 (stimme überhaupt nichts zu) bis 7 (stimme voll und ganz zu), inwieweit stimmen Sie folgender Aussage zu: "Ich habe den Eindruck, dass unter den in Deutschland lebenden Muslimen viele religiöse Fanatiker sind."	On a scale from 1 (strongly disagree) to 7 (strongly agree), to what extent do you agree with the following statement: “I have the impression that among Muslims living in Germany there are many religious fanatics.”	7-Punkte Skala (1 := Stimme überhaupt nicht zu — 7 := Stimme voll und ganz zu)	7-point scale (1 = Strongly disagree — 7 = Strongly agree)
soc.cap. st01	Manche Leute sagen, dass man den meisten Menschen trauen kann. Andere meinen, dass man nicht vorsichtig genug sein kann im Umgang mit anderen Menschen. Was ist Ihre Meinung dazu?	Some people say that most people can be trusted. Others think that you can’t be too careful when dealing with other people. What is your opinion on this?	[Den meisten Menschen kann man trauen; Man kann nicht vorsichtig genug sein; Das kommt drauf an]	[Most people can be trusted; You can’t be too careful; It depends]
soc.cap. sm01	Sind Sie derzeit Mitglied in einer Gewerkschaft?	Are you currently a member of a trade union?	[Ja; Nein]	[Yes; No]
soc.cap. sm02	Waren Sie früher einmal Mitglied in einer Gewerkschaft?	Have you ever been a member of a trade union in the past?	[Ja; Nein]	[Yes; No]
soc.ineq. im03	Auf einer Skala von 1 (sehr wichtig) bis 4 (unwichtig), wie beurteilen Sie die gegenwärtige Wichtigkeit der folgenden Eigenschaften und Umstände für einen Aufstieg in unserer Gesellschaft: "Bildung, Ausbildung".	On a scale from 1 (very important) to 4 (unimportant), how do you rate the current importance of the following characteristic or factor for upward mobility in our society: “Education, training.”	4-Punke Skala (1 := Sehr wichtig — 4 := Unwichtig)	4-point scale (1 = Very important — 4 = Unimportant)
soc.ineq. im08	Auf einer Skala von 1 (sehr wichtig) bis 4 (unwichtig), wie beurteilen Sie die gegenwärtige Wichtigkeit der folgenden Eigenschaften und Umstände für einen Aufstieg in unserer Gesellschaft: "Leistung, Fleiß".	On a scale from 1 (very important) to 4 (unimportant), how do you rate the current importance of the following characteristic or factor for upward mobility in our society: “Achievement, diligence.”	4-Punke Skala (1 := Sehr wichtig — 4 := Unwichtig)	4-point scale (1 = Very important — 4 = Unimportant)
soc.ineq. iw04	Inwieweit stimmen Sie folgender Aussage zu: "Der Staat muss dafür sorgen, dass man auch bei Krankheit, Not, Arbeitslosigkeit und im Alter ein gutes Auskommen hat."	To what extent do you agree with the following statement: “The state must ensure that people have a decent standard of living even in cases of illness, hardship, unemployment, and old age.”	4-Punkte Skala (1 := Stimme voll zu — 4 := Stimme überhaupt nicht zu)	4-point scale (1 = Completely agree — 4 = Do not agree at all)

values	vi06	Auf einer Skala von 1 (unwichtig) bis 7 (außerordentlich wichtig), wie wichtig ist Ihnen "Sozial Benachteiligten und gesellschaftlichen Randgruppen helfen"?	On a scale from 1 (unimportant) to 7 (extremely important), how important is “Helping socially disadvantaged people and marginalized groups” to you?	7-Punkte Skala (1 := Unwichtig — 7 := Außerordentlich wichtig)	7-point scale (1 = Unimportant — 7 = Extremely important)
values	vi07	Auf einer Skala von 1 (unwichtig) bis 7 (außerordentlich wichtig), wie wichtig ist Ihnen "Sich und seine Bedürfnisse gegen andere durchsetzen"?	On a scale from 1 (unimportant) to 7 (extremely important), how important is “Asserting yourself and your needs against others” to you?	7-Punkte Skala (1 := Unwichtig — 7 := Außerordentlich wichtig)	7-point scale (1 = Unimportant — 7 = Extremely important)
values	vi10	Auf einer Skala von 1 (unwichtig) bis 7 (außerordentlich wichtig), wie wichtig ist Ihnen "Sich politisch engagieren"?	On a scale from 1 (unimportant) to 7 (extremely important), how important is “Being politically active” to you?	7-Punkte Skala (1 := Unwichtig — 7 := Außerordentlich wichtig)	7-point scale (1 = Unimportant — 7 = Extremely important)

Table 4: **Overview of the variables selected as prediction tasks.** For each of the nine topical areas (*cat.*) we randomly select three variables (*var.*) as prediction tasks. We show the original ALLBUS questions and the corresponding response options (slightly adjusted for readability), as well as their English translations.