Guided Persona-based AI Surveys:
Can we replicate personal mobility preferences at scale using LLMs?

Ioannis Tzachristas*, Santhanakrishnan Narayanan and Constantinos Antoniou

Chair of Transportation Systems Engineering, TUM School of Engineering and Design, TUM, Germany

SHORT SUMMARY

This study explores the potential of Large Language Models (LLMs) to generate artificial surveys, with a focus on personal mobility preferences in Germany. By leveraging LLMs for synthetic data creation, we aim to address the limitations of traditional survey methods, such as high costs, inefficiency and scalability challenges. A novel approach incorporating "Personas" - combinations of demographic and behavioural attributes - is introduced and compared to five other synthetic survey methods, which vary in their use of real-world data and methodological complexity. The MiD 2017 dataset, a comprehensive mobility survey in Germany, serves as a benchmark to assess the alignment of synthetic data with real-world patterns. The results demonstrate that LLMs can effectively capture complex dependencies between demographic attributes and preferences while offering flexibility to explore hypothetical scenarios. This approach presents valuable opportunities for transportation planning and social science research, enabling scalable, cost-efficient and privacy-preserving data generation.

Keywords: artificial surveys, LLMs, mobility preferences, synthetic data, Persona-based

1 INTRODUCTION

Traditional survey methods face significant challenges, including high costs, inefficiencies and scalability limitations. As transportation systems evolve, there is an increasing demand for more efficient and adaptable data collection methods. In response, this study investigates the feasibility of using Large Language Models (LLMs) to enhance travel survey data collection (Liu et al., 2024; Tan et al., 2024).
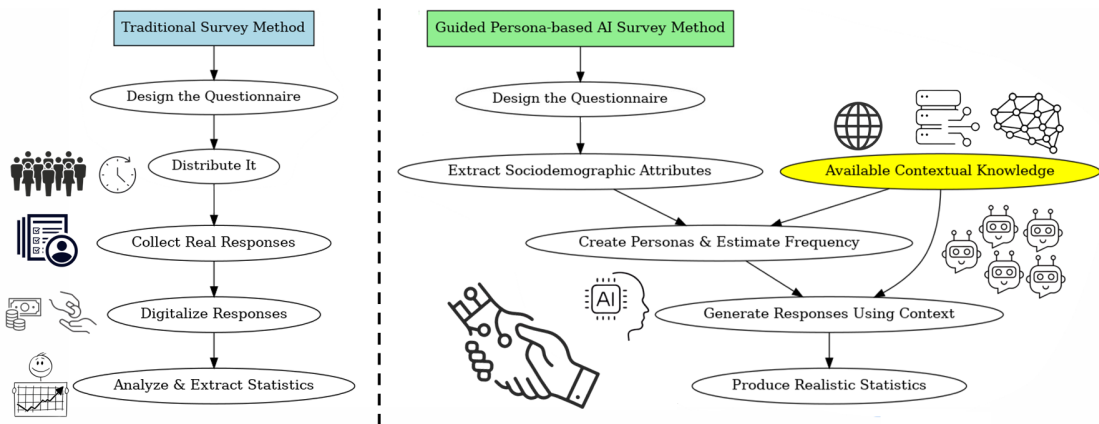

Figure 1: Traditional Survey Method vs Guided Persona-based AI Survey Method

We propose a novel guided Persona-based AI survey methodology[1], which, to the best of our knowledge, has not been previously applied to travel surveys. This approach systematically integrates sociodemographic and behavioural attributes to generate synthetic survey responses. To evaluate its effectiveness, we benchmark our method against other synthetic survey generation

---

[1] https://github.com/jtzach/LLM-based-AI-surveys-MiD2017.git

approaches using the MiD 2017 dataset (MiD, 2017), a comprehensive national mobility survey in Germany. Our findings demonstrate that the guided Persona-based approach significantly outperforms competing methods in terms of accuracy and consistency with observed mobility behaviours and patterns.

Existing studies in synthetic data generation have primarily focused on demographic simulation or imputation techniques, with limited applications to large-scale survey replication (Wang et al., 2024; Ma et al., 2024; Choenni et al., 2023; Bisbee et al., 2024). While these methods effectively replicate population structures, they often fail to capture intricate dependencies between demographic attributes and travel behaviours. For example, economic status may influence preferences for cycling or public transport, but existing methods may overlook such correlations unless explicitly predefined. Our guided Persona-based approach addresses these limitations by leveraging LLMs' contextual understanding to generate more context-aware and realistic synthetic data.

This study contributes to the advancement of artificial survey methodologies by offering a scalable, cost-effective and privacy-preserving alternative to traditional survey methods. The results highlight the potential of guided Persona-based AI surveys to support transportation planning and social science research by enabling the creation of high-fidelity synthetic datasets.

Why LLM-generated Artificial Surveys?

Artificial surveys offer numerous advantages over traditional survey methods:

- **Privacy-preservation**: By generating synthetic data, researchers can analyze population-level trends without accessing sensitive individual data (Livieris et al., 2024).

- **Cost-efficiency**: Synthetic data generation reduces the need for repeated data collection, significantly lowering costs.

- **Scalability**: LLMs can generate large datasets that encompass diverse scenarios, enabling analyses that were previously unfeasible (Tan et al., 2024).

- **Flexibility**: Researchers can simulate responses to questions that were not included in the original survey.

- **Adaptability**: Artificial surveys can be tailored to model specific sub-populations or explore "what-if" scenarios.

These benefits demonstrate the transformative potential of synthetic data in advancing survey methodologies and addressing long-standing challenges in social science research.

2 Methodology

Dataset Overview

The MiD 2017 dataset captures detailed mobility preferences in Germany, including demographic distributions, transport modes and trip frequencies. Key attributes include age groups, education levels, main activities, economic status and household types. We normalize the translated to English version of the dataset so that the percentage of "not specified" option is merged to the most popular responses of each question as shown in the Appendix (**Figure 4**).

Survey Generation Methods

For AI-survey generation, the GPT-4o API was utilized in a framework inspired by previous work (Tzachristas, 2024). Synthetic populations of 10,000 individuals were generated for all non-Persona-based methods. Persona-based methods involved creating distinct demographic combinations, leading to a total of 15,840 unique Personas, each defined by combinations of attributes such as age group, education level, main activity, economic status and household type.

The survey generation techniques employed a range of prompts tailored to each method. The prompts used for each method, as well as the LLM-generated code of the experiments are available here: `https://github.com/jtzach/LLM-based-AI-surveys-MiD2017.git`.
Figure 2 provides an overview of the methods.

Naive AI-survey:
A synthetic population for 2017 Germany was generated based on general demographic knowledge. Mobility-related survey responses were then simulated for each individual without any constraints or alignment with real-world benchmarks (Bisbee et al., 2024).

Structured AI-survey:
The synthetic population was aligned with demographic benchmarks provided by the MiD 2017 dataset. This method ensured a realistic representation of population structures, but responses were not explicitly aligned to the observed distributions of the MiD 2017.

Guided AI-survey:
This approach added constraints to both population structure and response averages, by considering the expected statistics of MiD 2017 responses.

Naive Persona-based AI Survey:
All possible combinations of demographic and behavioural attributes were identified to define Personas. Densities for each Persona were estimated using general demographic knowledge. Mobility-related responses were then simulated for each Persona.

Structured Persona-based AI Survey:
This method enhanced the Naive Persona-based AI Survey by aligning Persona densities with MiD 2017 demographic data.

Guided Persona-based AI Survey:
Building upon the Structured Persona-based AI Survey, this method incorporated expected response statistics from the MiD 2017 dataset.

The prompts used in the experiments were specifically crafted to guide GPT-4o's generation of realistic populations and survey responses. These prompts ranged from straightforward instructions (e.g. "Generate a population based on general demographic knowledge") to advanced directives that explicitly instructed the LLM to consider correlations between attributes (e.g. "Simulate mobility preferences while maintaining realistic correlations between age, household type and economic status based on MiD 2017 data"). This approach ensured high fidelity in response generation and allowed each method to address specific research goals while capturing intricate interdependencies between attributes.

Persona Methods for Artificial Survey Generation

In this section, we describe the Persona-based methods used in generating artificial surveys. This method revolves around defining a set of Personas based on a combination of demographic and behavioural attributes. We construct a partition of the population into Personas. We then calculate the population percentages of these Personas and use them for weighted analyses. Below, we formally define the key concepts and methods.

Definition of a Persona

A Persona is a tuple defined by a specific combination of demographic and behavioural attributes. Let the attributes be:

- **Age Group** (A): {14–17, 18–29, 30–39, 40–49, 50–59, 60–64, 65–74, 75–79, 80+}

- **Education Level** (E): {No Degree (yet), Low, Medium, High}

- **Main Activity** (M): {Full-time employee, Part-time employee, Employed (unspecified), Pupil, Student, Housewife/Househusband, Pensioner, Other}

- **Economic Status** (S): {Very Low, Low, Medium, High, Very High}

- **Household Type** (H): {Young singles, Middle-aged singles, Older singles, Young two-person households, Middle-aged two-person households, Older two-person households, Households with at least 3 adults, Households with at least 1 child under 6, Households with at least 1 child under 14, Households with at least 1 child under 18, Single parents}

A Persona P is then formally defined as:

$$P(A, E, M, S, H), \tag{1}$$

where $A \in$ Age Group, $E \in$ Education Level, $M \in$ Main Activity, $S \in$ Economic Status and $H \in$ Household Type.
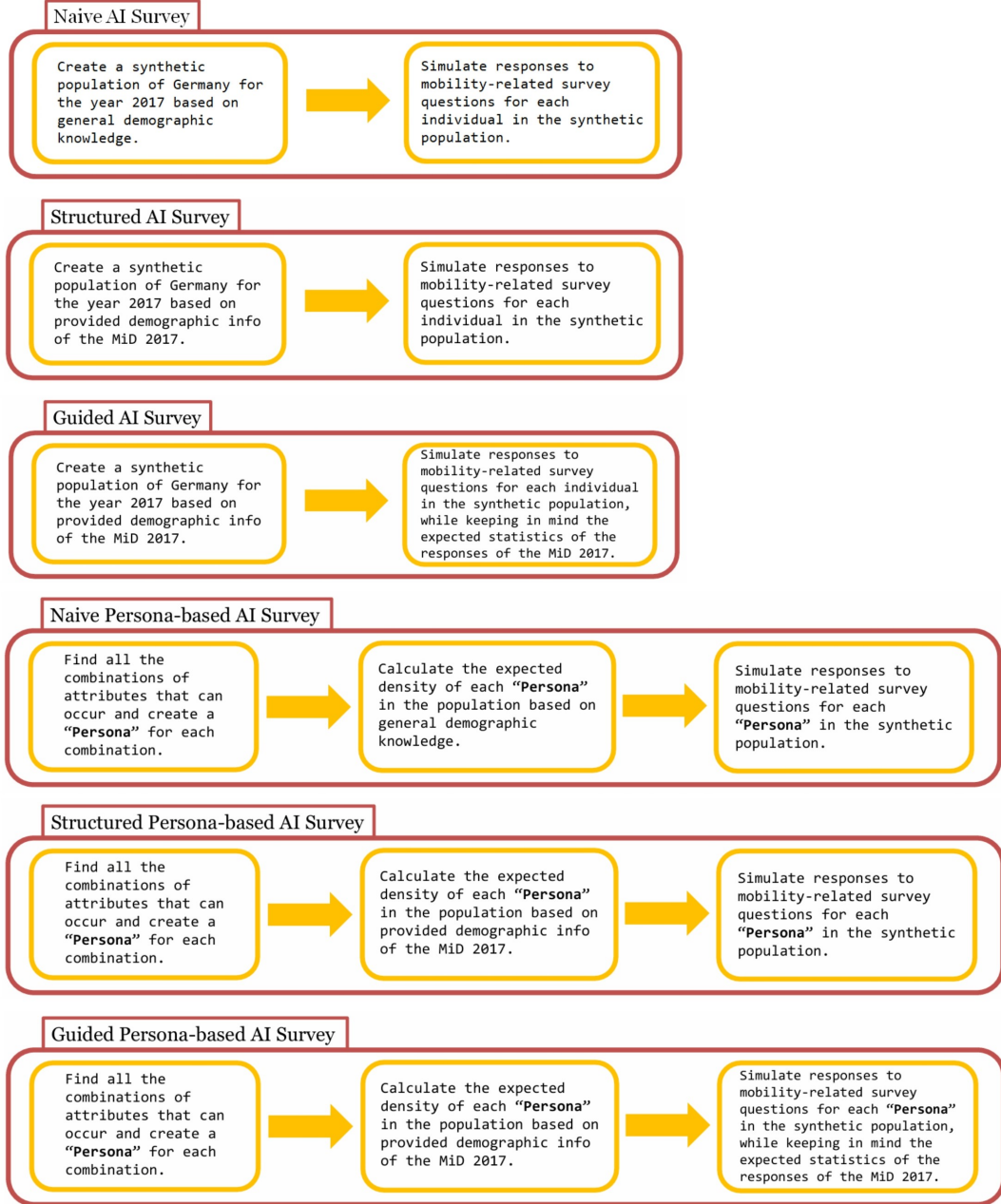
Figure 2: Overview of the Survey Generation Methods

Population Percentage of a Persona

The percentage of a Persona in the population is defined by accounting for the interdependencies among attributes using conditional probabilities. Let:

- $P_A(a)$ denote the probability of an individual belonging to age group $a \in A$.

- $P_E(e \mid a)$ denote the probability of an individual having education level $e \in E$ given their age group a.

- $P_M(m \mid a, e)$ denote the probability of an individual engaging in main activity $m \in M$ given their age group a and education level e.

- $P_S(s \mid a, e, m)$ denote the probability of an individual having economic status $s \in S$ given their age group a, education level e and main activity m.

- $P_H(h \mid a, e, m, s)$ denote the probability of an individual living in household type $h \in H$ given their age group a, education level e, main activity m and economic status s.

4

The percentage of Persona $P(a, e, m, s, h)$ in the population, denoted as π_P, is calculated as:

$$\pi_P = P_A(a) \cdot P_E(e \mid a) \cdot P_M(m \mid a, e) \cdot P_S(s \mid a, e, m) \cdot P_H(h \mid a, e, m, s), \tag{2}$$

where $a \in A$, $e \in E$, $m \in M$, $s \in S$ and $h \in H$.

This approach ensures that the correlations among attributes are respected, providing a more accurate representation of the population. The sum of π_P over all possible Personas equals 1, maintaining a complete representation of the population:

$$\sum_P \pi_P = 1. \tag{3}$$

Weighted Sums for Walking Preferences by Age

The walking preferences by age are calculated as a weighted sum of responses across Personas, where the weights are the population percentages of the Personas. Let:

- $R_i(a)$ be the response for walking preference i (e.g. "Completely Agree") for age group $a \in A$.

- π_P be the population percentage of Persona P.

- $R_i(P)$ be the probability that Persona P selects response i.

The weighted response for walking preference i for age group a is:

$$R_i(a) = \sum_{P \in A = a} \pi_P \cdot R_i(P), \tag{4}$$

where $P \in A = a$ indicates summation over all Personas whose age group is a. The responses for each age group are normalized to ensure they sum to 1:

$$\sum_i R_i(a) = 1, \quad \forall a \in A. \tag{5}$$

This method ensures that the preferences reflect both the demographic structure and the behavioural tendencies of the population.

3 RESULTS AND DISCUSSION

Evaluation Metrics

We evaluate the surveys based on the calculated percentages of each response for each attribute. The surveys were evaluated using:

- Statistical alignment with MiD benchmarks (distribution comparison).

- Error metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

- Jensen-Shannon (JS) Distance: A symmetric and normalized measure of divergence between distributions.

- Entropy: Measures the variability or uncertainty in a set of outcomes.

- Conditional Entropy: Assesses the remaining uncertainty in survey responses given demographic information.

- Cramér's V: Measures the strength of association between variables in real and synthetic data.

Table 1 summarizes the advanced statistical metrics for all survey methods.

Table 1: Advanced Statistical Metrics for AI-Generated Surveys

Survey Type	MAE	RMSE	JS Distance	Entropy	|Conditional Entropy|	Cramér's V
Naive AI Survey	10.53	11.35	0.19	3.54	0.24	0.35
Structured AI Survey	6.69	7.39	0.13	3.48	0.18	0.74
Guided AI Survey	3.37	4.65	0.067	3.28	0.014	0.68
Naive Persona-based AI Survey	9.02	10.28	0.157	3.48	0.19	0.73
Structured Persona-based AI Survey	8.53	9.73	0.152	3.48	0.18	0.73
Guided Persona-based AI Survey	0.03	0.17	0.0016	3.30	0.0005	1.00
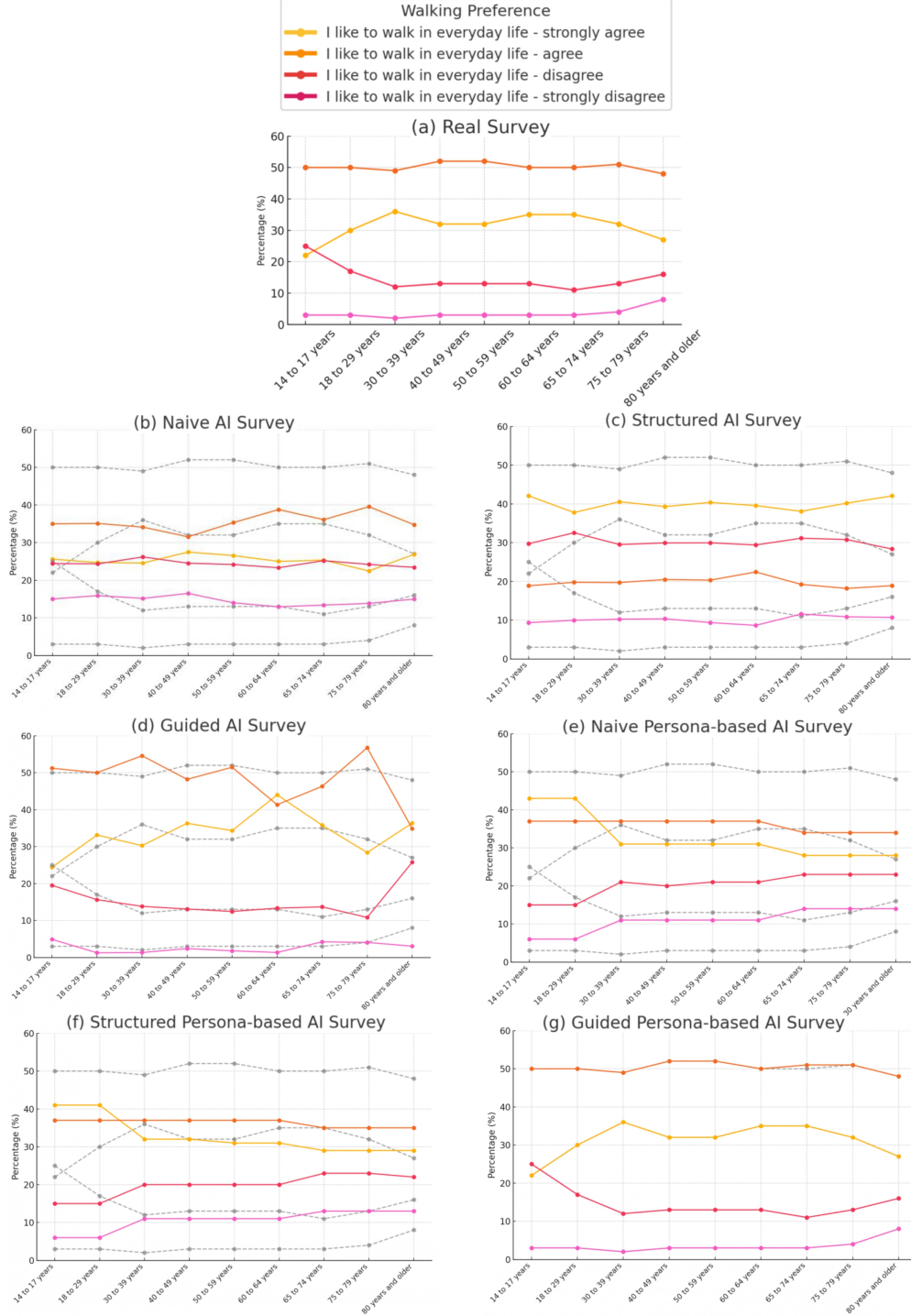
Figure 3: Comparative visualization of walking preferences by age group across real and synthetic surveys. Subplots: (a) Real Survey, (b) Naive AI Survey, (c) Structured AI Survey, (d) Guided AI Survey, (e) Naive Persona-based AI Survey, (f) Structured Persona-based AI Survey, (g) Guided Persona-based AI Survey.

6

Key Findings

- **Guided Persona-based AI Survey** exhibits the most outstanding performance among all methods. With the exceptionally low MAE (0.03), RMSE (0.17) and JS Distance (0.0016), it achieves near-perfect alignment with real survey data. These metrics highlight its unparalleled ability to model demographic-response dependencies effectively.

- **Guided AI Survey** also performs robustly, demonstrating relatively low error metrics (MAE = 3.37, RMSE = 4.65, JS Distance = 0.067). Although slightly less precise than the Guided Persona-based AI Survey, it offers a scalable solution with strong alignment across demographic patterns.

- **Structured Persona-based AI Survey** displays moderate performance with higher MAE (8.53) and RMSE (9.73) values. However, its Cramér's V value of 0.73 indicates a reasonable capacity to capture demographic patterns, albeit with greater variability.

- **Naive AI Survey** and **Structured AI Survey** exhibit limited alignment with real survey data. Their high MAE and RMSE values, coupled with lower JS Distance scores, underscore their struggles in capturing complex demographic-response relationships.

- **Naive Persona-based AI Survey** faces challenges in balancing variability and accuracy. With higher error metrics (MAE = 9.02, RMSE = 10.28) and moderate Cramér's V (0.73), it highlights the trade-offs in using an unconstrained Persona-based approach.

Comparative Visualization of Walking Preferences

Figure 3 illustrates the walking preferences across all age groups for both the real survey and synthetic methods. Subplot (a) represents the real survey, while (b)–(g) correspond to the six synthetic surveys. In subplots (b)-(g) the real survey results are also presented with grey lines, to visually illustrate the relative performance of the various methods.

4 Conclusions

The **Guided Persona-based AI Survey** emerges as the most effective method, demonstrating superior alignment with real survey data across all metrics. Its ability to incorporate both demographic alignment and response constraints allows for nuanced and realistic synthetic data generation. The **Guided AI Survey** also performs well, offering a scalable solution with robust error and alignment metrics.

In contrast, methods such as the **Naive AI Survey** and **Structured AI Survey** fall short in capturing intricate demographic-response dependencies, limiting their utility for high-fidelity synthetic data generation. Persona-based approaches, while versatile and diverse, require further optimization to balance variability with accuracy.

Future research should explore hybrid models that integrate the strengths of guided Persona-based techniques while enhancing computational efficiency and scalability. Such advancements will further validate the use of artificial surveys in diverse domains. Additionally, expanding the application to more cases, including different transportation modes from the MiD dataset as well as incorporating insights from other relevant datasets, can provide a more comprehensive evaluation.

Incorporating advanced statistical measures into the evaluation framework highlights the strengths and weaknesses of different synthetic survey methods. The Guided Persona-based AI Survey consistently outperforms other methods in alignment with real survey data, demonstrating the power of LLMs to generate realistic synthetic surveys. Future work should explore additional metrics, dynamic constraints and broader applications across various datasets and domains (Cho et al., 2024).

References

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. doi: 10.1017/pan.2024.5

Cho, S., Kim, J., & Kim, J. H. (2024). Llm-based doppelgänger models: Leveraging synthetic data for human-like responses in survey simulations. *IEEE Access*, *12*, 178917-178927. doi: 10.1109/ACCESS.2024.3502219

Choenni, S., Busker, T., & Bargh, M. S. (2023). Generating synthetic data from large language models. In *2023 15th international conference on innovations in information technology (iit)* (p. 73-78). doi: 10.1109/IIT59782.2023.10366424

Liu, R., Wei, J., Liu, F., Si, C., Zhang, Y., Rao, J., . . . Dai, A. M. (2024). *Best practices and lessons learned on synthetic data.* Retrieved from `https://arxiv.org/abs/2404.07503`

Livieris, I. E., Alimpertis, N., Domalis, G., & Tsakalidis, D. (2024). *An evaluation framework for synthetic data generation models.* Retrieved from `https://arxiv.org/abs/2404.08866`

Ma, B., Yoztyurk, B., Haensch, A.-C., Wang, X., Herklotz, M., Kreuter, F., . . . Assenmacher, M. (2024). *Algorithmic fidelity of large language models in generating synthetic german public opinions: A case study.* Retrieved from `https://arxiv.org/abs/2412.13169`

MiD. (2017). *Mid 2017 tabellenband deutschland.* Retrieved from `https://www.mobilitaet-in-deutschland.de/archive/pdf/MiD2017_Tabellenband_Deutschland.pdf`

Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., . . . Liu, H. (2024). *Large language models for data annotation and synthesis: A survey.* Retrieved from `https://arxiv.org/abs/2402.13446`

Tzachristas, I. (2024). *Creating an llm-based ai-agent: A high-level methodology towards enhancing llms with apis.* Retrieved from `https://arxiv.org/abs/2412.13233`

Wang, K., Zhu, J., Ren, M., Liu, Z., Li, S., Zhang, Z., . . . Wang, Y. (2024). *A survey on data synthesis and augmentation for large language models.* Retrieved from `https://arxiv.org/abs/2410.12896`

MiD Statistics

		Walking Preference				Cycling Preference				Driving Preference				Bus/Train Preference				Overnight Trips				
		strongly agree	agree	disagree	strongly disagree	strongly agree	agree	disagree	strongly disagree	strongly agree	agree	disagree	strongly disagree	strongly agree	agree	disagree	strongly disagree	0	1	2	3	4+
Age Group	14 to 17 years	22	50	25	3	32	42	19	8	26	45	18	12	12	44	35	8	44	28	12	7	8
	18 to 29 years	30	50	17	3	24	35	29	13	32	43	18	8	8	33	38	20	40	24	14	8	14
	30 to 39 years	36	49	12	2	27	38	25	10	32	45	17	7	7	27	41	25	36	24	15	10	15
	40 to 49 years	32	52	13	3	26	38	24	11	34	46	15	5	6	22	45	27	42	25	14	9	10
	50 to 59 years	32	52	13	3	26	38	25	12	32	47	15	6	6	23	45	26	44	26	14	7	9
	60 to 64 years	35	50	13	3	25	37	24	13	32	48	14	6	7	27	43	23	46	26	14	9	5
	65 to 74 years	35	50	11	3	25	35	23	18	34	48	13	5	7	30	41	22	47	28	14	7	4
	75 to 79 years	32	51	13	4	19	31	25	25	32	48	13	7	9	35	35	20	58	24	10	5	3
	80 years and older	27	48	16	8	12	19	31	38	28	44	15	13	9	37	31	23	71	20	5	3	1
Education Level	no degree (yet)	24	51	21	3	26	38	23	13	29	44	17	11	10	43	34	12	49	27	11	6	7
	low	28	53	14	5	19	36	26	19	32	49	12	7	6	26	42	26	63	23	8	3	3
	medium	31	52	14	3	23	38	27	13	36	46	13	5	6	25	43	26	44	27	14	7	8
	high	42	46	10	2	36	33	21	10	25	43	22	9	9	34	39	18	23	24	19	14	21
Main Activity	full-time employee (including trainees)	31	52	15	2	25	39	26	10	36	45	15	5	6	24	44	27	37	25	14	9	14
	part-time employee (11 to 35 hours per week)	37	50	10	2	30	36	23	10	30	48	16	6	6	26	44	24	38	29	17	8	7
	Employed person without details on the extent of education	36	47	13	4	29	40	18	13	32	46	15	7	7	31	42	21	47	22	14	9	7
	Pupil including preschool	22	52	24	2	30	40	22	9	27	49	15	9	11	44	36	9	45	28	13	7	8
	Student	41	47	11	1	39	30	21	10	20	41	26	13	12	48	32	8	28	23	16	12	21
	housewife/househusband	36	48	13	3	22	32	28	17	31	47	16	6	6	26	43	24	55	25	11	5	4
	Pensioner not employed	32	50	13	5	21	30	25	24	32	48	13	8	8	32	38	21	55	26	11	6	3
	other activity	33	50	13	4	23	33	27	16	28	44	18	10	11	29	38	22	58	20	11	5	7
Economic Status	very low	29	53	15	4	23	35	26	17	26	43	18	12	11	33	38	18	67	17	6	3	6
	low	30	53	13	3	22	35	26	17	30	47	15	7	8	29	39	23	57	25	9	4	5
	medium	33	50	14	3	24	36	26	16	32	45	15	7	5	29	41	23	45	27	14	7	7
	high	33	52	13	2	28	39	24	9	34	47	14	4	7	24	45	26	31	26	17	11	14
	very high	37	46	14	2	31	35	23	11	34	43	18	5	7	25	44	24	22	23	17	13	24
Household Type	young people living alone	34	44	20	2	30	30	28	13	29	41	19	12	11	34	33	21	32	24	16	10	18
	middle-aged singles	34	49	13	3	27	34	25	14	27	42	19	11	9	29	41	22	49	19	11	8	13
	older singles	32	51	13	4	19	29	26	26	28	43	17	12	12	38	33	17	60	23	8	5	4
	young two-person households	34	51	13	2	25	38	27	11	31	44	17	8	9	30	39	22	34	22	15	11	17
	middle-aged two-person households	33	52	13	3	25	38	25	13	33	46	16	5	6	24	42	27	38	27	15	8	11
	older two-person households	35	50	12	4	24	34	24	18	34	50	12	4	6	29	42	23	47	28	14	8	4
	Households with at least 3 adults	28	54	15	4	23	37	27	14	34	47	14	5	5	26	44	24	47	26	12	7	8
	Households with at least 1 child under 6 years	38	49	12	1	29	38	23	10	32	46	16	5	6	24	44	26	37	27	15	9	12
	Households with at least 1 child under 14 years	30	53	14	2	27	40	24	8	35	47	14	4	5	23	46	26	36	30	16	8	10
	Households with at least 1 child under 18 years	27	50	19	3	26	39	26	10	34	48	14	5	5	29	41	24	46	24	15	6	8
	Single parents	34	52	13	2	27	37	26	10	34	42	16	8	11	23	47	19	47	22	12	9	10

Figure 4: Normalized MiD 2017 dataset