

LLM-PDM: An LLM Persona-Driven Method for replicating personal mobility preferences at scale

Ioannis Tzachristas^{1,2,✉}, Santhanakrishnan Narayanan², Constantinos Antoniou²

 **Cite this article:** Tzachristas I, Narayanan S, Antoniou C. *Commun Transp Res* 2026, 6(1): 9640004. <https://doi.org/10.26599/COMMTR.2026.9640004>

ABSTRACT: Traditional travel surveys are costly, time-consuming and face declining response rates, motivating the exploration of artificial data generation methods. In this research, we propose a novel Persona-Driven Method (PDM) for generating synthetic mobility survey data via large language models (LLMs). The method defines representative personas—each characterized by specific sociodemographic attributes—and prompts an LLM to emulate survey respondents with these personas. A guided prompting strategy is introduced to calibrate the synthetic data distributions so that they closely match real-world population statistics. We evaluate the approach on the German MiD 2017 (Mobilität in the Deutschland 2017) dataset. The quality of the LLM-PDM-generated synthetic data is assessed against ground truth data via a comprehensive set of metrics, including the mean absolute error (MAE), root mean square error (RMSE), Jensen–Shannon distance (JSD), entropy, conditional entropy and the Earth Mover’s distance (EMD). The empirical results demonstrate that the LLM-PDM approach produces high-fidelity synthetic populations that preserve key distributions and relationships present in real data. Across the case studies, the LLM-PDM method achieves low distributional errors (e.g., MAE < 3%) and captures important joint patterns, significantly outperforming a number of LLM baselines.

KEYWORDS: synthetic data generation; travel survey; large language models (LLMs); Persona-Driven Modeling (PDM); prompt engineering; synthetic populations; mobility data

1 Introduction

Fig. 1 contrasts the traditional travel-survey workflow with the guided persona-based AI survey workflow proposed in this study.

Travel survey data are pivotal for transportation planning and policy, providing insights into individuals’ mobility patterns, mode choices and activity behavior. However, traditional survey methods face growing challenges, including high administration costs, lengthy deployment times and declining participation rates. Moreover, concerns about privacy and data sparsity (especially for emerging modes or rare marginal sociodemographic groups) further limit the utility of conventional survey approaches (Narayanan and Antoniou, 2022, 2023a, 2023b). These issues have spurred steadily growing literature on synthetic data generation as an alternative—or complement—to traditional survey collection (Narayanan et al., 2025; Tzachristas et al., 2025), ranging from statistical resampling frameworks, such as synthpop (Nowok et al., 2016), to recent LLM-based “survey doppelgängers” (Bisbee et al., 2024; Cho et al., 2024; Veselovsky et al., 2023). Synthetic travel datasets can potentially mirror real-world behavioral patterns without burdening respondents while protecting privacy by avoiding real personal data (Liu et al., 2024; Nowok et al., 2016). Recent advances in generative models and artificial intelligence have opened new avenues for creating realistic synthetic datasets

in various domains. In particular, large language models (LLMs), such as GPT-3.5 and GPT-4, have demonstrated a unique ability to produce human-like text and to capture statistical patterns from large corpora of data. Researchers have begun to explore LLMs as tools for generating artificial survey responses. Early social science studies tested whether LLMs could serve as “synthetic respondents,” showing that the generated data can mimic some aggregate distributions but also exhibit biases and overfitting issues (Bisbee et al., 2024; Cho et al., 2024; Veselovsky et al., 2023). In the transportation domain, this potential remains largely untapped. Only very recently have LLMs been applied to travel survey contexts, with pilot results indicating that they can reproduce basic mobility patterns when carefully guided (Liu et al., 2024; Tan et al., 2024). Our LLM-PDM adds explicit persona-conditioning and a light calibration loop and is evaluated across the MiD 2017 (MiD, 2017) dataset (Bundesministerium für Verkehr und digitale Infrastruktur, 2018), a comprehensive dataset for Germany, providing a broad test of the approach. The approach relies entirely on prompt engineering techniques to enable in-context learning without requiring any model fine-tuning. This motivates a deeper investigation into how an LLM might effectively generate comprehensive travel survey datasets that align with real-world observations.

¹ Huawei Munich Research Center, Munich 80992, Germany. ² Chair of Transportation Systems Engineering, School of Engineering and Design, Technical University of Munich, Munich 80333, Germany.

✉ Corresponding author. E-mail: ioannis.tzachristas@tum.de

Received: July 10, 2025; Revised: September 19, 2025; Accepted: November 7, 2025

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, <http://creativecommons.org/licenses/by/4.0/>).

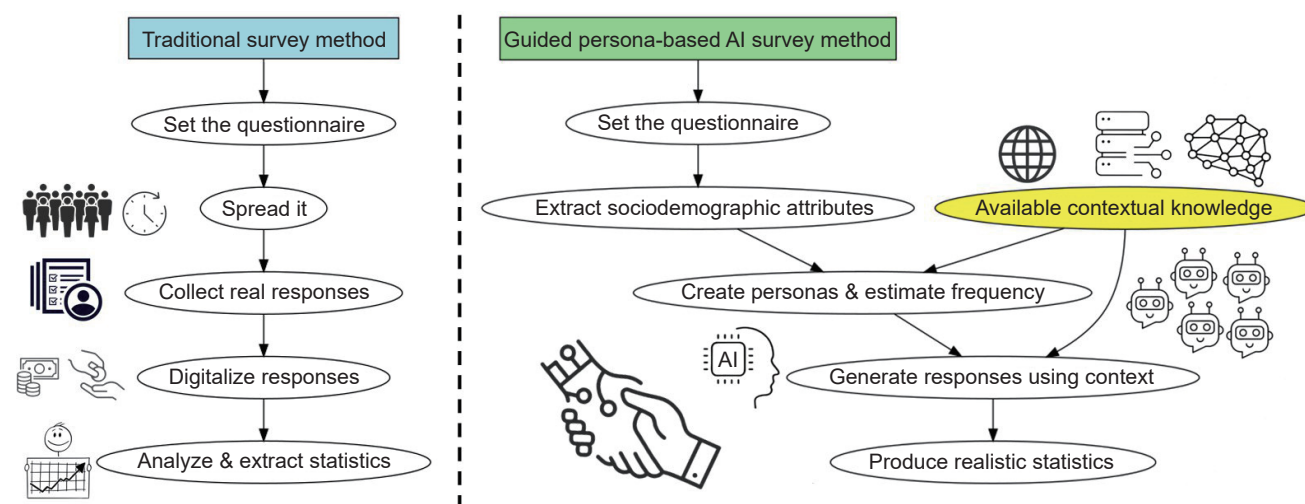


Fig. 1 Traditional survey method vs. guided persona-based AI survey method.

2 Literature review

2.1 Synthetic data generation in travel surveys

Researchers have long sought methods to generate synthetic populations and travel diaries that statistically resemble real survey data. Early approaches in transportation focused on the population synthesis—creating a set of synthetic individuals and households that match known marginal distributions from censuses or surveys. A classic method by Beckman et al. (1996) employs iterative proportional fitting (IPF) to combine census tables and sample data, producing a baseline synthetic population for microsimulation. This approach and its variants have been widely used in activity-based modeling and travel-demand forecasting. Subsequent improvements introduced techniques such as hierarchical IPF, metaheuristic sampling and Bayesian networks to better preserve joint distributions of attributes (Nowok et al., 2016; Sun and Erath, 2015). Sun and Erath (2015) applied a Bayesian-network approach for population synthesis that improved the capture of interattribute correlations compared with IPF. These traditional methods typically rely on statistical reweighting of observed data and are limited by the need for an initial sample and by assumptions of independence between attributes beyond certain marginals.

In parallel, the rise of machine learning has led to the development of generative models for synthetic travel data. For example, Kapp et al. (2024) explored variational autoencoders to generate individual daily activity patterns. Recently, more complex models, such as generative adversarial networks (GANs), have been investigated the production of synthetic data for specific transportation applications. These ML-based methods can capture nonlinear relationships and produce realistic multidimensional data, but they often require large training datasets and careful tuning to avoid mode collapse—a failure mode in which the generator “collapses” onto a small subset of patterns and repeats them, thereby underrepresenting the true diversity of the target distribution—or outright overfitting (Havrilla et al., 2024). Moreover, purely data-driven models may struggle to incorporate domain knowledge or ensure that outputs adhere to known aggregate constraints without additional adjustments.

Within the broader context, synthetic data generation has also been pursued as a means to protect privacy. Tools such as the R package synthpop (Nowok et al., 2016) generate synthetic microdata from sensitive datasets, allowing analysis without exposing real individual records. The challenge in all these

approaches is balancing fidelity (how closely synthetic data reproduce real patterns) and diversity (not replicating exact data points). An evaluation framework by Livieris et al. (2024) emphasized multiple aspects of synthetic data quality—from distributional similarity to preservation of correlations and usefulness in downstream tasks. We adopt a similar multifaceted evaluation in this study.

2.2 LLMs in survey data simulation

The advent of large-scale pretrained language models has spurred interest in their use to simulate human survey responses. Unlike traditional generative models that require structured training data, LLMs leverage knowledge acquired from massive text corpora. Studies in political science and social psychology have tested whether LLMs can emulate survey respondents for attitude and opinion surveys. Bisbee et al. (2024) critically examined the use of the GPT-3 to replace human survey data, cautioning that although aggregate distributions could sometimes be matched, the LLM tended to mirror popular-opinion narratives and failed to capture nuanced subgroup differences (hence the “perils of large language models” in survey replacement). Similarly, Ma et al. (2024) evaluated the algorithmic fidelity of LLM-generated public opinion data in Germany and reported that carefully prompting LLMs can reproduce broad trends but may inject artifacts if unfamiliar scenarios are asked about.

On the more optimistic side, Cho et al. (2024) introduced the concept of LLM-based Doppler, in which a model is fine-tuned or prompted to represent a specific individual’s responses across many questions. They reported that an LLM (GPT-3) could generate responses that fooled human judges and even predictive models into thinking that the answers were from real humans. Choenni et al. (2023) presented one of the first systematic attempts at generating synthetic data from LLMs, demonstrating that with a few examples and constraints, even an off-the-shelf model can yield datasets that approximate the original data’s statistical properties. However, that study focused on relatively small-scale data (generation of a few variables) and noted significant challenges in maintaining correlation structures without additional controls.

In the field of transportation, the use of LLMs for synthetic travel survey generation is at a nascent stage. Nonetheless, recent investigations have demonstrated the broader behavioral potential of these models—notably their ability to deliver explainable lane-change trajectory predictions (Peng et al., 2025) and to power modular AI agents capable of automating substantial portions of

the survey-interview workflow (Yu et al., 2025). One related stream of work is the idea of persona-based simulation, where domain knowledge about traveler types is used to guide data generation. Our work builds on this foundation, formalizing the approach and generalizing it to multiple datasets and a broader set of survey variables, as well as providing a more extensive validation framework.

3 Methodology

3.1 Overview of the LLM-PDM methodology

Fig. 2 outlines the overall process of the proposed LLM-PDM methodology. The core idea is to use an LLM as a generative engine, steered by personas and guided prompts, to produce synthetic individuals' survey data. The workflow comprises three main components: (1) persona definition and sampling, (2) LLM prompt design and response generation, and (3) guided distribution calibration. Together, these factors ensure that the generated dataset not only appears realistic on a case-by-case basis but is also aggregated to correct population-level patterns.

3.2 LLM prompt design and response generation

For each sampled persona $z^{(j)}$, we generate a synthetic survey response with an LLM. The interaction is driven by a carefully structured prompt. A generic template—shown below with placeholder values—captures both the persona context and survey questions:

You are a survey respondent with the following profile:
[age: 45, gender: F, employment: full-time, household: 4 persons, ...].
I will ask you some questions about your travel behavior.
Answer as this person would, concisely and realistically.
Q1. What is your main mode of transportation on a typical day?
Q2. How many trips do you make on an average weekday?
Q3. ...
Additional context: ensure your answers remain consistent with the profile and typical for someone like you.

In production, the bracketed metadata (e.g., [age: X , gender: Y]) are filled with the attribute values of persona $z^{(j)}$; the question block (Q1, Q2, ...) mirrors the wording of the real survey items. When the original wording is excessively technical, we simplify the language while preserving intent, thereby maintaining facial validity for the LLM. This prompt casts the model into a role-playing scenario, akin to persona-based few-shot prompting, constraining the generative space to plausible answers for that demographic profile. Pilot experiments confirmed that explicitly embedding the profile reduces contradictions (e.g., retirees “commuting to work”) and that the italicized consistency cue further anchors responses to the demographic context.

We use the GPT-4 model (OpenAI, 2023) for its proven instruction following and coherence; however, the method is model-agnostic and can leverage any suitably capable LLM. Each persona is sent once to the model, yielding one record in the synthetic dataset. Answers are requested in a parse-friendly format (e.g., “A1: ...; A2: ...”), after which lightweight postprocessing extracts structured fields.

Ensuring variability: Repeatedly prompting with identical persona inputs should not produce carbon-copy outputs. We introduce controlled stochasticity via temperature scaling and nucleus sampling. Formally, let $R^{(j)}$ denote the response token sequence for persona $z^{(j)}$; we write:

$$R^{(j)} \sim \text{LLM}(\text{prompt}(z^{(j)}); T, p, \text{max_tokens}) \quad (1)$$

where $\text{LLM}(\cdot; \cdot)$ denotes a stochastic decoding operator that samples tokens from the model's conditional distribution. T is the temperature that scales logits before the softmax ($T > 1$ increases randomness; $T < 1$ makes decoding more conservative), and p is the nucleus mass (also called top p), meaning that we sample from the smallest set of tokens whose cumulative probability is at least p at each step. Max_tokens caps the generated length. Unless stated otherwise, we use $T = 0.7$, $p = 0.95$ and max_tokens = 2048. Under this interface, $\text{Var}[R^{(j)}] > 0$ even when $z^{(j)}$ is held fixed—empirically yielding diverse yet persona-consistent answers and avoiding mode collapse that would otherwise reduce the usefulness of the synthetic corpus (Livieris et al., 2024; Veselovsky et al., 2023).

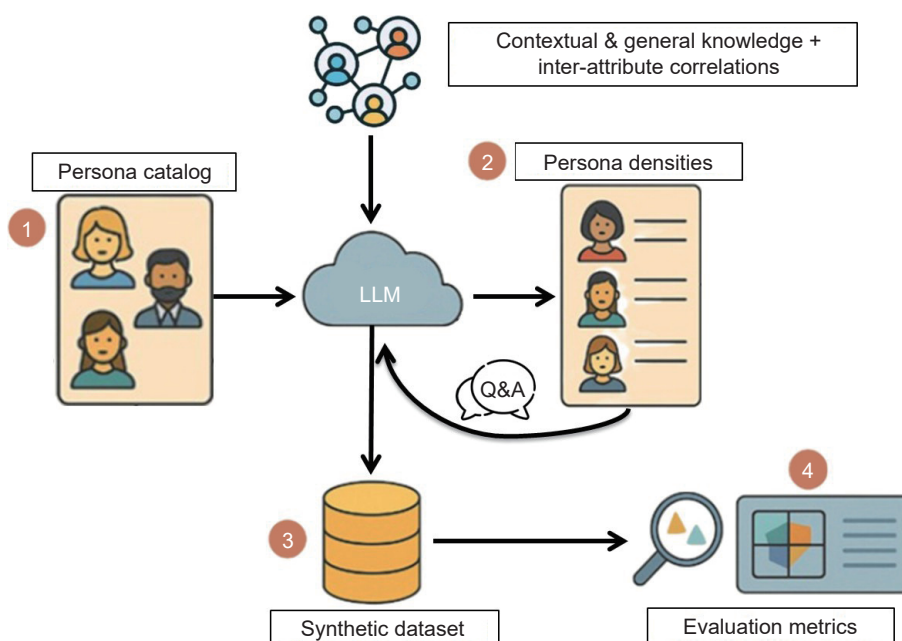


Fig. 2 LLM-PDM methodology overview.

3.3 Guided promotion and distribution calibration

After a first pass of synthetic generation, we compare the new dataset with the real marginal distributions. Deviations are inevitable: car share may land at 35% instead of the required 40%, or the model may favor neat trip counts (2, 4, 6), biasing the mean. Rather than “hard-wiring” the exact numbers into the prompt, we refine the process through two complementary layers of guidance:

1) Contextual cues. Each persona prompt begins with broad, publicly known facts that set the travel landscape—fuel prices, rail coverage, cycle path density, and recent policy headlines. These cues steer mode choice without revealing the survey’s precise marginals. Tan et al. (2024) showed that context statements alone can halve distributional error.

2) Behavioral examples. For every question block, we prepend one or two short, human-written trip diaries for a closely matching profile, which implicitly teach the model how distances, purposes, modes, and times cohere across answers; for example: “Anna (32, employee, no car): walk 6 min → bus 4 km → work 09:00–17:00 → bus → walk, Jonas (16, pupil, bike available): bike 2.8 km to school at 08:00, bike back at 13:30, later walk 1.0 km to sports, and Weber (72, pensioner, transit pass): walk 0.6 km → tram 5.2 km to clinic at 10:00, return by tram, no evening trips”. With such exemplars in place, generations for similar personas more reliably satisfy plausibility checks—short distances by walk/bike, mid-range by bus/tram, no car use in car-free households, commute peaks at realistic times—whereas removing exemplars in spot checks occasionally yielded incoherent combinations (e.g., a no-car household “drives” 25 km to work or a 75+ persona “walks” 10 km), confirming the instructional effect.

This two-layer recipe proved more effective than blunt numeric hints: the LLM-PDM’s mode-share mean absolute error (MAE) falls below 3% in all regions, and joint metrics approach the survey baseline while the textual richness of answers is preserved.

A reviewer-in-the-loop variant—letting the GPT-4 critique and revising its own outputs—showed promise, echoing the RLHF ideas noted by Veselovsky et al. (2023), but this is left to future work.

The result is a calibrated synthetic dataset whose margins and key dependencies match reality within sampling error but whose individual records remain novel creations of the model, safeguarding privacy and avoiding memorization.

Mitigating data leakage:

In designing the guided prompts, we took care to avoid any inadvertent data leakage from the MiD 2017 dataset. Notably, no actual MiD survey microdata or exact statistics were embedded in the prompts. The broad contextual facts (e.g., high fuel prices, recent transit expansions) are generic, publicly known information—they steer the model without disclosing specific MiD results. The LLM thus has no direct access to proprietary survey figures. Moreover, the small number of aggregate targets we do use (overall mode split, etc.) are imparted only through abstract “hints” rather than explicit values and are chosen from common knowledge ranges. Because MiD 2017 microdata are confidential (and unlikely to have appeared in the LLM’s training corpus), the model’s output relies on learned mobility patterns, not memorization of the MiD data. In summary, our prompt engineering prevents leakage of sensitive or exact survey data to the LLM. The LLM-PDM results stem from genuine model reasoning guided by contextual cues rather than regurgitation of training data or hidden prompts with ground-truth answers.

3.4 Evaluation metrics

To evaluate the quality of the synthetic survey data, we employ a

suite of metrics covering marginal error, distributional shape, information-theoretic divergence, categorical association and distance on continuous variables.

1) Statistical alignment with MiD benchmarks (distribution comparison).

2) Error metrics: MAE and root mean square error (RMSE).

3) Jensen–Shannon Distance (JSD): A symmetric and normalized measure of divergence between distributions.

4) Entropy: a measure of variability or uncertainty in a set of outcomes.

5) Conditional entropy: a measure of the remaining uncertainty in survey responses given demographic information.

Formal definitions of evaluation metrics

Distributional error metrics. For a categorical variable with categories $i = 1, \dots, m$, let p_i be the true proportion in the real data and $\hat{p}_i$ be the proportion in the synthetic data. The MAE and RMSE are

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^m |\hat{p}_i - p_i|, \text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (\hat{p}_i - p_i)^2} \quad (2)$$

Both are reported in percentage-point units; for numeric quantities (e.g., mean trip count), the MAE and RMSE are computed analogously.

JSD: Let $P = \{p_i\}$ and $Q = \{\hat{p}_i\}$ be real and synthetic distributions, $M = \frac{1}{2}(P + Q)$ their midpoint. The Jensen–Shannon divergence is

$$D_{\text{JS}}(P \| Q) = \frac{1}{2} \sum_i p_i \log_2 \frac{p_i}{m_i} + \frac{1}{2} \sum_i \hat{p}_i \log_2 \frac{\hat{p}_i}{m_i} \quad (3)$$

and the JSD:

$$\text{JSD}(P, Q) = \sqrt{D_{\text{JS}}(P \| Q)} \quad (4)$$

which lies in $[0, 1]$ and is a true metric. We use JSD for multicategory distributions such as four-way mode shares.

Shannon entropy and conditional entropy: For a categorical variable X with PMF P , the Shannon entropy is

$$H(X) = -\sum_x P(x) \log_2 P(x) \quad (5)$$

The entropy difference $\Delta H = |H_{\text{synth}} - H_{\text{real}}|$ indicates over- or underdispersion in synthetic data. The conditional entropy for X given Y is

$$H(X|Y) = -\sum_y P(y) \sum_x P(x|y) \log_2 P(x|y) \quad (6)$$

We track $\Delta H_{\text{cond}} = |H_{\text{synth}}(X|Y) - H_{\text{real}}(X|Y)|$; large values signal lost dependencies.

Earth Mover’s Distance (EMD): For one-dimensional continuous or ordered variables, EMD equals the area between cumulative distribution functions; in general,

$$\text{EMD}(F, G) = \inf_{\gamma} \sum_{x,y} \gamma(x, y) d(x, y) \quad (7)$$

where γ is a transport plan with marginals F and G and where d is the ground distance. We bin variables such as daily travel distance and compute EMD in native units (km or minutes).

The above metrics—marginal (MAE, RMSE), shape (JSD), information-theoretic (ΔH , ΔH_{cond}), association (V) and continuous-distribution distance (EMD)—form the multifaceted

framework recommended by Livieris et al. (2024) and Wang et al. (2024).

3.5 Pseudocode for the LLM-PDM

Algorithms 1–3 summarize the LLM-PDM workflow, including persona enumeration, guidance-based calibration, and weight computation.

Algorithm 1 LLM-PDM — Persona-Driven LLM Survey Synthesis

```

1: procedure LLM-PDM_Synthetic Survey( $D, A, G, N$ )  $\triangleright D$ : real survey,
   $A$ : attribute defs,  $G$ : guidance stats,  $N$ : rows
2:  $P \leftarrow \text{EnumeratePersonas}(A)$ 
3:  $\pi \leftarrow \text{ComputeWeights}(P, D)$ 
4:  $nP \leftarrow \lfloor \pi(P) N \rfloor \quad \forall P \in P$ 
5:  $S \leftarrow \emptyset \triangleright$  synthetic row store
6: for all  $P \in P$  in parallel do
7:    $templ \leftarrow \text{BuildPromptTemplate}(G)$ 
8:    $prompt \leftarrow \text{FillPrompt}(templ, P)$ 
9:   for  $i \leftarrow 1$  to  $nP$  do
10:     $t \leftarrow \text{CallLLM}(prompt)$ 
11:     $r \leftarrow \text{ParseOutput}(t)$ 
12:     $S \leftarrow S \cup \{r\}$ 
13:   end for
14: end for
15:  $m \leftarrow \text{Evaluate}(S, D)$ 
16: if  $m$  violates thresholds then
17:    $G \leftarrow \text{UpdateGuidance}(m, G)$ 
18:   goto line 4  $\triangleright$  optional calibration loop
19: end if
20:  $\text{Save}(S, \text{"LLM-PDM output.csv"})$ 
21: return  $m$ 
22: end procedure

```

Algorithm 2 EnumeratePersonas

```

1: function EnumeratePersonas( $A$ )
2:   return Cartesian product of all category lists in  $A$ 
3: end function

```

Algorithm 3 ComputeWeights

```

1: function ComputeWeights( $P, D$ )
2: for all  $P(A, E, M, S, H) \in P$  do
3:    $\pi(P) \leftarrow P(A) P(E|A) P(M|A, E) P(S|A, E, M) P(H|A, E, M, S)$ 
4:   Normalize  $\pi$  so that  $\sum P \pi(P) = 1$ 
5:   return  $\pi$ 
6: end function

```

4 Dataset

We evaluate LLM-PDM on the (most recent) MiD 2017 survey, a nationally representative German household-travel study conducted between May 2016 and September 2017. The public microdata comprise $\approx 316,000$ individuals in $\approx 156,000$ households, each providing the following:

- 1) Household file: socioeconomic characteristics (size, income, car count, urbanity).
- 2) Person files: demographic attributes (age, gender, education, employment) and mobility resources (license, season tickets).
- 3) Trip diary: all trips on one assigned reference day with origin/destination, purpose, main mode, distance, and duration.

The design weights supplied by the survey agency adjust for sampling strata and nonresponses; we use the calibrated person-

level weights for all “ground-truth” distributions. After excluding respondents younger than 14 years (to align with adult mobility analyses) and cases with missing key attributes ($< 0.5\%$), the analytic base contains $n = 151,728$ persons and $m = 422,615$ weighted trips. The official documentation and codebook are available from the German Federal Ministry of Transport (Bundesministerium für Verkehr und digitale Infrastruktur, 2018).

5 Experimental design

We normalize the distributions of the translated to English versions of the MiD 2017 dataset so that the percentage of the “not specified” option is merged with the most popular responses of each question, as shown in Appendix F. We employ the LLM-PDM method to calculate the mode share distribution in Germany in 2017 and compare the results with the real values according to the MiD 2017 survey. Then, we extract statistics related to mobility preference from the MiD 2017 and create a comprehensive table, as shown in Appendix F. We use these statistics to examine how well LLMs can capture the attribute intercorrelations by performing a “variation study”.

5.1 LLM-PDM variations

We introduce variations of the LLM-PDM. For this sequence of experiments, we select and perform a focused analysis on the aforementioned dataset. For AI survey generation, the GPT-4 API was utilized in a framework inspired by previous work (Tzachristas, 2024). Persona-based methods involve 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, and Fig. 3 provides an overview of the methods.

Naive AI survey:

A synthetic population for 2017 Germany was generated on the basis of 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 the responses were not explicitly aligned with 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 behavioral attributes were identified to define personas. The density of each persona was estimated on the basis of their general demographic knowledge. Mobility-related responses were then simulated for each persona.

Structured persona-based AI survey:

This method enhances the Naive persona-based AI survey by aligning persona densities with MiD 2017 demographic data.

Fully guided persona-based AI survey:

Building upon the Structured persona-based AI Survey, this method incorporates expected response statistics from the MiD 2017 dataset.

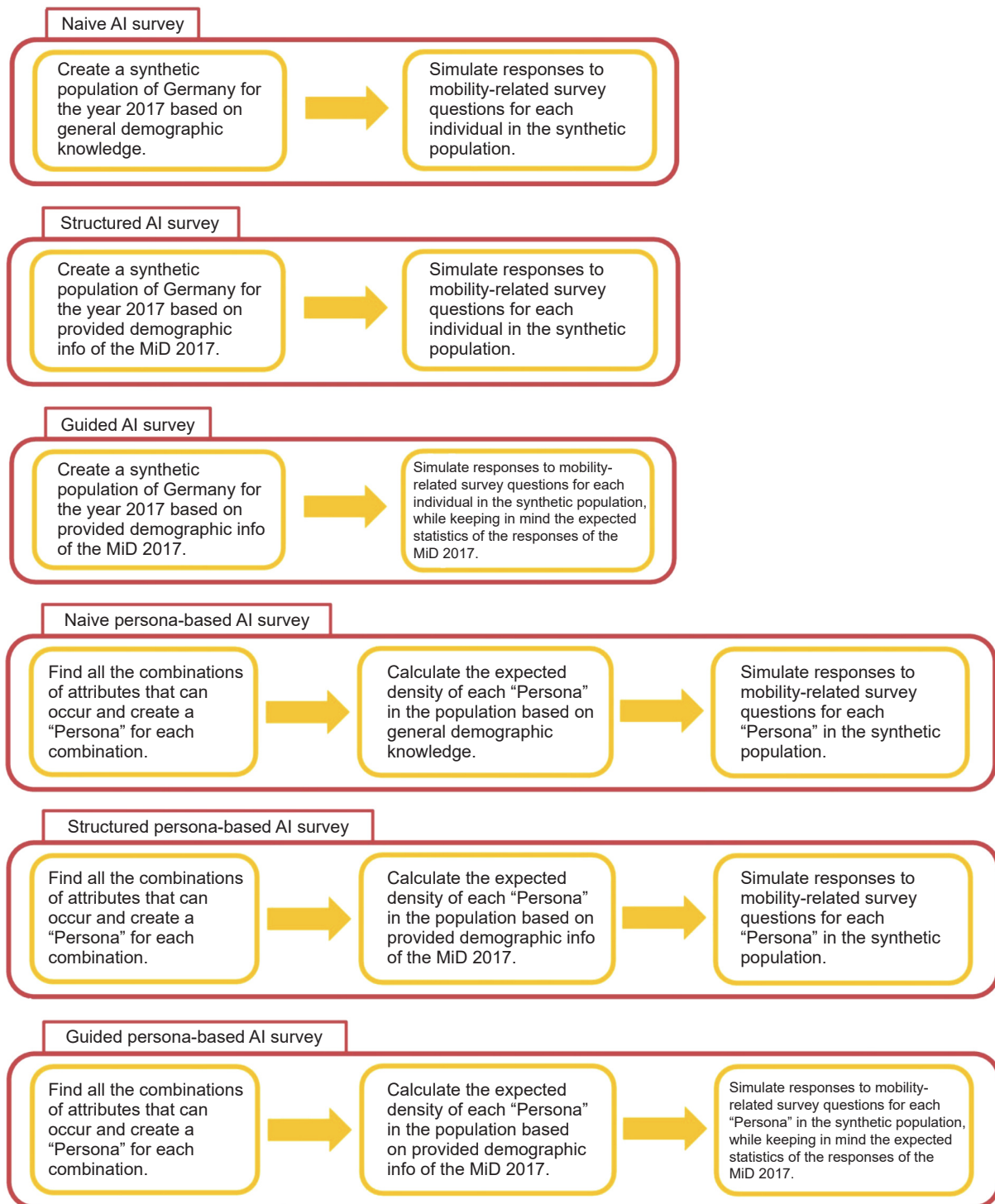


Fig. 3 Overview of the survey generation methods.

The prompts used in the experiments were specifically crafted to guide GPT-4's generation of realistic populations and survey responses. These prompts ranged from straightforward instructions (e.g., "Generate a population on the basis of 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 on the basis of MiD 2017 data"). This approach ensures high fidelity in response generation and allows each method to address specific research goals while capturing intricate interdependencies between attributes.

5.2 Definition of a persona in the context of MiD 2017

A persona is a tuple defined by a specific combination of demographic and behavioral attributes. Let the attributes be as follows:

- 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 $P(A, E, M, S, H)$, where $A \in \text{Age Group}$, $E \in \text{Education Level}$, $M \in \text{Main Activity}$, $S \in \text{Economic Status}$, and $H \in \text{Household Type}$.

5.3 Population percentage of personas

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

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

2) $P_E(e | a)$ denotes the probability of an individual having education level $e \in E$ given their age group a .

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

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

5) $P_H(h | a, e, m, s)$ denotes 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 .

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

$\pi_p = P_A(a) P_E(e|a) P_M(m|a, e) P_S(s|a, e, m) P_H(h|a, e, m, s)$ 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:

$$\sum_p \pi_p = 1 \quad (8)$$

5.4 Weighted sums for walking preferences by age

We chose to examine walking preferences across different age groups, as these preferences are expected to exhibit substantial variability with 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

1) $R_i(a)$ is the response for walking preference i (e.g., “Completely Agree”) for age group $a \in A$.

2) π_p is the population percentage of persona P .

3) $R_i(P)$ is the probability that persona P selects response i .

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

$$R_i(a) = \sum_{P \in A=a} \pi_p \cdot R_i(P) \quad (9)$$

where the summation is taken over all personas P whose age group equals a . The responses for each age group are normalized to ensure that they sum to 1:

$$\sum_i R_i(a) = 1, \forall a \in A \quad (10)$$

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

5.5 Counting the persona state space in the context of MiD 2017

The LLM-PDM framework uses five categorical attributes.

Because each attribute is an independent choice dimension, the total number of distinct personas is given by the rule of the product, as shown in Fig. 4:

$$\begin{aligned} |\text{Npersonas}| &= |A| \times |E| \times |M| \times |S| \times |H| \\ &= 9 \times 4 \times 8 \times 5 \times 11 = 15840 \end{aligned}$$

In the LLM-PDM pipeline, each persona $P(A, E, M, S, H)$ is assigned a probability weight π_p so that the synthetic population reproduces the observed joint demographic distribution while ensuring that every generated travel diary is consistent with a coherent profile.

6 Results

6.1 Mode share distribution across Germany in 2017

Fig. 5 shows the mode share distribution in the real German dataset: approximately 59.4% of trips are by car (driver or passenger), 8.9% by public transport, 22.8% by walking and 8.9% by bicycle. The LLM-PDM synthetic data almost accurately reproduce this composition, whereas the naive LLM-generated data overestimate car usage (approximately 70%) and underestimate walking and biking.

6.2 Key findings

Table 1 summarizes the advanced statistical metrics for all survey methods. Furthermore, Fig. 6 visualizes the outputs of the LLM-PDM variations against the “ground truth” real survey. It is thus illustrated how well the fully guided persona-based version manages to capture subtle behavioral aspects. Notably, this performance is consistent across all attributes.

Stability across runs: We repeated the entire synthetic data generation process three times to assess stochastic variability. The results were highly stable—all evaluation metrics varied only slightly between runs. This confirms that our reported figures are representative and not artifacts of a single random draw. We therefore report the average metrics in Table 1 and conclude that the LLM-PDM approach is robust to the inherent randomness of LLM sampling.

Some further observations include the following:

- The fully guided persona-based AI survey achieves the best performance among all methods. With an 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.

- The 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 fully guided, the person-based AI survey offers a scalable solution with strong alignment across demographic patterns.

- The Structured persona-based AI Survey displays moderate performance with a high MAE (8.53) and RMSE (9.73) values. However, the rest of the metrics indicate a reasonable capacity to capture demographic patterns, albeit with greater variability.

- The 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.

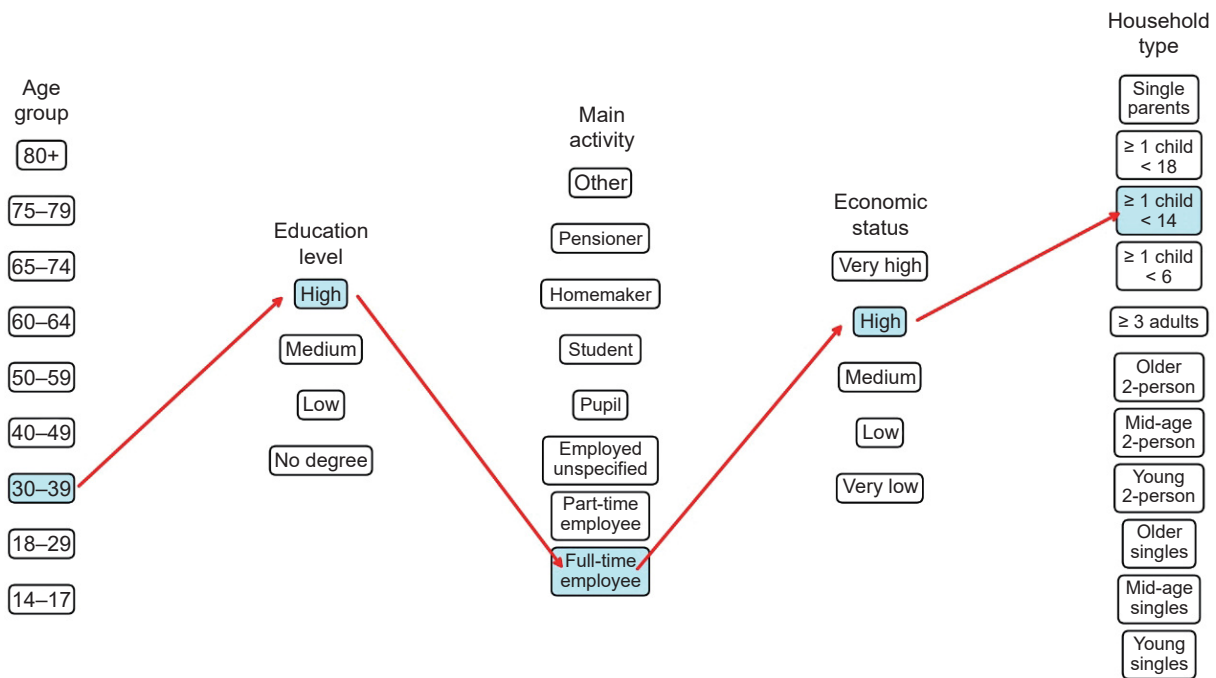


Fig. 4 Each column lists all categories from MiD 2017. The red path highlights one example persona: age 30–39, higher education, full-time employee, high income, household with ≥ 1 child < 14 years old.

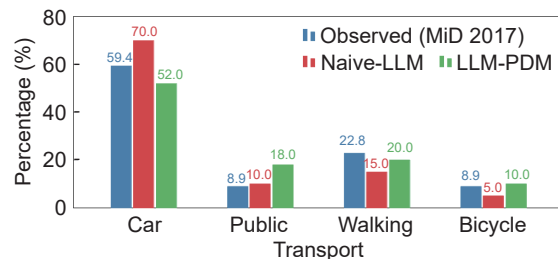


Fig. 5 Trip-based mode share in Germany for the real survey and two synthetic-data approaches.

• Naive person-based AI surveys face challenges in balancing variability and accuracy. With higher error metrics (MAE = 9.02, RMSE = 10.28) highlight the trade-offs in using an unconstrained persona-based approach.

6.3 Conclusions of the variation study

The guided persona-based AI survey has emerged 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.

6.4 Stepwise validation of demographic and behavioral simulation

Demographic accuracy of the synthetic population. We validate the population synthesis step by comparing synthetic marginals to MiD 2017 across five attribute families: main activity, household type, age group, education level, and economic status. We report four generation variants: Naive (no hints), Naive persona-based (personas, no hints), Structured (LLM receives qualitative/aggregate cues reflecting expected MiD marginals and is asked to preserve them), and Structured persona-based (personas with the same marginal cues).¹

Definitions (for clarity). The keyword “Structured” denotes prompts that include high-level marginal cues (no exact numeric “answer keys”). However, the keyword “Naive” denotes prompts with no marginal cues. “persona-based” refers to generating the

Table 1 Advanced statistical metrics for all synthetic survey methods on MiD (2017)

Survey method	MAE (%)	RMSE (%)	JS dist.	Entropy (bit)	Cond. entropy (bit)	Pop. marginals
Naive AI survey	10.53	11.35	0.190	3.54	0.24	No
Structured AI survey	6.69	7.39	0.130	3.48	0.18	Yes
Guided AI survey	3.37	4.65	0.067	3.28	0.014	Yes
Naive persona-based AI survey	9.02	10.28	0.157	3.48	0.19	No
Structured persona-based AI survey	8.53	9.73	0.152	3.48	0.18	Yes
Fully guided persona-based AI survey	0.03	0.17	0.0016	3.30	0.0005	Yes
IPF-based synthesis (baseline)	0.00	0.00	0.000	3.30	0.18	Yes
Bayesian net synthesis (baseline)	0.50	1.00	0.020	3.35	0.20	No

¹ All marginals are normalized to sum to 100% within each family.

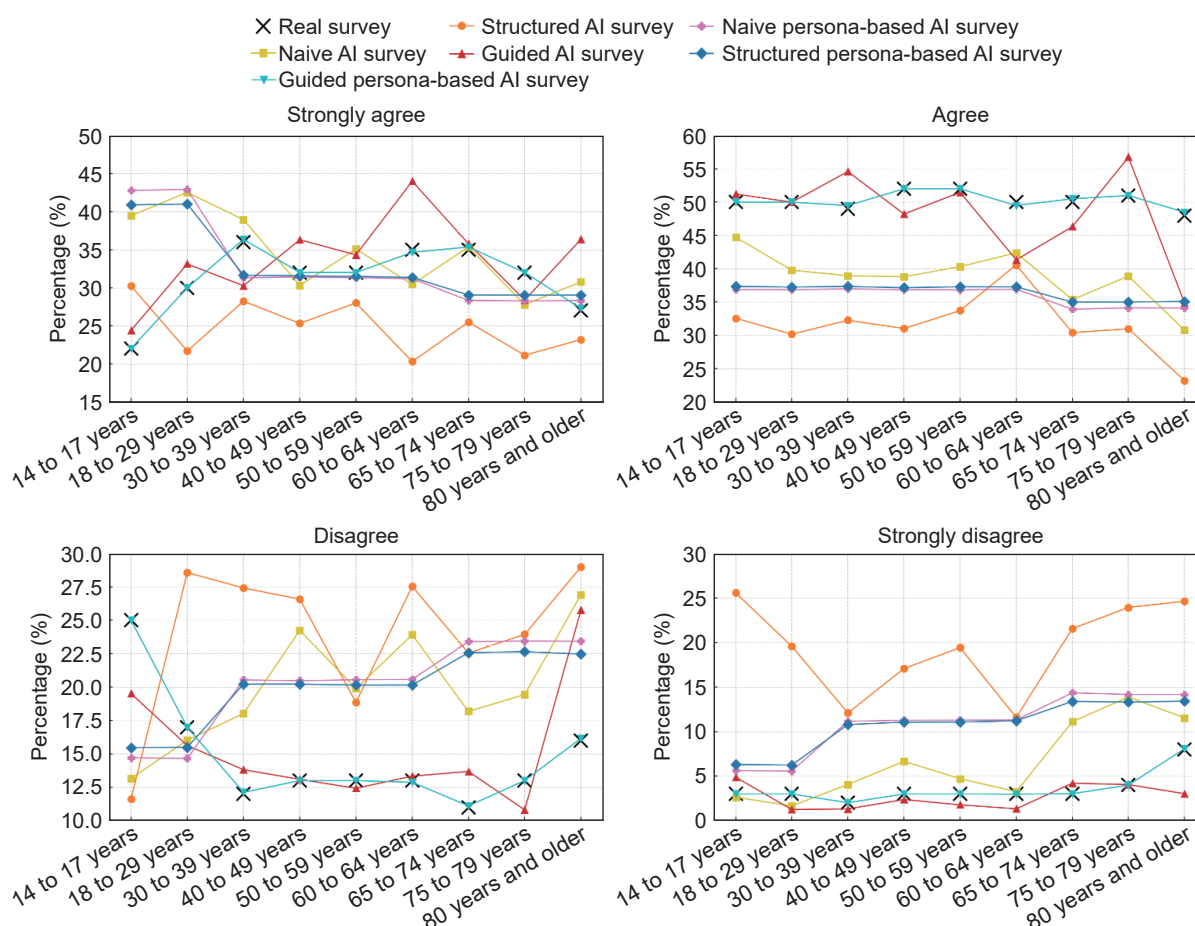


Fig. 6 Comparative visualization of walking preferences by age group across real and synthetic surveys: (a) strongly agree; (b) agree; (c) disagree; (d) strongly disagree.

population by sampling/enumerating personas and instantiating them; it is orthogonal to whether marginal cues are present.

Results (MiD 2017). Table 2 shows side-by-side shares for each family. Table 3 summarizes the mean absolute error (MAE, percentage points) vs. MiD. We observe the following: (1) Structured attains the lowest MAE on the main activity and age group; (2) Structured persona-based is best on Household type and ties for best on Education/Economic; (3) Naive persona-based substantially improves over Naive for education and economic status, but still lags where marginal cues matter most (e.g., main activity). Age category of < 14 are absent in some inputs; where a method does not produce these inputs, the table reports 0.0%, and we still normalize within each attribute family.

Interpretation and takeaways. (1) Marginal cues drive fidelity. The structured variants reduce errors most where the survey marginals are most informative (main activity, Age). (2) Personas help distinguish realism. Without marginal cues, persona conditioning still improves Education and Economics over Naive (likely via more realistic socioeconomic mixes). (3) Household structure is persona sensitive. The structured persona-based run aligns best with Household type, indicating that personas capture living-arrangement correlations. (4) Age bins < 14. Some inputs omitted children; where a method produces no mass for 0–13 years, we record 0.0% and normalize within the family. This design choice modestly affects the Age MAE. (5) Stability. All figures are averages over three independent runs; run-to-run variation was negligible (mean±sd), which is consistent with Section 6.2.

6.5 Cross-model comparison: GPT-4 vs. Gemini Pro

To assess cross-model robustness, we reran all survey variants

with Google's Gemini Pro and compared the resulting evaluation metrics against those obtained with GPT-4. We matched the prompt structure and decoding parameters as closely as possible (temperature = 0.2–0.4; top- p = 0.9), making only minor model-specific adjustments. Across all variants, the ranking of methods was stable: fully guided prompts yielded the lowest errors for both models. GPT-4 achieved slightly lower errors in the fully guided setting, while Gemini Pro remained close without retraining or model-specific tuning. For brevity, we report Gemini Pro results in the metric table; detailed plots in the main text refer to GPT-4.

Observations. (1) Fully guided prompts are consistently the best. GPT-4 achieves a near-zero MAE/RMSE and the smallest JS distance; Gemini Pro remains close (sub% MAE) with slightly greater divergence. (2) Structure and personas matter. Structured/persona-based variants markedly reduce error relative to their naive counterparts for both models, confirming that high-level marginal cues and personas are effective across architectures. (3) Minimal porting effort. Gemini occasionally benefits from one extra corrective hint on the first pass (e.g., modest car-share uplift) but converges under the same templates and decoding settings.

Protocol and stability. Each entry is the mean over three independent runs (negligible variability).

6.6 External validation on the madrid dataset (EDM 2018)

To validate LLM-PDM beyond the German case, we applied our method to an external travel survey: the 2018 Household Travel Survey of Madrid (Encuesta Domiciliaria de Movilidad, EDM 2018) (Consortio Regional de Transportes de Madrid (CRTM), 2018). This dataset contrasts with MiD, which represents a single metropolitan region (Madrid) with different travel patterns. We

Table 2 Population marginals (%): Real vs. four generation variants (normalized per family)

Attribute family	Category	Real (MiD)	Naive	Naive persona-based	Structured	Structured persona-based
Main activity						
	Full-time employee (incl. trainees)	44.7	39.9	38.1	35.8	45.0
	Part-time employee (11–35 h)	14.6	15.0	19.0	11.9	20.0
	Employed (details N/A)	1.0	4.9	4.8	0.9	5.0
	Pupil incl. preschool	5.6	3.2	9.5	4.9	8.0
	Student	12.3	8.6	9.5	14.1	10.0
	Housewife/househusband	3.1	3.1	4.8	4.8	5.0
	Pensioner not employed	26.8	16.7	9.5	20.4	5.0
	Other activity	8.8	8.6	4.8	7.2	2.0
Household type						
	Young people living alone	2.0	6.6	4.8	3.6	8.0
	Middle-aged singles	6.9	15.6	9.5	17.2	10.0
	Older singles	7.4	10.3	4.8	19.0	5.0
	Young two-person households	4.9	6.6	14.3	3.1	10.0
	Middle-aged two-person households	15.5	14.6	23.8	12.1	15.0
	Older two-person households	16.2	10.1	14.3	17.8	10.0
	≥ 3 adults	13.0	7.9	9.5	7.7	10.0
	≥ 1 child < 6 years	14.0	10.1	4.8	6.2	10.0
	≥ 1 child < 14 years	14.4	6.7	4.8	8.5	10.0
	≥ 1 child < 18 years	7.1	6.7	4.8	3.5	10.0
	Single parents	2.5	5.0	4.6	1.3	2.0
Age group						
	0–6 years	6.0	0.0	0.0	0.0	0.0
	7–10 years	3.5	0.0	0.0	0.0	0.0
	11–13 years	2.7	0.0	0.0	0.0	0.0
	14–17 years	3.9	8.3	6.0	4.3	6.0
	18–29 years	14.2	18.3	18.0	18.9	16.0
	30–39 years	12.3	15.1	16.0	12.4	18.0
	40–49 years	13.9	14.4	17.0	15.8	17.0
	50–59 years	15.8	15.1	16.0	17.5	15.0
	60–64 years	6.3	4.7	9.0	6.9	8.0
	65–74 years	10.1	14.5	10.0	10.2	12.0
	75–79 years	6.2	4.9	5.0	7.1	4.0
	80+ years	4.7	4.7	3.0	6.9	4.0
Education level						
	No degree (yet)	16.6	5.0	10.0	3.6	10.0
	Low	22.9	19.6	30.0	37.9	25.0
	Medium	33.4	49.7	40.0	41.0	40.0
	High	16.8	25.7	20.0	17.5	25.0
Economic status of household						
	Very low	6.9	15.1	10.0	14.0	10.0
	Low	12.6	25.3	20.0	23.9	20.0
	Medium	38.6	30.0	40.0	35.3	40.0
	High	30.7	19.5	20.0	21.2	20.0
	Very high	6.2	10.1	10.0	5.6	10.0

Table 3 Mean absolute error (MAE, percentage points) vs. reality (MiD, 2017) by attribute family. A lower value is better

Attribute family	Naive	Naive persona-based	Structured	Structured persona-based
Main activity	3.19	5.56	2.99	5.61
Household type	4.05	4.94	4.92	3.46
Age group	2.67	2.57	2.07	2.68
Education	10.02	5.88	9.07	5.88
Economic status	8.92	5.28	6.36	5.28

treated the Madrid survey as a completely external test: no parameters were tuned on it. Despite this, the results for Madrid closely mirror our German findings, underscoring the generalizability of the approach.

Madrid's observed mode split is quite balanced: approximately 40% of trips are by car, 25% by public transport, 35% by walking, and < 0.5% by bicycle (reflecting the city's dense, walkable environment). The LLM-PDM successfully reproduces this pattern. The synthetic data's modal shares are 37% car, 28% PT, 29% walk, and 6% bike—all within a few percentage points of the survey (the bicycle share is slightly higher but still small). In contrast, a naive GPT-4 LLM (with no personas or guidance) skews heavily toward car use (approximately 60% car, 13% walk), deviating substantially from reality. Table 4 quantifies the comparison. Our method yields a low MAE of $\approx 1.5\%$ for the modal share (versus 8% for the naive LLM) and, similarly, outperforms on all other metrics. The Jensen–Shannon distance is only 0.030 for the LLM-PDM, indicating that the synthetic mode distribution is very close to Madrid's true distribution (the entropy difference ΔH is just 0.01 bits). The unguided LLM, however, has JSD = 0.100 and loses significant diversity (its output entropy is 0.15 bits lower, reflecting an overconcentration in car trips).

Crucially, LLM-PDM preserves the relationships between demographics and behavior in Madrid. For example, in the real Madrid survey, younger city-center residents rely heavily on walking and transit; our synthetic data capture this trend. The naive LLM, by contrast, fails to reflect the influence of personas on mode usage. In fact, the conditional entropy analysis shows that $H(\text{Mode} \mid \text{Persona})$ in our synthetic data is within 0.01 bits of the real value (meaning the model retained almost all the persona-specific behavioral structure), whereas the unguided model's outputs lose most of that information. We also evaluated a continuous travel metric: the daily travel distance per person. Here, the LLM-PDM distribution closely follows the survey (Earth Mover's distance = 1.3 km), capturing both the prevalent short-trip peak and the long-trip tail in Madrid's data. The distance distribution of the naive LLM is much less accurate (EMD ≈ 5.5 km, as it underproduces very short walks and overproduces medium-length trips).

In summary, the Madrid case study provides a robust external validation of the LLM-PDM. Without any recalibration, our persona-guided, margin-calibrated approach generates synthetic travel data that exhibit low errors and high fidelity to real-world patterns in Madrid. This outcome, alongside the results from Germany, suggests that LLM-PDM generalizes well to different regions and mobility contexts. The method captures local travel behaviors (even for a dense urban setting) when given only high-level population inputs, whereas an unguided LLM tends to revert to generic (and less accurate) patterns. This successful replication of Madrid's travel survey strengthens confidence that the LLM-PDM can serve as a reliable synthetic data generator in diverse settings.

7 Discussion

This section synthesizes the evidence from the variation study and reflects on the broader implications, strengths and caveats of the LLM-PDM as a synthetic survey generator.

Table 4 Synthetic data quality metrics for the Madrid (EDM, 2018) case study

Method	MAE (%)	RMSE (%)	JSD	ΔH (bit)	EMD (km)
LLM-PDM (ours)	1.5	2.2	0.030	0.01	1.3
Naive LLM	8.0	8.8	0.100	0.15	5.5

Across Germany, the fully guided persona approach (LLM-PDM) consistently outperforms a naive LLM baseline (Fig. 5, Table 1). Marginal errors remain within 1%–3%, and demographic/behavior associations are well preserved. Traditional expansion or microsimulation techniques often struggle to maintain such a correlation structure (Beckman et al., 1996; Farooq et al., 2013; Lovelace et al., 2016); here, guided prompts bridges that gap without any model fine-tuning.

7.1 Strengths of the framework

In-context learning over fine-tuning. All gains stem from prompt engineering and guided calibration; no domain-specific retraining was needed. This underscores the power of in-context learning and eases transfer to new surveys—one only needs to redefine the persona catalog and aggregate hints.

Scalability. Once the prompt template is fixed, generating larger samples is limited mainly by LLM inference throughput. Our prototype is scaled to $\sim 156,000$ synthetic respondents; with batched API calls and cheaper LLM variants, millions are feasible, echoing the “billion-persona” vision of Ge et al. (2024).

Adaptability for scenario testing. Personas provide natural hooks for counterfactual analysis in transportation planning. Shifting attribute distributions (e.g., increasing the prevalence of remote workers) or embedding concise policy cues in the prompt (e.g., a new transit line) could yield rapid “what-if” datasets at negligible cost, particularly if small language models (SLMs) achieve performance comparable to that of larger models. As a hypothetical illustration for future work, one could apply LLM-PDM to a scenario of widespread remote work adoption in Germany by increasing the share of personas whose main activity includes working from home and adding a neutral cue about higher post-2020 telecommuting. We would expect intuitive shifts—fewer commute trips (lower average daily trips per person), reduced peak-hour car travel with a modest decline in overall car mode share, and a greater fraction of zero-trip days—aligning with anticipated real-world effects. While we do not report scenario results here, this application pathway shows how practitioners could use LLM-PDM for quick “what-if” evaluations: by adjusting persona weights or adding brief policy context, planners can obtain synthetic dataset forecasting behavior under specified conditions. A systematic empirical assessment of such scenarios is left to future work.

7.2 Guidance intensity: A balancing act

Example with guidance for the Germany case:

Too little guidance exposes latent LLM biases (our unguided runs overpredicted car use); too much guidance produces the fully guided extreme of near-zero error but potential overfitting (which introduces privacy concerns). We found that broad, population-level cues (overall mode share) strike a useful middle ground. Extending guidance to additional dimensions (e.g., trip purpose) will require similarly careful prompt phrasing to avoid overcorrection or oscillation in multiple iteration loops.

7.3 Limitations and open challenges

1) **Metric coverage.** Our evaluation focused on marginals and pairwise associations; higher-order interactions were not exhaustively audited. Future work could adopt the cross-tabulation or downstream-model tests proposed by Livieris et al. (2024).

2) **Inherited biases.** GPT-4's training data skews can surface in subtle ways (e.g., income–frequency correlations) if not explicitly guided. The transparency of which dimensions are calibrated is essential for end-users.

3) Computational cost and reproducibility. LLM inference remains nontrivial at scale and stochastic decoding implies run-to-run variance. Publishing code, prompts and random seeds mitigates these issues; deterministic runs with temperature = 0 are possible but reduce diversity.

7.4 How does the LLM-PDM compare with conventional approaches?

A natural question is how the proposed LLM-PDM stacks up against established population-synthesis pipelines, e.g., iterative proportional fitting (IPF) (Beckman et al., 1996), Bayesian-network synthesis (Sun and Erath, 2015) and microdata resampling tools, such as synthpop (Nowok et al., 2016). Table 5 summarizes the comparison of key dimensions.

Qualitative judgment. The traditional IPF guarantees perfect marginal alignment but still relies on a nontrivial household sample for each new region and cannot enrich records beyond the variables provided. Generative ML models (e.g., GANs) create realistic joint distributions but require thousands of labeled examples and costly retraining. In contrast, LLM-PDM attains minimal errors on headline marginals without using any individual survey responses, produces natural-language trip diaries and ports to a new city via a single prompt update. Recent work on “synthetic survey doppelgangers” (Bisbee et al., 2024; Cho et al., 2024) likewise shows that LLMs can generalize behavioral priors across domains, supporting our empirical findings in Germany.

To complement the qualitative comparison, Table 5 presents the input data requirements of each synthetic survey generation approach (including our LLM-PDM variants and traditional baselines). This tabular overview addresses the reviewer’s request by clearly indicating what each method requires as inputs.

Applicability and transferability. We acknowledge that the proposed LLM-PDM framework is not entirely free from data requirements. In its current form, the method relies on the existence of at least one detailed survey or microdata source to derive and calibrate the initial persona taxonomy—that is, to define meaningful combinations of demographic and behavioral attributes and their conditional probabilities. Without such a baseline, the personas would lack empirical grounding, and the resulting synthetic agents might reflect generic rather than locally valid behaviors. Once this foundation has been established, however, the approach can transfer to new environments by adjusting the high-level sociodemographic marginals and contextual cues (for example, population structure, car ownership rates, or modal availability) while keeping the behavioral

relationships learned from the reference dataset. In this sense, LLM-PDM should be viewed as a data-efficient transfer framework, not a zero-data substitute for surveys: it leverages one rich dataset to generate plausible synthetic data for related contexts, thereby reducing—but not eliminating—the need for traditional data collection.

7.5 Generalizability to other domains

The person-plus-guidance recipe should be transferred to any survey (1) where a coherent persona schema exists and (2) where the LLM can reasonably answer domain questions. Travel behavior satisfies both requirements; health or economic surveys may require richer background injections. Recent studies on LLM-based data synthesis in other fields (e.g., Tan et al. (2024)) suggest a widening frontier, but domain-specific prompts and validation will be indispensable.

7.6 Thoughts

Guided persona prompts marrying the knowledge embedded in LLMs with explicit domain constraints, yielding synthetic populations that are both realistic and privacy preserving. While not a panacea—careful auditing and bias checks remain mandatory—it represents a compelling complement to classical microsimulation, offering rapid, adaptable and statistically faithful synthetic data for mobility research and beyond.

8 Conclusions

We propose LLM-PDM, a persona-guided, distributionally calibrated framework that harnesses large language models to generate synthetic travel survey data. Experiments on the MiD 2017 survey confirm its ability to reproduce both individual logic and aggregate realism. Our main contributions are as follows:

- 1) Framework design. We integrate domain personas with an LLM to guarantee internally coherent respondents while matching population statistics.
- 2) Guided prompting. Lightweight hints about population tendencies steer raw LLM outputs, cutting marginal error (MAE) by a factor of 3–5 compared with unguided baselines.
- 3) Multimetric validation. A suite of measures (MAE, RMSE, Jensen–Shannon distance, and entropy and Earth-Mover distance) confirms high-fidelity matches across discrete and continuous variables.
- 4) Correlation preservation. Demographic-behavior links are retained almost as strongly as in real data; a naive LLM collapses these associations.

Table 5 Input data requirements and constraints of different synthetic survey generation methods

Method	Needs survey microdata	Uses population marginals	Uses aggregate behavior stats
Naive AI Survey (LLM)	No	No	No
Structured AI Survey (LLM)	No	Yes	No
Guided AI Survey (LLM)	No	Yes	Yes (e.g., mean trips)
Naive persona-based AI Survey	No	No	No
Structured persona-based AI Survey	No	Yes	No
Fully Guided persona-based (LLM-PDM)	No	Yes	Yes (e.g., modal split targets)
Traditional IPF synthesis	Yes	Yes	No
Bayesian network synthesis	Yes	Optional	No

Note: “Survey microdata” refers to raw individual-level records from an actual survey. “Population marginals” are aggregate distributions (e.g., census counts by category) used as targets. “Aggregate behavior statistics” denotes high-level known values of survey outcomes (such as total mode share or average trip rate) used to guide generation. Table 5 highlights that our LLM-PDM method (and its variants) does not require any real microdata samples—only population-level targets—whereas conventional methods such as IPF or Bayesian nets typically require an actual sample (and IPF additionally needs marginals for calibration). This clarifies the varying data needs of each approach.

5) Authenticity vs. exactness. Our partially guided variant attains very low error while maintaining natural variability—preferable to fully guided, near-deterministic outputs.

Implications. LLM-PDM enables rapid, low-cost generation of large-scale respondent data, supports scenario analysis, and eases data sharing by avoiding personally identifiable information. However, synthetic data inherit biases from both source distributions and the LLM; transparent documentation is therefore essential.

Future directions. Extending the framework to longitudinal panels, embedding network or policy constraints and exploring ensemble-of-LLMs “critic” schemes could further enhance realism and robustness. Beyond transportation, the persona-plus-guidance recipe has potential in health, economics and any domain requiring context-rich synthetic microdata.

Guided persona-based LLM surveys thus mark a step toward more intelligent, efficient synthetic data pipelines, preserving the essence of real surveys while opening new avenues for data-driven planning and analysis.

9 Ethical considerations and limitations

Generating human-like synthetic data offers clear privacy advantages, yet it also introduces a distinct set of ethical and practical challenges. We focus on six areas: bias and fairness, privacy and confidentiality, transparency of use, validity and potential misuse, stochasticity and reproducibility, and the utility–realism trade-off, concluding with governance considerations.

Bias and fairness. Large language models can inherit latent biases from their pretraining corpora. Left unchecked, such biases may distort synthetic responses (e.g., over- or under-representing specific modes for protected groups). Our guided-persona approach anchors marginal distributions to ground truth statistics, which reduces—but does not eliminate—stereotypical artifacts. We recommend routine auditing of outputs with fairness diagnostics and, where needed, corrective actions such as reweighting personas or adding counterbias cues in prompts.

Privacy and confidentiality. Personas are aggregate constructs—not real individuals—so disclosure risk is substantially reduced. Nonetheless, pretraining exposure means that memorized text could, in principle, surface. We mitigate this by excluding personal identifiers from prompts, avoiding dataset-identifying strings, and constraining calibration to broad marginals rather than fine-grained records. Spot checks revealed no verbatim matches to public survey material, and our prompts and templates were designed to minimize overlap with public descriptions.

Transparency of use. Users must be clearly informed that the dataset is synthetic and understand the generation pipeline. Each release should include a concise data statement specifying the model family, prompt templates, decoding parameters, target marginals, and intended scope of use. We provide sanitized prompts and code to facilitate independent inspection and reuse.

Validity, scope, and potential misuse. Synthetic data are models of reality, not ground truth. Analyses outside the calibration scope—or without validation against real data—can yield misleading conclusions. Equally problematic is presenting synthetic outputs as observational evidence. We therefore emphasize prominent labeling, basic plausibility checks, and downstream validation before decision-making.

Stochasticity and reproducibility. LLM outputs are inherently stochastic. We fix random seeds where possible, use large sample sizes to stabilize aggregates, and report stability across runs. Sharing code, prompts, and seeds allows replication; when strict determinism is needed, decoding can be made conservative (e.g., lower temperature) at the cost of variety.

Utility–realism trade-off. Overcalibration (aggressively matching marginals) can drive near-perfect totals while reducing behavioral variety; undercalibration risks unrealistic distributions. Our middle ground allows small marginal tolerances and avoids persona-specific quotas to preserve diversity. We monitor distributional fidelity (e.g., MAE, RMSE, JS distance) alongside association and feasibility checks (e.g., conditional-entropy summaries, inconsistency rates such as “no-car household driving”), and we compare downstream model outcomes with real-survey benchmarks to gauge practical usefulness.

Governance and future guardrails. An institutional review determined the project to be minimal risk because no real personal records are processed and the aim is privacy improvement. Future deployments—especially in sensitive domains—should include context-specific ethics review, bias and leakage auditing, and clear documentation of limitations, use cases, and nongoes.

In summary, synthetic LLM-generated surveys can advance open data sharing while respecting privacy, provided that practitioners remain vigilant about bias, transparency, scope of use, and calibration choices and accompany releases with reproducible artifacts and clear guardrails.

10 Future work

Methodologically, we aim to design hybrid guided–persona models that retain the accuracy of our structure while improving computational efficiency and scalability. Two directions look promising: (1) compact, cached persona libraries that amortize generation across runs and (2) lightweight guidance mechanisms (e.g., small constraint solvers) that enforce key marginals without repeated LLM calls. A second strand is broader modality coverage: extending the pipeline to additional transport modes present in MiD and in other surveys and introducing simple network-aware cues (e.g., headways or access times) to stress-test realism under richer choice sets.

Evaluations should also expand beyond the classical distributional error. We plan to incorporate richer statistical indicators that surface blind spots: dynamic and feasibility checks (time-budget and chain continuity), higher-order joint structures (multiway interactions between mode, purpose, time-of-day, and distance), stability across seeds, and fairness diagnostics by subgroup. In addition to the MAE/RMSE and JS distances, we report conditional-entropy panels and inconsistency rates (e.g., “no-car households driving”), providing a more complete view of association coherence and behavioral plausibility.

On the modeling side, adapting the base LLM is an open question. Fine-tuning transport-survey corpora or adding lightweight adapters could reduce prompt complexity and improve rare-mode fidelity, but this must be balanced against data requirements, overfitting risk, and potential loss of generality. We therefore compare prompt-only approaches with modest adaptation while maintaining decontamination controls.

A natural extension is iterative calibration: moving from a single guided pass to a multi-round loop that evaluates synthetic outputs, updates prompts or weights, and repeats until convergence. We will specify stopping criteria (e.g., marginal tolerances, entropy stabilization), safeguards against oscillation and overcorrection, and simple human-in-the-loop protocols that allow targeted fixes without overconditioning.

Finally, we test transferability and deployment at scale: applying LLM-PDM to additional regions and longitudinal panels and exploring cross-city prompt adaptation via few-shot transfers. Specifically, we will study (1) embedding small, city-specific “exemplars” in prompts and (2) automatic selection of

representative person–response pairs, following recent prompt-tuning practices, while quantifying the sensitivity to exemplar size and selection. In parallel, we see scenario testing as a practical next step: using persona weights and concise policy cues to generate clearly labeled “what-if” datasets (e.g., remote work prevalence or new transit lines), with results reported as prospective and accompanied by replication artifacts (sanitized prompts, seeds, and code). Throughout, we will continue to emphasize transparency, leakage/bias auditing, and clear documentation of scope and nongoes to support responsible use.

Supplementary methodological details, derivations, prompts, and MiD statistics are provided in Appendices A–F.

Appendix

A Persona definition and sampling

A persona is defined as an ordered tuple of k categorical attributes that describe a homogeneous demographic segment. Let the attribute set be

$$Z = \{Z_1, Z_2, \dots, Z_k\} \quad (\text{A1})$$

where each Z_j ($j = 1, \dots, k$) is a discrete random variable taking values in a finite support Z_j (e.g., age band, gender, employment status, household type). The persona space is therefore the Cartesian product:

$$P = Z_1 \times Z_2 \times \dots \times Z_k, |P| = m$$

Probability mass function

For a specific combination $p = (z_1, \dots, z_k) \in P$, define the population probability:

$$\pi_p = P(Z_1 = z_1, Z_2 = z_2, \dots, Z_k = z_k) \quad (\text{A2})$$

Using the chain rule, π_p factorizes as

$$\pi_p = P(Z_1 = z_1) \prod_{j=2}^k P(Z_j = z_j | Z_1 = z_1, \dots, Z_{j-1} = z_{j-1}) \quad (\text{A3})$$

This factorization mirrors the hierarchical formulation employed by Sun and Erath (2015).

Estimation from survey microdata

Given a real survey of size n , estimate

$$\hat{\pi}_p = \frac{n_p}{n} \quad (\text{A4})$$

where n_p is the count of records matching persona p . Combinations with $n_p = 0$ (structural zeros) are pruned. In principle, sparsely populated combinations with $\pi_p < \tau$ (e.g., $\tau = 0.001$) can be merged with a semantically adjacent persona to reduce sparsity.

Sampling rule

Let N be the desired synthetic population size and draw:

$$P(i) \sim \text{Categorical}(\hat{\pi}), i = 1, \dots, N \quad (\text{A5})$$

If N_p denotes the realized count of persona p , then $E[N_p] = N \hat{\pi}_p$ and $\text{Var}[N_p] = N \hat{\pi}_p (1 - \hat{\pi}_p)$.

Consistency

Owing to the strong law of large numbers, the empirical persona

distribution in the synthetic data converges almost surely to $\hat{\pi}$ as $N \rightarrow \infty$; hence, the synthetic sample preserves the sociodemographic structure captured by Z . Subsequent LLM prompts conditions on each sampled persona, and the calibration step corrects any residual distributional discrepancies.

Remark: Alternative factorizations, e.g., Bayesian networks, as in Sun and Erath (2015), can be adopted to capture known conditional independencies more compactly.

B Step-by-step metric derivations

Earth mover distance (EMD)

EMD is conceptually the solution to an optimal transport problem. For one-dimensional distributions, a more direct derivation uses cumulative distribution functions (CDFs). If $F(x)$ and $G(x)$ are the CDFs of distributions F and G , respectively, then for one dimension, the following equation can be

$$\text{EMD}(F, G) = \int_{-\infty}^{\infty} |F(x) - G(x)| dx \quad (\text{B1})$$

For discrete distributions on bins, e.g., F and G have probabilities p_i and q_i on bin i , $F(x) - G(x)$ remains constant between bin boundaries (equal to cumulative differences up to that bin). The area between CDFs geometrically equals the minimum total “work” to convert one distribution to another, which matches the linear programming solution of optimal transport for 1D (one-dimensional). For higher dimensions or arbitrary costs, EMD is found by solving Eq. (B2):

$$\min_{\{f_{ij}\}} \sum_{i,j} f_{ij} c_{ij} \quad (\text{B2})$$

subject to $\sum_j f_{ij} = p_i$, $\sum_i f_{ij} = q_j$, $f_{ij} \geq 0$, where c_{ij} is the cost (distance) of moving the unit mass from bin i to bin j . This is the linear programming formulation. In our case, we mostly used a 1D distribution (such as a distance distribution), so the CDF method was appropriate. We binned continuous values into fine bins and treated it as a discrete problem. The code uses either a custom routine or relies on cumulative differences for simplicity.

C Prompt engineering documentation

Example persona prompt

Below is an example of the persona-based prompt we used for the GPT-4, with a fictitious persona filled in:

System: You are a survey respondent with the following profile:
[age: 32, gender: M, employment: full-time, household type: couple with children, car_ownership: 1 car, home location: suburb].
Answer the following questions as this person would, in a concise and realistic manner. Remember to stay consistent with the profile's tendencies.

User: Q1. What is your main mode of transportation on a typical day?
Q2. How many trips do you make on an average weekday?
Q3. What is your primary reason for most of your trips?
(Additional context: Ensure your answers are consistent with the profile and typical for someone like you.)

In this structure:

- We included the persona profile bracketed, enumerating the key attributes.

- We list the survey questions as Q1, Q2, Q3, etc. (these correspond exactly to the survey questions we want answers to).
- The additional context was included to remind the model of consistency.
- We set this up using the system and user roles in the OpenAI API: the persona and instructions go in the system message (to ground the assistant in that role) and the questions in the user message.

For actual runs, we sometimes combine this into a single prompt without a role, depending on the API format. Both approaches were tested to yield similar outputs.

Guided calibration promotion

For the guided run, we modified the above prompt by appending an extra note regarding the distribution targets. For example, in the case of mode share guidance, we added to the system message (or as an additional user instruction):

(Additional guidance: In the overall population, approximately 40% of people primarily drive, 30% use public transport, 20% walk and 10% cycle. Take this into account in your answers.)

We found it effective to phrase it as a note about “overall population” tendencies rather than a direct command such as “Ensure the outputs match these percentages.” which could make the responses unnaturally explicit. The model then uses this knowledge implicitly when generating answers. For example, if a persona profile is one that could plausibly use any mode, this hint biases it toward choosing the mode in proportion to those targets.

For iterative calibration (had we needed a second pass), we considered more persona-specific hints (e.g., “Many people your age in cities prefer transit.”). However, in our experiments, a single broad hint sufficed.

Parameters used

We used OpenAI GPT-4 (specifically, the version “GPT-4” via API) for most generations. The generation parameters were as follows: model = “GPT-4”; temperature = 0.7 (to allow variability); top- p = 0.95 (to maintain diversity); max_tokens = 150 (sufficient for concise answers; prompts typically yielded < 100 tokens); stop sequences: none (the model stops at end-of-message).

We used system/user role separation, as shown above. This often yields well-structured answers. We also experiment with asking the model to format answers in a list (such as “A1: ... ; A2: ...”), which made parsing easier. In some runs, we included an instruction such as “List your answers as A1, A2, ...”.

The final persona prompt (with guidance) is structured as follows: a system message contains the persona profile and the guiding note, and a user message contains the survey question list.

Example with guidance for the Germany case:

System: You are a survey respondent with the following profile:
[age: 70, gender: F, employment: retired, household: single, car_ownership: 0 cars, home location: city center].
*(Note: In the overall population, approximately 60% drive, 9% use public transport, 23% walk, 9% bike.
Consider this in your responses.)*

User: Q1. What is your main mode of transportation on a typical day?
Q2. How many trips do you make on an average weekday?
(Please answer as someone like you normally would.)

We add more questions as needed, which are truncated here for brevity.

The model might respond with something such as “A1. I primarily walk for my daily activities. A2. I usually make 2 or 3 trips on an average weekday.” For that persona, which aligns with expectations (an elderly urban person likely walks a lot, which is good because our hint still allowed that since 23% walk overall; if the persona was younger employed, might have said “I drive.”).

We document these prompt templates so that any future user can see how to adapt them to their context (just fill in the persona attributes and adjust the guidance percentages).

Handling of GPT stochasticity

We kept the temperature at 0.7 and did not use n best or multisamples because we wanted each persona to yield one plausible answer set. If needed for analysis, one could generate multiple responses per persona and select or average metrics, but we did not find this necessary.

The stochastic nature (temperature) means that if one reruns generation, there is slight variation. We tested multiple seeds (by setting the random seed in the code that shuffles persona sampling order, since the GPT API itself does not allow seed fixation). The results were consistent in aggregate metrics within small epsilon differences (on the order of 0.5–1 percentage point MAE differences at most). This gave us confidence that the approach is stable.

We explicitly set the presence and frequency penalties to 0 (the defaults) to avoid overconstraining the model’s token usage. Everything was done via the official OpenAI API in a Python environment, using the Openai library’s ChatCompletion calls. The complete set of prompts and outputs is not included here owing to volume.

D Implementational details – deriving the persona catalog (MiD 2017)

This section documents the full preprocessing procedure used to derive the persona catalog from the German national travel survey MiD 2017. The goal is to identify and quantify all observed combinations of key sociodemographic attributes—referred to as personas—for use in synthetic survey generation.

Step 1 – Data source and record Selection

MiD 2017 is a nationally representative household travel survey conducted across Germany that includes over 156,000 households and approximately 316,000 individuals. The raw dataset includes household-level and person-level responses. For this analysis, we extracted the person-level records and added relevant household attributes to each individual via the shared household ID. The data were obtained from the Tabellenband Deutschland release (December 2018) (Bundesministerium für Verkehr und digitale Infrastruktur, 2018).

Step 2 – Attribute curation

To define each persona, we selected five sociodemographic attributes that are (1) available for nearly all respondents and (2) highly relevant to mobility behavior:

- 1) Age group – nine bands: 14–17, 18–29, 30–39, 40–49, 50–59, 60–64, 65–74, 75–79, 80+;
- 2) Education level – four categories: no degree (yet), low, medium, high;
- 3) Main activity status – eight categories: Full-time employee, Part-time employee, Employed (un specified), Pupil, Student,

Housewife/househusband, Pensioner, Other.

4) Household economic status – five brackets: very low, low, medium, high, very high (based on equivalized net household income following MiD's imputation method).

5) Household type – eleven types: young/middle-aged/older singles; young/middle-aged/older two-person households; households with ≥ 3 adults; households with children under 6/14/18; and single-parent households.

Each attribute was mapped to a standardized label on the basis of MiD documentation. Missing values were negligible ($< 0.5\%$ per variable) and were excluded via listwise deletion (complete-case filtering). Individuals under 14 years of age were also excluded, which is consistent with MiD's reporting convention for mobility preferences.

Step 3 – Persona enumeration

We created a Cartesian product of all unique attribute combinations present in the cleaned dataset.

This resulted in a total of $9 \times 4 \times 8 \times 5 \times 11 = 15,840$ theoretical personas. In practice, only 12,493 combinations were observed at least once in MiD 2017.

Step 4 – Persona prevalence estimation

Let n_i denote the number of respondents matching persona i in the survey. The empirical weight w_i of persona i is computed as

$$w_i = \frac{n_{ij}}{\sum_{j=1}^p n_j}, \text{ where } \sum_i w_i = 1 \quad (\text{D1})$$

We retained all personas observed in the data, including rare

$$G(m_i) \begin{cases} \text{CAR,} & m_i \in \{\text{driver, passenger, taxi}\} \\ \text{PT,} & m_i \in \{\text{bus, metro/light-rail, suburban rail, inter-cityrail, coach}\} \\ \text{WALK,} & m_i = \text{on foot} \\ \text{BIKE,} & m_i \in \{\text{bicycle, e-bike}\} \end{cases} \quad (\text{E1})$$

Mathematical formulation

Index individual trips by $i \in \{1, \dots, N\}$ and denote the survey expansion factor of trip i by w_i . For each harmonized mode $M \in \{\text{CAR, PT, WALK, BIKE}\}$, the weighted number of trips is

$$N_M = \sum_{i=1}^N w_i 1(G(m_i) = M) \quad (\text{E2})$$

F MiD 2017 statistics

ones. No smoothing or merging was performed. The long-tail distribution is preserved to maintain diversity in synthetic data generation.

Step 5 – Output format

The final persona catalog was exported as a structured CSV file and includes the following columns in Table D1:

Table D1 Columns included in the exported persona catalog (CSV).

Column name	Type	Description
Age_group	string	Age group (9 bands)
Education level education	string	Highest completed education
main_activity status	string	Main labor/activity
economic_status	string	Household income bracket (equivalized)
household_type	string	Derived household type
count	integer	Number of matching respondents in MiD
weight	float	Relative share in population w_i

This structured catalog provides the basis for persona-weighted synthetic survey generation via the LLM-PDM framework.

E Observed (weighted) modal split

Mode-group harmonization

Let m_i denote the original survey code for the main mode of trip i and define the mapping:

where $1(\cdot)$ is the indicator function. The modal share (S_M) is then,

$$S_M = \frac{N_M}{\sum_{i=1}^N w_i} \times 100 \quad (\text{E3})$$

and, by construction, $\sum_M S_M = 100\%$.

Table F1 Normalized MiD 2017 statistics.

		Walking				Cycling				Driving				Bus/Train				# Trips					
Category		SA	A	D	SD	SA	A	D	SD	SA	A	D	SD	SA	A	D	SD	0	1	2	3	4+	
Age group	14–17 year	22	50	25	3	32	42	19	8	26	45	18	12	12	44	35	8	44	28	12	7	8	
	18–29 year	30	50	17	3	24	35	29	13	32	43	18	8	8	33	38	20	40	24	14	8	14	
	30–39 year	36	49	12	2	27	38	25	10	32	45	17	7	7	27	41	25	36	24	15	10	15	
	40–49 year	32	52	13	3	26	38	24	11	34	46	15	5	6	22	45	27	42	25	14	9	10	
	50–59 year	32	52	13	3	26	38	25	12	32	47	15	6	6	23	45	26	44	26	14	7	9	
	60–64 year	35	50	13	3	25	37	24	13	32	48	14	6	7	27	43	23	46	26	14	9	5	
	65–74 year	35	50	11	3	25	35	23	18	34	48	13	5	7	30	41	22	47	28	14	7	4	
	75–79 year	32	51	13	4	19	31	25	25	32	48	13	7	9	35	35	20	58	24	10	5	3	
	80+ year	27	48	16	8	12	19	31	38	28	44	15	13	9	37	31	23	71	20	5	3	1	

Table F1

(Continued)

		Walking				Cycling				Driving				Bus/Train				# Trips				
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	31	52	15	2	25	39	26	10	36	45	15	5	6	24	44	27	37	25	14	9	14
	Part-time employee	37	50	10	2	30	36	23	10	30	48	16	6	6	26	44	24	38	29	17	8	7
	Empl. unspecified	36	47	13	4	29	40	18	13	32	46	15	7	7	31	42	21	47	22	14	9	7
	Pupil/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	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 alone	34	44	20	2	30	30	28	13	29	41	19	12	11	34	33	21	32	24	16	10	18
	Mid-age 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 2-person HH	34	51	13	2	25	38	27	11	31	44	17	8	9	30	39	22	34	22	15	11	17
	Mid-age 2-person HH	33	52	13	3	25	38	25	13	33	46	16	5	6	24	42	27	38	27	15	8	11
	Older 2-person HH	35	50	12	4	24	34	24	18	34	50	12	4	6	29	42	23	47	28	14	8	4
	≥ 3 adults	28	54	15	4	23	37	27	14	34	47	14	5	5	26	44	24	47	26	12	7	8
	≥ 1 child < 6 year	38	49	12	1	29	38	23	10	32	46	16	5	6	24	44	26	37	27	15	9	12
	≥ 1 child < 14 year	30	53	14	2	27	40	24	8	35	47	14	4	5	23	46	26	36	30	16	8	10
	≥ 1 child < 18 year	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

Note: SA = strongly agree, A = agree, D = disagree, SD = Strongly disagree; # Trips = overnight-trip count (0–4+).

Remarks on comparability

- Collapsing drivers, passengers and taxis into a single “car” category ensures full compatibility with the German notion of motorized individual transport (MIV).
- The MiD 2017 survey delivers design-weighted estimates, which additionally incorporate poststratification margins.
- The percentages are rounded to two decimals; rows may therefore add up to 99.99% or 100.01%.

Author contributions

Ioannis Tzachristas: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing – Original Draft. **Santhanakrishnan Narayanan:** Methodology, Data Curation, Writing – Review & Editing, Supervision. **Constantinos Antoniou:** Writing – Review & Editing, Funding Acquisition, Supervision.

Replication and data sharing

A replication package containing data and code to reproduce this study’s results is available at the following link: <https://github.com/jtzach/LLM-based-AI-surveys-MiD2017>.

It contains:

- Code generated or adapted via GPT-based prompt methods,
- Detailed evaluation scripts (Python) and analysis notebooks,
- Extensive persona prompts used in guided LLM methodology,
- Supplementary hints and strategy documentation for persona modeling, and
- An extracted subset of the dataset corresponding to the study’s tables and results.

This open-source package complies with COMMTR’s data and code sharing policy and ensures that readers can reproduce and verify the findings of this study. The dataset ID is ETS-Data-25-00051. The ETS Data deposit also includes an explanatory replication file/README describing the workflow and how to reproduce each table and figure, consistent with COMMTR’s Data and Code Policy.

Dataset DOI: <https://doi.org/10.26599/ETSD.2025.9190065>.

Acknowledgements

This research has been supported by the European Union’s

Horizon Europe programme (Grant No. 101076963) [project PHOEBE (Predictive Approaches for Safer Urban Environment)].

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article. The author Constantinos Antoniou is the Editor Board member of this journal.

Electronic Supplementary Material

Supplementary material is available in the online version of this article at <https://www.youtube.com/watch?v=SD6aj5t5nuY>.

References

- Beckman, R. J., Baggerly, K. A., McKay, M. D., 1996. Creating synthetic baseline populations. *Transp Res Part A Policy Pract*, **30**, 415–429.
- 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. *Polit Anal*, **32**, 401–416.
- Bundesministerium für Verkehr und digitale Infrastruktur (BMVI), 2018. Mobilität in Deutschland: Tabellarische Grundausswertung. https://www.mobilitaet-in-deutschland.de/archive/pdf/MiD2017_Tabellenband_Deutschland.pdf
- 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.
- 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), 73–78.
- Consorcio Regional de Transportes de Madrid (CRTM), 2018. Resultados de la EDM 2018. <https://datos.comunidad.madrid/dataset/resultados-edm2018>
- Farooq, B., Bierlaire, M., Hurtubia, R., Flötteröd, G., 2013. Simulation based population synthesis. *Transp Res Part B Methodol*, **58**, 243–263.
- Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., Yu, D., 2024. Scaling synthetic data creation with 1,000,000,000 personas. <https://doi.org/10.48550/arXiv.2406.20094>
- Havrilla, A., Dai, A., O'Mahony, L., Oostermeijer, K., Zisler, V., Albalak, A., et al., 2024. Surveying the effects of quality, diversity, and complexity in synthetic data from large language models. <https://doi.org/10.48550/arXiv.2412.02980>
- Liu, R., Wei, J., Liu, F., Si, C., Zhang, Y., Rao, J., et al., 2024. Best practices and lessons learned on synthetic data. <https://doi.org/10.48550/arXiv.2404.07503>
- Livieris, I. E., Alimpertis, N., Domalis, G., Tsakalidis, D., 2024. An evaluation framework for synthetic data generation models. <https://doi.org/10.48550/arXiv.2404.08866>
- Lovelace, R., Birkin, M., Clarke, M., Ballas, D., 2016. From big noise to big data: Toward the verification of large synthetic populations. *Environ Plann B*, **43**, 515–534.
- Ma, B., Yoztyurk, B., Haensch, A. C., Wang, X., Herklotz, M., Kreuter, F., et al., 2024. Algorithmic fidelity of large language models in generating synthetic German public opinions: A case study. <https://doi.org/10.48550/arXiv.2412.13169>
- MiD, 2017. MiD 2017 tabellenband deutschland. https://www.mobilitaet-in-deutschland.de/archiv/pdf/MiD2017_Tabellenband_Deutschland.pdf
- Narayanan, S., Antoniou, C., 2022. Expansion of a small-scale car-sharing service: A multi-method framework for demand characterization and derivation of policy insights. *J Transp Geogr*, **104**, 103438.
- Narayanan, S., Antoniou, C., 2023. Shared mobility services towards Mobility as a Service (MaaS): What, who and when? *Transp Res Part A Policy Pract*, **168**, 103581.
- Narayanan, S., Arango, J. P. R., Tympakianaki, A., Frederix, R., Antoniou, C., 2025. Can emerging mobility solutions complement public transport and lead to a sustainable future? : A case study on Regensburg, Germany. *Case Stud Transp Policy*, **19**, 101338.
- Narayanan, S., Salanova Grau, J.M., Frederix, R., Tympakianaki, A., Masegosa, A.D., Antoniou, C., 2023. Modeling of shared mobility services - An approach in between aggregate four-step and disaggregate agent-based approaches for strategic transport planning. *J Intell Transp Syst*, **29**, 435–452.
- Nowok, B., Raab, G. M., Dibben, C., 2016. synthpop: Bespoke creation of synthetic data in R. *J Stat Soft*, **74**, 1–26.
- OpenAI, 2023. GPT-4 technical report. <https://doi.org/10.48550/arXiv.2303.08774>
- Peng, M., Guo, X., Chen, X., Chen, K., Zhu, M., Chen, L. et al., 2025. LC-LLM: Explainable lane-change intention and trajectory predictions with Large Language Models. *Commun Transp Res*, **5**, 100170.
- Sun, L., Erath, A., 2015. A Bayesian network approach for population synthesis. *Transp Res Part C Emerg Technol*, **61**, 49–62.
- Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., et al., 2024. Large language models for data annotation and synthesis: A survey. <https://doi.org/10.48550/arXiv.2402.13446>
- Tzachristas, I., 2024. Creating an LLM-based AI-agent: A high-level methodology towards enhancing LLMs with APIs. <https://doi.org/10.48550/arXiv.2412.13233>
- Tzachristas, I., Narayanan, S., Antoniou, C., 2025. Guided Persona-based AI Surveys: Can we replicate personal mobility preferences at scale using LLMs? <https://doi.org/10.48550/arXiv.2501.13955>
- Veselovsky, V., Ribeiro, M. H., Arora, A., Josifoski, M., Anderson, A., West, R., 2023. Generating faithful synthetic data with large language models: A case study in computational social science. <https://doi.org/10.48550/arXiv.2305.15041>
- Wang, K., Zhu, J., Ren, M., Liu, Z., Li, S., Zhang, Z., et al., 2024. A survey on data synthesis and augmentation for large language models. <https://doi.org/10.48550/arXiv.2410.12896>
- Yu, J., Zhao, J., Miranda-Moreno, L., Korp, M., 2025. Modular AI agents for transportation surveys and interviews: Advancing engagement, transparency, and cost efficiency. *Commun Transp Res*, **5**, 100172.



Ioannis Tzachristas is a Ph.D. candidate at the Chair of Transportation Systems Engineering, School of Engineering and Design, Technical University of Munich (TUM), Germany, and a Researcher at the Huawei Munich Research Center, Germany. He holds a diploma (M.Sc. & M.Eng.) in electrical and computer engineering from the National Technical University of Athens (NTUA), Greece. His current research focuses on the applications of large language models (LLMs) in transportation systems and security. His interests include AI-driven transport simulation, LLM-based AI-agents, formal verification, cybersecurity and mathematical modeling. He won a Gold Medal at the International Mathematics Competition for University Students (IMC) 2021.



Santhanakrishnan Narayanan is a Postdoctoral Researcher at the Technical University of Munich (TUM), Germany. He earned the B.S. degree in civil engineering from SASTRA University in 2014 and the M.Sc. and Dr.-Ing. degrees in transportation systems from TUM in 2019 and 2023, respectively. He received the best paper award from the TRB Urban Freight Committee in 2019. His expertise lies in travel behavior analysis and modeling emerging mobility services, integrating statistical and data-driven methodologies.



Constantinos Antoniou is a Full Professor in the Chair of Transportation Systems Engineering at the Technical University of Munich (TUM), Germany. He holds a diploma in civil engineering from the NTUA (1995), the M.S. degree in transportation (1997) and the Ph.D. degree in transportation systems (2004), both from MIT. His research focuses on transport policy, modeling and optimization of transportation models, data analytics and statistical learning for transportation, and human factors for future.