

Can Large Language Models Replace Human Survey Respondents? Evidence from a Comprehensive Transportation Survey

Adeeba Naz

The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin, TX 78712, USA
Email: an36723@my.utexas.edu

Dale Robbennolt

The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin, TX 78712, USA
Email: dar4836@utexas.edu

Abdul Rawoof Pinjari

Indian Institute of Science
Department of Civil Engineering
Centre for Infrastructure, Sustainable Transportation, and Urban Planning (CiSTUP)
Bengaluru 560012, India
Email: abdul@iisc.ac.in

Ram M. Pendyala

Arizona State University
School of Sustainable Engineering and the Built Environment
660 S. College Avenue, Tempe, AZ 85287-3005, USA
Email: ram.pendyala@asu.edu

Chandra R. Bhat (Corresponding Author)

The University of Texas at Austin
Department of Civil, Architectural and Environmental Engineering
301 E. Dean Keeton St. Stop C1761, Austin, TX 78712, USA
Email: bhat@mail.utexas.edu

ABSTRACT

Large Language Models (LLMs) are increasingly being explored as a potential means of generating travel survey data, particularly as traditional household activity-travel surveys face growing challenges related to declining response rates, increasing costs, respondent burden, and sampling limitations. While recent studies suggest that state-of-the-art LLMs can generate human-like responses in narrowly defined travel choice settings, such as mode choice among existing transportation alternatives, little is known about their ability to produce coherent, human-like responses across an entire transportation survey spanning multiple behavioral domains and including emerging transportation technologies. To address this gap, we use Gemini 2.5 Pro to generate synthetic survey responses for each respondent in the 2025 U.S. Nationwide Transportation Heartbeat of America (THA) survey based solely on the respondent's observed individual, household, and geographic characteristics. We then evaluate the resulting dataset along three complementary dimensions: (a) its ability to reproduce aggregate trends in travel behavior, (b) its ability to maintain internal response consistency, while also remaining appropriately sensitive to question wording; and (c) its ability to replicate meaningful behavioral differences across population subgroups. The results show that the LLM-generated responses reasonably reproduce aggregate patterns for relatively simple outcomes, such as driver status and vehicle ownership. However, for more behaviorally complex outcomes, the synthetic responses exhibit compressed variation, excessive concentration in dominant response categories, unrealistic attitudinal consistency, and limited ability to reproduce behavioral heterogeneity across population subgroups. Overall, the findings suggest that, while current LLMs show promise as tools for complementing traditional survey methods, they should not yet be viewed as substitutes for human respondents in comprehensive transportation surveys.

Keywords: Large Language Models, Data Generation, Synthetic Survey Data, Travel Behavior, Transportation Surveys

1. INTRODUCTION

Large Language Models (LLMs) have been emerging as powerful tools with promising applications across a wide range of domains. They are being explored based on capabilities extending well beyond traditional natural language processing tasks, to include pattern recognition, logical reasoning, and decision-making (Naveed et al., 2025). LLMs are pre-trained on large and diverse corpora, such as news articles, web-based discussion groups, and books, enabling them to absorb a wide range of human opinions and perspectives (Zhao et al., 2026). This allows LLMs to internalize patterns of human thought, language, and social behavior, enabling them to potentially emulate key aspects of human cognition and decision-making (Wang et al., 2024). While not necessarily guaranteeing their performance in making realistic behavioral inferences, this capacity to learn associations between human attributes, contexts, and responses supports the idea that LLM outputs may vary with contextual information. In this sense, their value lies not only in the diversity of ideas they encode, but also in their potential sensitivity to social and geographic context-dependent narratives (Li et al., 2024a).

The aforementioned abilities of LLMs arise from their generative architecture and the way they process contextual information. Specifically, as deep generative neural networks equipped with in-context learning (ICL), they are optimized to predict the next word (or part of the word) in a sequence based on preceding context. This predictive emphasis enables modern LLMs to adapt their responses to different scenarios, personas, or question formats, making them suitable for interactive applications and potentially capable of emulating human respondents in survey-like settings (Brown et al., 2020; Chang et al., 2024). The natural question, then, is whether LLMs can serve as viable surrogates for human respondents in survey-based research (Aher et al., 2023; Brand et al., 2023). From a methodological standpoint, recent advances in LLM development make this possibility increasingly plausible, particularly given the growing challenges with survey data collection (including being costly, time-consuming, and resource-intensive, as well as facing challenges related to respondent fatigue, declining response rates, and difficulties in reaching underrepresented populations; see National Research Council, 2013; Bonnel and Munizaga, 2018; Wang et al., 2023a).

This possibility has prompted research on whether LLM-generated survey datasets, sometimes referred to as “silicon samples,” could complement traditional survey methods or serve as an auxiliary data source. While many existing studies examining this avenue have originated in the areas of political forecasting (Argyle et al., 2023; Li et al., 2024b), public opinion formation (Hua et al., 2024; Lee et al., 2024; Ma, 2025), economic decision-making (Horton, 2023), psychological behavior patterns (Demszky et al., 2023; Dillion et al., 2023), and social studies (Simmons, 2023; Geng et al., 2024), similar studies have also begun to emerge in the travel behavior field. Examples include those assessing the potential of LLMs to provide responses related to travel mode choice (Mo et al., 2023; Liu et al., 2024; Alsaleh and Farooq, 2025; Nishida et al., 2025), trip planning (Chaudhuri et al., 2025; Karmakar et al., 2025; Shao et al., 2025), value of travel time (VOT) estimation (Yan et al., 2026), route choice modeling (Wang et al., 2025), and travel diary generation (Li et al., 2024c). Collectively, these studies suggest that LLMs can reproduce certain travel behavior patterns under specific task settings, especially when provided with detailed profiles, examples, and/or contextual information. However, they also indicate that discrepancies remain between simulated and observed human responses. For instance, Liu et al. (2024) examined GPT-3.5 and GPT-4 and found that when the LLM was given only the relevant information about a mode choice instance (just the exogenous variables) and asked to predict the traveler’s most probable mode choice (a zero-shot approach), its predictions were behaviorally

misaligned with observed human responses. Alignment with observed responses improved when the prompt in GPT-4 included a few example mode choice instances (with both exogenous variables and the corresponding observed outcome; referred to as a few-shot approach), or when the prompt included a traveler persona representing the traveler’s underlying habits, valuations, attitudes, and behavioral preferences (a persona loading approach). However, studies have found that the effectiveness of few-shot prompting depends on the demonstration examples provided, and may even decline when too many examples are included (Tang et al., 2025). In another study, Yan et al. (2026) analyzed VOT using GPT-4o, Gemini-2.5 Pro, and Claude-Sonnet-4, showing that all three demonstrated human-like sensitivity to travel purpose, income, and time-cost trade-offs. But the LLMs differed substantially in the magnitude and direction of sensitivity to travel contexts and income elasticity, suggesting the need for rigorous validation of LLMs before their application.

The studies just discussed, and many other transportation studies discussed later in Section 2, have examined the predictive performance of LLMs in a variety of relatively specific choice contexts (such as travel mode choice, route choice, or VOT estimation). But, to the best of the authors’ knowledge, no prior study has systematically evaluated whether LLMs can generate coherent responses across a full-length transportation survey covering multiple behavioral domains. In such full-length survey settings, LLMs will need to (a) replicate population trends across multiple dimensions simultaneously, while also (b) maintaining individual-level covariations and human-like consistency across the many responses (at least at a general level, as has been established in investigations of human responses in surveys; see Krosnick et al., 2018; Bessette et al., 2021), and (c) preserving subgroup heterogeneity. Whether LLM-generated responses can reliably meet each of these three criteria when generating full-length survey responses, particularly in the context of travel behavior, remains largely underexplored.

Motivated by this gap, this paper seeks to assess the ability of a state-of-the-art LLM (Gemini 2.5 pro; *Google Gemini*, 2026) to generate human-like survey responses when conditioned on respondent characteristics. Using data from the Transportation Heartbeat of America (THA) survey (Robbennolt et al., 2025), we generate a corresponding synthetic response for each human respondent in the THA survey, enabling a direct comparison between human and LLM-generated data. We evaluate the resulting LLM-generated dataset along multiple dimensions, exploring responses both in the context of currently available/familiar mobility options (such as walking and public transportation use) and niche and emerging topics (such as potential use of electric vehicles and autonomous vehicles). Through this systematic comparison, the study identifies both the strengths and limitations of a state-of-the-art LLM as synthetic survey data generator and provides guidance on how such models may (or may not) be integrated into survey-based design, data collection, and analysis workflows in travel behavior research.

The remainder of this paper is organized as follows: Section 2 provides an overview of the literature on data generation using LLMs, focusing specifically on the three evaluation criteria just discussed. Then, Section 3 describes the data collection process for a baseline sample of human respondents used as a reference for comparison as well as the collection of the LLM-generated data. Section 4 then describes the data comparison process between these two datasets while Section 5 presents the main results. Finally, Section 6 discusses broader implications for policy and future data collection, and Section 7 concludes with recommendations and suggestions for future research efforts.

2. LITERATURE OVERVIEW

A growing number of studies are now examining the extent to which LLMs can mimic human survey responses. While some studies have supported their use as surrogates for human respondents, others have raised concerns regarding their validity. Among these earlier studies, Argyle et al. (2023) demonstrated that the Generative Pre-trained Transformer -3 (GPT-3) can reproduce the viewpoints of demographically varied subpopulations within the domain of U.S. politics and public opinions.¹ Similarly, Kozlowski et al. (2024) showed that GPT-3-simulated respondents representing different ideological profiles can replicate observed partisan differences in COVID-19 attitudes at rates significantly greater than chance. Conversely, other studies have raised important methodological concerns regarding the reliability of LLM-generated survey data. For instance, using questions on abortion and unauthorized immigration from the 2020 American National Election Studies (ANES), Li et al. (2025a) found that the LLM-generated responses exhibited severe homogenization, particularly underrepresenting minority opinions. Similarly, Boelaert et al. (2025) concluded that the LLMs used in their study (GPT, Llama, and Mixtral) could not reliably replicate human research subjects for opinion or attitudinal research due to topic-dependent biases and unrealistic homogeneity. Collectively, these studies suggest that while LLMs show promise as tools for generating synthetic survey data at scale, their use as substitutes for human respondents warrants further investigation. Building on this emerging body of work, the remainder of this section reviews three areas that parallel the investigation we undertake: (a) the ability of LLM to identify overall trends in attitudes, preferences, and behaviors, (b) the ability of LLM to balance response consistency and sensitivity to item wordings (particularly in the context of attitudinal responses), and (c) the ability of LLM to distinguish preferences and behaviors across population subgroups.

2.1 Identifying Overall Trends in Attitudes, Preferences, and Behaviors

The first line of basic face-validity assessment is to examine whether survey responses generated by LLMs accurately capture population-level trends. This concern applies both to general-purpose pre-trained models and models fine-tuned for specific domains, since any biases present in the underlying training data may persist or even be reinforced in the generated responses (Bolukbasi et al., 2016; Li et al., 2025b). This issue is particularly concerning because of limited transparency surrounding the training data used to develop many large-scale pre-trained LLMs (Longpre et al., 2024). Unlike human survey datasets that are deliberately constructed for research purposes using pre-specified and well-documented sampling strategies, the training data used for LLMs are largely drawn from publicly available web content, including social media platforms, online forums, and other user-generated text (Liu et al., 2025a). Consequently, the patterns learned by LLMs reflect the types of information and narratives that appear most frequently in online text rather than information collected through structured measurement. Indeed, most contemporary LLMs rely heavily on “Common Crawl,” one of the largest publicly available web-scraped datasets for training purposes, even though web crawling processes do not incorporate explicit mechanisms to ensure mirroring of the “entire web” (Baack, 2024). As a result, such sources may overrepresent certain demographic groups, viewpoints, and styles of expression, while underrepresenting others (Santurkar et al., 2023; Li et al., 2024d).

¹GPT-3 is an LLM based on the Transformer neural network architecture (Vaswani et al., 2017). GPT-3 was pretrained on a large corpus of text using a self-supervised learning objective, enabling it to perform a broad range of natural language tasks, including zero-shot and few-shot prediction, without task-specific parameter updates (Brown et al., 2020).

Furthermore, even if LLMs were trained on and were able to access all web content, empirical evidence shows that a relatively small minority of users generate a disproportionate share of online content. For example, on platforms such as Facebook, Amazon, and Twitter, about 5% of users are responsible for roughly half of all posts and edits. Similarly, the first versions of half of all English-language Wikipedia articles were created by only 0.04% of registered editors (Baeza-Yates, 2018). Thus, web-based text corpora often reflect highly active contributors rather than the broader population. Additionally, differences in internet usage across demographic groups can further affect this representativeness. For instance, the training dataset used for GPT-2 was constructed by scraping text from webpages linked in posts of Reddit (Radford et al., 2019), a platform whose daily active users are disproportionately located in the United States, are male, and young (Hernandez, 2025). Such demographic imbalances can influence which perspectives and topics become prevalent in LLM training data. As a result, web-scale training corpora may overrepresent highly visible online viewpoints, even including harmful or exclusionary views (Bender et al., 2021). When models learn from such data, their outputs may reproduce or amplify these patterns rather than reflecting broader population views.

Finally, reliance on historical web-based training data raises questions about whether LLMs possess sufficient knowledge to generate reliable responses about emerging or future trends. Humans can reason about unfamiliar or hypothetical situations through mental simulation and counterfactual thinking, even when direct experience or data are limited (Byrne, 2007, 2017). In contrast, LLMs learn patterns from text available in their training corpora, limiting their knowledge to topics already discussed online to some extent. Consequently, their performance may be poor in attitudes toward emerging or context-specific issues, such as new transportation technologies/policy instruments and under-represented local mobility concerns. Indeed, empirical evidence shows that LLM performance depends strongly on how frequently relevant information appears in their training corpora, with models performing better on widely represented knowledge than on less documented topics (Holmes et al., 2023; Kandpal et al., 2023; Sun et al., 2024). Even if these emerging topics are represented, their coverage may be at a high-level or in a reactionary manner, reflecting vocal opinions rather than the broader population. For example, early opinions about new technology are usually shaped by discussions driven by a small group of early adopters (Frattoni et al., 2014). Thus, emerging technologies often experience a “peak of inflated expectations,” where discourse is dominated by overenthusiasm despite modest real-world adoption (Borup et al., 2006; Shi and Herniman, 2023), followed by (typically) a phase of relative disillusionment before stabilizing at more realistic levels (Dedehayir and Steinert, 2016). Consequently, if training data disproportionately capture early hype, LLMs may overestimate technology adoption. Conversely, for technologies with limited real-world deployment and few first-hand usage narratives, public discourse may be more strongly shaped by uncertainty, perceived risk, and lack of trust (Kenesei et al., 2022). In such cases, LLM-generated responses may emphasize uncertainty and non-adoption. Predicting adoption for these technologies may therefore be particularly challenging because LLM performance depends on the prevalence and quality of relevant information in the training data (Kandpal et al., 2023). This issue is especially relevant for transportation, an evolving field shaped by emerging technologies such as electric vehicles (EVs), autonomous vehicles (AVs), and urban air mobility (UAM), whose future adoption and impacts remain uncertain (Huang et al., 2023). Hence, this study seeks to examine whether LLMs can reproduce the adoption intention of emerging transportation modes.

2.2 Balancing Response Consistency and Sensitivity to Item Wordings

A second concern in LLM-generated survey data is whether LLMs can maintain internal consistency across multiple survey items while remaining sensitive to wording and context. In traditional surveys, human respondents exhibit both of these properties to a meaningful degree. Basically, psychological studies suggest that humans possess an intrinsic drive to maintain cognitive harmony and coherence among their beliefs, attitudes, and behaviors as if their opinions exist in internally consistent clusters (Morvan and O'Connor, 2017), even if their responses may still vary slightly due to differences in wording, framing, or their frame of mind at different points in the survey (Krosnick et al., 2018). For LLM-generated data to be analytically useful, LLMs must replicate this behavior by maintaining a relatively stable persona and producing coherent responses across the entire survey instrument, while still allowing reasonable variation due to item wordings and context. Although there is growing evidence regarding the ability of these models to replicate human text and maintain context-specific consistency across long prompts (Wang et al., 2023b), the extent to which LLMs truly achieve this balance remains an open question. Prior research shows that LLM outputs can be highly sensitive to subtle variations in prompt wording, and LLMs may not be able to hold semantic consistency in long conversations. Thus, even minor paraphrasing while preserving the central meaning may lead to different or incorrect responses, especially in long conversations (Ye et al., 2023; Raj et al., 2025).

This issue is especially important for studies relying on latent attitudinal constructs, as capturing these aspects of behavior requires both structured consistency across related items within the responses representing a single respondent as well as systematic variation across respondents. A good deal of research in the psychological domain (as well as across other fields, such as transportation) has employed such latent attitudinal constructs and developed techniques to capture underlying attitudes while reflecting meaningful differences in preferences across the indicators used for measurement. In the context of LLM-based survey responses, some recent work has used similar techniques using multi-item instruments, such as psychological learning scales and health-related surveys, and evaluated whether the resulting responses reproduce expected attitudinal latent factor structures (Vogelsmeier et al., 2025; Villarreal-Zegarra and Bellido-Boza, 2026). However, a limitation of these existing studies that have specifically assessed response consistency by examining attitudinal latent constructs is that they do not include human comparison datasets and instead assess the plausibility of the latent constructs resulting from LLM-generated data against theoretical expectations and prior psychometric findings. Furthermore, despite the fact that many transportation planning studies incorporate such attitudinal constructs (Bhat and Dubey, 2014; Kroesen et al., 2017), to the best of our knowledge, this issue has not yet been examined in the transportation domain. We address this gap by evaluating whether LLM-generated responses maintain internal coherence across related survey items and whether the latent constructs derived from these responses align with those observed in human data.

2.3 Distinguishing Preferences and Behaviors among Population Subgroups

A final concern regarding LLM-generated synthetic data is whether they can accurately reproduce behavioral differences across population subgroups. This capability is essential in transportation research because heterogeneity in preferences and behaviors across population subgroups is central to understanding travel demand and designing targeted interventions (Zheng et al., 2021; Barri et al., 2022). If synthetic data do not preserve this heterogeneity, downstream analyses may homogenize the population and produce misleading conclusions.

Argyle et al. (2023) labeled this issue as “algorithmic fidelity,” which they defined as the degree to which complex relationships between ideas, attitudes, and socio-cultural contexts within a model mirror those observed across human population subgroups. They provided one of the strongest early arguments that LLMs can exhibit algorithmic fidelity. They conditioned GPT-3 on socio-demographic backstories from real American National Election Studies (ANES) respondents. Then, they assessed algorithmic fidelity for a vote prediction task mainly through one-variable subgroup checks (such as comparing the voting behavior of men in the ANES data with that of men in the generated data) and reported high correspondence between GPT-3 and ANES vote choices for most of these groups. Bisbee et al. (2024), using ChatGPT 3.5 Turbo-generated persona responses matched to respondents from the 2016-2020 ANES, found that synthetic responses often approximated human responses in aggregate, but exhibited substantially less variation than the human respondents. Across subgroups defined by race/ethnicity and party identification, ChatGPT-generated responses were often more extreme than those of comparable human respondents. In particular, synthetic responses exaggerated out-group antipathy along partisan and ideological lines, while evaluations of racial and religious groups exhibited compressed variation and subgroup-level distortions. Similarly, Santurkar et al. (2023) examined how closely LLM responses align with public opinion across the overall U.S. population as well as within sixty demographic groups and found substantial misalignment. The study also identified specific groups whose perspectives were particularly poorly represented in model outputs, including older adults and widowed individuals. Consistent with the Santurkar et al. (2023) study, a recent industry report applying the same demographic-persona approach to political polling found that while the resulting synthetic samples could often recover the direction of a national partisan signal, they did not preserve the individual- or subgroup-level structure needed to replace real survey data (Morris et al., 2026).

Finally, a small but growing body of work has begun to address this question of algorithmic fidelity in transportation contexts, though these studies often evaluate heterogeneity in a rather limited way. Li et al. (2024c) tested algorithmic fidelity indirectly by examining whether adding age, income, and occupation improves the realism of generated travel diaries. They found that finer-grained demographic combinations produced trajectories that more closely resembled real-world mobility patterns, demonstrating that individual attributes meaningfully shape LLM-generated behavior. Yan et al. (2026) took a more direct approach by examining how LLM-generated VOT estimates vary across traveler profiles and trip contexts, including income, gender, age, educational attainment, and travel purpose. Their findings showed that LLMs responded to these factors in directions broadly consistent with human evidence, though the magnitude of income sensitivity differed notably from human benchmarks. Liu et al. (2025b) incorporated socio-demographic information through a persona inference and loading framework. In their approach, they first inferred personas from detailed observations that included respondents’ socio-demographic characteristics and repeated observed choices, and then used a learned persona-loading function based on behavioral embeddings to condition LLM-based mode choice predictions. They showed that this approach outperformed lightweight LLM-based alignment methods in both aggregate mode-share prediction and individual choice accuracy. Sameen et al. (2025) synthesized qualitative traveler personas from demographic, household, spatial, and travel survey data using an LLM and then assigned quantitative scores to theory-guided latent attitudes (including time sensitivity, cost sensitivity, tech affinity, pro-car attitude, environmental concern, convenience/comfort, and spontaneity). They then provided a limited subgroup-level validation of the LLM-generated personas by examining whether the synthesized latent attitudes varied

plausibly across ridesourcing users and non-users, as well as across age, income, and vehicle-availability segments, and concluded that the generated personas captured meaningful behavioral differences across traveler groups.

Taken together, these earlier studies offer encouraging early evidence that LLMs can reflect meaningful demographic variation in travel behavior. However, they share an important limitation: each evaluates heterogeneity across a narrow set of demographic variables and a single behavioral outcome, without systematically assessing how LLM-generated responses align with human behavior across a wide range of outcomes and demographic characteristics. The current study addresses this gap directly by estimating discrete choice models for a broad range of travel-related outcomes and evaluating whether LLM-generated responses reproduce the sensitivity of these outcomes to individual, household, and geographical characteristics.

2.4 Present Study in Context

Recent transportation studies have begun exploring whether LLMs can support travel behavior analysis by predicting or generating different forms of mobility behavior. The body of existing transportation-focused studies using LLMs for travel behavior analysis is summarized in Table 1, highlighting several aspects of each study with relevance to the current analysis. This includes (1) the scope of the LLM-generated output (including the specific mobility dimension considered, and the general data/analysis approach), (2) the study goal (identified as being either for the prediction of specific outcome behaviors or data generation, as discussed further in the next paragraph), (3) the exogenous variables provided to the LLM as the basis for the LLM responses, (4) the comparisons made to validate the LLM-generated data (comparisons with existing literature versus comparisons with actual individual-level ground truth data), (5) the assessment criteria used (that is, which of the three considerations as discussed in Sections 2.1 through 2.3 were used for assessment), and (6) whether each study considered emerging transportation mobility topics.

In the context of Table 1, there are several salient aspects of the current study. First, regarding the scope and goals of the data collection effort, many existing studies employ LLMs to predict a specific outcome within a predefined choice setting (see the “Prediction” sub-column under the “Study goal” main column of Table 1). These studies generally focus on fairly isolated behavioral outcomes, such as a single mode choice instance (Mo et al., 2023; Chen et al., 2024; Liu et al., 2024; Alsaleh and Farooq, 2025; Liu et al., 2025b; Nishida et al., 2025), several of which are targeted for integration as predictive modeling components into larger transportation planning frameworks (Mo et al., 2023; Alsaleh and Farooq, 2025; Liu et al., 2025b; Nishida et al., 2025). In contrast, and more closely aligned with the current study, a second set of studies uses LLMs in a more generative role, producing synthetic mobility, behavioral, or travel-planning outputs (see the “Data generation” sub-column). Within this group, some studies generate direct-use travel plans or itineraries as an end product for the traveler (Chaudhuri et al., 2025; Karmakar et al., 2025; Shao et al., 2025), while others generate behavioral or survey-like data intended for downstream modeling and analysis (Li et al., 2024c; Liu et al., 2025c; Sameen et al., 2025; Wang et al., 2025; Yan et al., 2026). The current study is more closely aligned with the latter set of studies, as it aims to generate synthetic responses that can be used for travel behavior analysis. But rather than focusing on specific choice tasks or isolated transportation behaviors as in earlier studies, we consider the case of a complete transportation survey covering a wide range of attitudes and behaviors, enabling us to evaluate LLM performance across various transportation domains as well as investigate consistency in LLM responses across the entire survey.

Table 1: Recent Travel Behavior Modeling Studies Using LLMs

Study	Scope of LLM-generated output	Study goal		Exogenous variables provided				Comparison		Assessment criteria			Consideration of emerging topics
		Prediction	Data generation	Socio-demographic	Geographic	Context-specific	Attitudes/personas	Literature	Ground truth	Aggregate trends	Response consistency	Subgroup heterogeneity	
Mo et al. (2023)	- Mode choice and trip purpose predictions with explanations - Text embeddings of travel scenarios as features for supervised models	•		•		•			•				•
Chen et al. (2024)	Decision and reasoning to wait and continue by metro or abandon metro during delays	•				•			•				
Liu et al. (2024)	- Mode choice predictions and reasonings - Persona-loaded prompts from LLM-inferred traveler traits	•		•		•	•		•	•			•
Alsaleh and Farooq (2025)	Mode choice predictions and rationales across SP and RP datasets	•		•		•			•	•			•
Liu et al. (2025b)	- Mode choice predictions - Personas inferred from socio-demographics and repeated observed choices - Persona-loaded prompts selected using learned behavioral embeddings	•		•		•	•		•	•		•	•
Nishida et al. (2025)	Mode choice model generalizable across alternative sets	•		•		•			•				•
Chaudhuri et al. (2025)	Multi-day itineraries under persona and constraint settings		•	•	•	•	•						
Karmakar et al. (2025)	Disruption-aware travel itineraries revision		•		•	•	•						
Shao et al. (2025)	- Personalized multi-day travel plans - Structured user preference profiles inferred from review histories		•		•	•	•						
Li et al. (2024c)	- Mobility patterns inferred from anonymized trajectories and individual profiles - Personalized travel diaries through motivation reasoning		•	•	•	•			•	•		•	
Liu et al. (2025c)	- Daily activity schedules, tours and travel plans - Plan revisions from simulated feedback		•	•	•	•	•						
Sameen et al. (2025)	- Structured persona with motivation, behavioral tags, and brief justification - Personas scored on theory-grounded latent attitudes		•	•	•			•				•	
Wang et al. (2025)	Repeated daily route choice and reasoning		•	•	•	•	•	•		•			
Yan et al. (2026)	SP rankings for 13 driving alternatives with explanations		•	•		•		•		•		•	
Current Study	Survey responses on travel behaviors, vehicle ownership, mode-use, teleworking, package receiving, policy preferences, and emerging technology familiarity/adoption intention		•	•	•				•	•	•	•	•

Second, compared with prior transportation studies using LLMs, our approach differs in the exogenous variables provided to the model and the way traveler characteristics are represented. Many previous studies provide LLMs with socio-demographic information, geographic details, context-specific attributes (such as travel times and costs by alternative modes in a mode choice prediction effort), and sometimes predefined attitudes or personas (see the main column of Table 1 labeled “Exogenous variables provided”). Some studies additionally incorporate attitudes as inferred behavioral traits (such as time and cost sensitivity, flexibility, or comfort preferences) to construct personas, and some others provide attitudes as explicit persona descriptions (including traveler types such as “laid-back travelers” or “adventure seekers,” as well as traveler spending preferences such as “luxury” or “budget-conscious”). In contrast, we consider attitudes while allowing the LLM to implicitly construct a persona for each respondent rather than assigning predefined personas or specifying latent preferences a priori. The LLM prompt includes socio-demographic and geographic attributes available from the THA survey dataset (including age, gender, employment status, educational attainment, race, ethnicity, income level, household composition, housing tenure type, housing unit type, and zip code). By including the respondent’s zip code, we explicitly anchor each synthetic agent within a local geographic context. Thus, our approach examines whether LLMs can generate contextually grounded and coherent responses across an entire survey instrument by generating a persona and associated attitudes for each respondent, using only observable respondent characteristics.

Third, we consider the Transportation Heartbeat of America (THA) survey, a U.S. national travel behavior survey undertaken in late 2024 and early 2025 (see Robbennolt et al., 2025), as an empirical benchmark (reference) dataset. Synthetic responses for each individual represented in the THA survey are generated by Gemini 2.5 Pro, conditioned on individual/household socio-demographics and their respective geographic contexts.² This enables direct comparison between human and model-generated responses at both aggregate and subgroup levels across the survey. As shown in the main column of Table 1 labeled “Comparison,” some existing studies compare their results with literature or existing theory rather than with ground truth data. Also, among those studies that have made direct comparisons with ground truth data, it is typical to use publicly available datasets, most commonly the Swissmetro dataset (see Mo et al., 2023; Liu et al., 2024; Alsaleh and Farooq, 2025; Liu et al., 2025b; Nishida et al., 2025). This dataset was collected in 1998 (not aligning with the timing of collection of LLM-based data/predictions or the knowledge available to these models) and has since been widely available through prior literature and online repositories. Consequently, the possibility that the dataset itself, or behavioral patterns derived from it, appeared in LLM training corpora cannot be ruled out. Hence, strong performance of LLMs on such public benchmark data may overestimate the capabilities of LLM data generation. Alsaleh and Farooq (2025) acknowledge this issue by comparing datasets with different levels of accessibility (open-access: Swissmetro, London Passenger Mode Choice, and proprietary restricted-access: Brightwater survey). However, their results do not directly isolate prior exposure effects, as performance differences are instead attributed to choice-set complexity and stated-versus revealed-preference survey design, with no accessibility-specific analysis reported. To address this concern, the present study uses the more recent (and not yet publicly available) THA

²Gemini 2.5 Pro is a natively multimodal “thinking” model developed by Google DeepMind that uses a sparse mixture-of-experts transformers architecture and supports text, image, video, and audio inputs. It was pretrained on a large and diverse corpus comprising publicly available web documents, code in multiple programming languages, images, audio, and video. Its long-context processing and advanced reasoning capabilities enable it to perform complex reasoning, coding, and multimodal understanding tasks (Comanici et al., 2025).

survey as the reference dataset and generates responses using a version of Gemini with a knowledge cutoff date (the specific point in time when an LLM's training data concludes) aligning with the THA survey data collection period (see Section 3). This eliminates the possibility that the generated responses reflect exact training patterns rather than inferred respondent behavior, while aligning the survey response timing with the model's knowledge base.

Fourth, in terms of assessment criteria (the sixth main column of Table 1), we undertake a detailed investigation of (1) the ability of LLM-generated survey data to reproduce aggregate trends of transportation behaviors (see Section 5.1), (2) the ability of the LLM to generate attitudinal profiles and respond with an appropriate level of consistency across the survey (see Section 5.2), and (3) the ability of the LLM to respond to the provided exogenous variables to capture subgroup heterogeneities in transportation behaviors (see Section 5.3). While a good number of existing studies have made these types of comparisons in aggregate (see the “Aggregate trends” sub-column under the main column labeled “Assessment Criteria” in Table 1), we are aware of no existing studies of data generation in the transportation context that consider consistency of attitudinal responses, and only a few that consider heterogeneity in responses across population subgroups. While some existing studies in other domains have developed latent constructs using LLM-generated data (as discussed in Section 2.2), these studies generally do not include human comparison datasets and instead assess plausibility against theoretical expectations and prior psychometric findings. In contrast, we compare attitudinal latent constructs generated using the LLM to those generated using the corresponding benchmark data, enabling an assessment of how well the LLM reproduces the latent attitudinal structures and associated relationships. In terms of subgroup heterogeneity, we estimate comparable models of several transportation outcomes as functions of exogenous socio-demographic and geographic characteristics using reference and LLM-generated datasets. As discussed in Section 2.3, previous studies have examined related forms of heterogeneity, but with a more limited scope. We go beyond that by considering a wider range of transportation behavioral outcomes, identifying differences in the ability of the LLM to make predictions for each outcome.

Finally, and aligning with this last point, we include a variety of both well-established and emerging transportation topics to assess LLM performance across a range of outcomes. While a few existing studies have attempted to assess such emerging topics (see the final column of Table 1), quite a few of these have used the Swissmetro dataset, which (as already discussed) is a relatively dated reference dataset. Instead, we consider currently emerging and evolving topics such as the adoption of EVs, AVs, and UAM, as well as relevant transportation policy topics. The consideration of a variety of outcome types enables us to identify the ability of LLMs to accurately reproduce future behaviors, rather than simply representing existing trends, an important element in a rapidly evolving transportation environment.

3. DATA COLLECTION AND DATA GENERATION

3.1 Reference Survey Dataset Collection and Overview

This study uses data from the first wave of the Transportation Heartbeat of America (THA) Survey, conducted in the United States between October 2024 and January 2025. The survey was administered through an online panel managed by a commercial firm and was designed to reflect the U.S. adult population (see Robbennolt et al., 2025). To achieve this, a quota-based sampling approach was used, with target proportions set across a range of socio-demographic characteristics. This helped ensure that the sample captures adequate variation across key groups defined by race, age, gender, education, household income, area type (urban or rural), and census division. The

survey was collected only from adults aged 18 and above living in the U.S. The survey included several attention checks and quality control questions throughout to ensure response reliability. Specifically, two attention questions were included, instructing participants to select a specific response option to show that they were paying attention. Additionally, several consistency checks were employed to ensure response reliability (for instance, respondents who reported a vehicle with a purchase year more than one year prior to the model year were excluded; similarly, individuals reporting children in their household as employed full-time or drivers were excluded). Finally, a minimum timing cutoff of 10 minutes was applied to remove low-quality responses (the median response time in the final sample was 18.15 minutes). The final dataset consists of 8,030 individual responses and serves as the reference dataset for all subsequent analyses. Descriptive statistics for this reference dataset are provided in Table 2 alongside corresponding 5-year estimates from the American Community Survey (ACS) for 2024 (U.S. Census Bureau, 2026).

Table 2: Descriptive Statistics of the Reference Sample and 2024 ACS 5-Year Estimates

Variable	Reference dataset	ACS	Variable	Reference dataset	ACS
Individual characteristics			Household characteristics		
<i>Age</i>			<i>Annual household income</i>		
18-24 years old	9.7	11.9	\$49,999 or less	39.2	34.3
25-34 years old	15.1	17.4	\$50,000 to \$100,000	30.5	29.1
35-44 years old	19.3	16.9	\$100,000 and above	30.3	36.6
45-54 years old	17.1	15.6	<i>Presence of children</i>		
55-64 years old	15.6	16.2	Household without children under 18	64.4	71.5
65 years old and above	23.2	22.0	Household with children under 18	35.6	28.5
<i>Gender</i>			<i>Housing tenure type</i>		
Male	46.6	48.8	Owner-occupied housing units	59.2	60.0
Female	53.1	51.2	Renter-occupied housing units	35.5	34.8
Other	0.3	-	Others	5.3	5.2
<i>Educational attainment</i>			<i>Housing unit type</i>		
High school or less	33.0	36.4	Single family housing	66.6	62.4
Some college	29.5	27.9	Apartment	19.8	32.4
Bachelor or higher	37.5	35.7	Others	13.6	5.2
<i>Employment status</i>			Residential location by census division in the U.S.³		
Employed full-time	45.6	45.6	New England	8.3	4.7
Employed part-time	11.1	14.6	East North Central	13.1	14.2
Others	43.3	39.8	East South Central	9.4	5.8
<i>Race</i>			West North Central	8.2	6.4
White	63.0	61.0	West South Central	13.2	12.1
Black	14.9	12.2	Pacific	13.0	16.0
Others	22.1	26.8	South Atlantic	12.9	20.4
<i>Ethnicity</i>			Mountain	8.6	7.6
Hispanic	20.1	19.3	Middle Atlantic	13.3	12.8
Non-Hispanic	79.9	80.7			

³New England: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont; East North Central: Illinois, Indiana, Michigan, Ohio, and Wisconsin; East South Central: Alabama, Kentucky, Mississippi, and Tennessee; West North Central: Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, and South Dakota; West South Central: Arkansas, Louisiana, Oklahoma, and Texas; Pacific: Alaska, California, Hawaii, Oregon, and Washington; South Atlantic: Delaware, District of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia; Mountain: Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming; Middle Atlantic: New Jersey, New York, and Pennsylvania.

The dataset provides detailed information on multiple aspects of travel behavior and preferences. These include attitudes toward transportation and mobility, lifestyle and personality traits, current travel behavior, views on transportation investments and policies, familiarity with emerging technologies, and patterns of online activity participation. In addition, built environment characteristics, such as population density, employment density, road network density, job accessibility, low wage worker density, and distribution of vehicles, were appended to each respondent from the Environmental Protection Agency Smart Location Database based on the respondent’s ZIP code (Chapman et al., 2021).

Overall, the survey statistics and ACS estimates shown in Table 2 are reasonably close in the aggregate. Besides, in the current context (including for aggregate comparisons between the reference and LLM-generated datasets), even the relatively minor differences are not of concern because demographic details for each individual (from the THA survey) were provided to Gemini as inputs for data generation (as discussed in the following section). Thus, with matching demographic variables in the reference and LLM-generated datasets, all comparisons at the population and sub-population levels may be made directly without the need for any weighting.

3.2 Synthetic Dataset Generation by LLM and Overview

Synthetic survey responses were generated using Gemini 2.5 Pro. This model was appropriate for the current context because it is suitable for the complex reasoning tasks involved with detailed survey responses, and had a knowledge cutoff in January 2025 (the model was trained only on data available as of January 2025, and was unable to access any information/data after this cutoff; Hundertmark, 2025), aligning almost exactly with the THA survey collection window. For the data generation, each individual respondent from the THA survey was simulated by providing the model with a respondent profile created using a comprehensive set of socio-demographic attributes replicating those solicited in the THA survey: zip code, age, gender, race, ethnicity, level of educational attainment, employment status, household income, housing tenure type, housing unit type, and household composition. These profiles were formatted into structured text descriptions that were provided to the model as background information for each response.

For each agent, the LLM was then prompted to complete the entire THA survey instrument as if it were that individual. The prompt included (i) the agent’s socio-demographic profile, (ii) detailed instructions emphasizing that responses must be selected strictly from the predefined answer options, and (iii) the full list of survey questions presented in sequence. Importantly, the prompt explicitly instructed the model to answer all applicable questions and to skip a question only when the survey logic required it. For example, if an agent was specified as unemployed, the model was instructed not to answer employment-related follow-up questions. It was also asked to report its answers in a consistent format, listing the question ID followed by the selected response.

The same quality control checks were included for the LLM-generated responses as were used in the THA survey (including the trap questions and consistency checks, but not the timing checks). Notably, Gemini passed all data quality checks for every respondent. Gemini also consistently reproduced the input demographic and location information for each respondent when asked in the survey and correctly reproduced the respondent’s state of residence based on their provided zip code for every respondent when asked as part of the survey.

4. DATASET COMPARISON METHODOLOGY

The assessment of the differences between the reference dataset and LLM-generated dataset follows the three-step structure described above to compare (1) the overall aggregate trends in

transportation behaviors and outcomes, (2) the level of internal consistency in attitudinal responses assessed using a confirmatory factor analysis, and (3) the sensitivity of the LLM to socio-demographic and geographic variables through an individual-level modeling effort. These assessment methods are described in the remainder of this section.

4.1 Comparing Overall Aggregate Trends in Attitudes, Preferences and Behaviors

First, the two datasets were compared in terms of their aggregate distribution. The statistical significance of distributional differences between the datasets was initially evaluated using Pearson’s Chi-square statistic. The Chi-square statistic assesses the extent to which observed categorical response frequencies differ from the frequencies expected under distributional equivalence between the two datasets. The test statistic for each variable of interest is given by:

$$\chi^2 = N \sum_{k=1}^K \frac{(p_{THA,k} - p_{GEMINI,k})^2}{(p_{THA,k} + p_{GEMINI,k})} \quad (1)$$

where $N = 8,030$ is the number of individuals in the dataset, k is the number of possible outcome categories, $p_{THA,k}$ is the proportion of individuals selecting the k^{th} outcome category in the reference dataset, and $p_{GEMINI,k}$ is the proportion of individuals selecting the k^{th} outcome category in the LLM-generated dataset. The resulting test statistics may be compared to the critical chi-square value with $(k - 1)$ degrees of freedom to test for a difference in the distributions. However, due to the large size of the datasets, it should be noted that a Chi-square distribution test will often turn up significant even under relatively modest differences in category shares (Sullivan and Feinn, 2012).

Accordingly, Cramér’s V was also calculated for each outcome distribution, which provides a standardized effect-size measure of association for categorical variables while accounting for contingency table dimensionality (Cramér, 1999). Cramer’s V (specifically for the current case of comparing two datasets with the same size) is calculated as:

$$V = \sqrt{\frac{\chi^2}{2N}} \quad (2)$$

where χ^2 denotes the Pearson Chi-square statistic calculated above. The statistic ranges from 0 to 1, where values closer to 0 indicate stronger similarity between the reference and LLM-generated distributions and larger values indicate greater distributional divergence. Unlike Chi-square statistics, Cramér’s V provides a scale-independent measure of the substantive magnitude of categorical differences and is therefore more appropriate for evaluating distributional similarity or divergence for large samples.

In our analysis, we considered a total of 14 mobility-related dimensions for the aggregate comparison between the THA and LLM-generated datasets, as discussed further in Section 5.1.

4.2 Comparing Levels of Internal Consistency of Attitudinal Responses

To examine whether related attitudinal responses exhibit coherent common-factor structures across respondents, we estimate corresponding confirmatory factor analysis (CFA) models using the reference and the LLM-generated datasets. This effort serves two purposes. First, it allows us to assess whether the LLM generates response patterns across related questions that support coherent

latent constructs comparable to those observed in the reference data. Second, it allows us to assess the sensitivity of the model to differences in item wording by comparing responses across the indicators associated with each latent construct. As noted in Section 2.2, human respondents are expected to exhibit some variation in their responses depending on how an item is worded, while remaining broadly consistent with their underlying beliefs. Realistic LLM-generated responses should similarly reflect such variation, while maintaining a high level of consistency across statements representing similar underlying attitudes or beliefs.

We consider three specific latent constructs: (1) pro-transit orientation, (2) pro-environmental disposition, and (3) perception of online shopping (in)accessibility. These three constructs were selected because they span distinct categories of attitudinal and lifestyle-related constructs that are routinely used in mobility literature to explain travel behaviors. The first two are more directly connected to transportation attitudes, capturing mode-related preferences and broader lifestyle-related values. Notably, both are commonly used in travel behavior analysis and have been shown in the literature to be associated with observable respondent characteristics, including individual- and household-level demographics (Lavieri et al., 2017; Chng et al., 2019; Etezady et al., 2021; Shaw et al., 2021). These constructs may therefore be relatively easier for the LLM to infer from the respondent profile. The third relates more to specific accessibility perceptions, in this case for online shopping. As such, it is more experience based and likely to be shaped by a number of factors such as privacy and security concerns, prior online shopping experience, perceived ease of use, and local service quality, many of which are not directly observed in the respondent profile. Thus, this third latent construct provides a meaningful test of whether LLM-generated data can capture experience-based constructs that are harder to infer from a respondent profile, a capability that would need to hold if LLM-generated data are to be useful for the many mobility applications (such as e-commerce and delivery demand or accessibility planning) where such experience-based constructs are central.

The first pro-transit construct is based on ordinal-scale responses to three indicator queries eliciting preferences for (a) transit-accessible residential locations, (b) positive perceptions toward public transportation as a travel mode, and (c) productive use of time spent in, on, or waiting for transportation vehicles. The second pro-environment construct is similarly based on ordinal-scale responses to three other indicator queries reflecting preferences for (a) less-polluting transportation modes, (b) active travel propensity, and (c) broader commitments toward environmentally friendly lifestyles. And the final construct, online shopping (in)accessibility perception construct is also based on ordinal-scale responses to three other indicator queries corresponding to (a) challenges associated with products or services ordered through online platforms, (b) ordering difficulties, and (c) unreliable deliveries.

The indicators for each of the latent constructs were determined through an exploratory factor analysis (EFA) of a broader set of candidate survey items using the THA data. Then, the selected set of indicators for each latent construct was used for the CFA using both datasets. Appendix A provides the detailed mathematical formulation of the CFA framework, which includes a structural equation modeling component (that captures the effect of individual- and household-level demographics on the latent constructs, as well as any correlations across the latent constructs themselves) and a measurement equation modeling component (that captures the loadings of the latent constructs on the propensities underlying the ordinal indicators).

To provide an intuitive and scale-comparable assessment of how the LLM associates respondent characteristics with the attitudinal constructs, we compare the effects of the exogenous

variables in the structural component of the CFA model. Specifically, each latent construct l for individual i is expressed as

$$z_{li}^* = \alpha_l' \mathbf{w}_i + \eta_{li} \quad (3)$$

Because the CFA models are estimated separately for the reference and LLM-generated datasets, the scales of z_l^* may differ across the two datasets, and the raw elements of α_l are therefore not directly comparable. Accordingly, the effects are normalized separately for each latent construct and dataset. For individual i , the predicted latent score is first computed as

$$\hat{z}_{li}^* = \hat{\alpha}_l' \mathbf{w}_i \quad (4)$$

The variance of the $\hat{z}_{li}^*$ is then calculated across individuals i (labeled as $Var(\hat{z}_l^*)$). Because the variance of η_{li} is fixed at one for identification, the total standard deviation of the latent construct is computed as

$$s_l = \sqrt{Var(\hat{z}_l^*) + 1} \quad (5)$$

Then, the elements of the normalized coefficient vector $\tilde{\alpha}_l = \hat{\alpha}_l / s_l$ represent the expected effects of a one-unit change in each exogenous variable on z_l^* , expressed relative to the total standard deviation of that latent construct. Comparing these normalized effects across the reference and LLM-generated datasets allows us to assess whether the LLM translates the characteristics on which each synthetic respondent is conditioned into attitudinal orientations in a manner consistent with the relationships observed in the reference data.

4.3 Comparing Preferences and Behaviors among Population Subgroups

Finally, the two datasets are compared based on their ability to differentiate behaviors among population subgroups. Specifically, rather than making single-group comparisons of the behavioral outcomes, we use a predictive modeling approach to identify behavioral differences across the population. Comparing the results of models based on the reference and the LLM-generated data highlights important ways that the LLM incorporates the provided persona details and the extent to which this aligns with real-world findings. To keep the analysis focused and streamlined, in this disaggregate part of the analysis, we select eight of the 14 dimensions examined in the “aggregate trends” section. Thus, in total, sixteen models are estimated, eight using the reference dataset and eight using the LLM-generated dataset. Of these eight dimensions, six take the form of a traditional univariate ordered-response probit model. But for the remaining two dimensions, corresponding to intentions to use electric vehicles (EVs) and autonomous vehicles (AVs) (full self-driving vehicles), only those familiar with each technology were asked about adoption intention. Thus, in these two cases, to accommodate any unobserved factors that may simultaneously impact knowledge with these technologies and the actual adoption intentions, we jointly model a binary selection equation for EV/AV knowledge and the adoption outcome (which is only available for those with this prior knowledge) for the THA data. However, the corresponding joint models could not be estimated using the LLM-generated data because, as discussed in Section 5.1 later, the LLM responses indicated near-universal EV and AV familiarity, leaving insufficient variation in the familiarity outcome. Therefore, univariate ordered-response probit models were estimated for EV and AV adoption intentions using the LLM-generated data.

Based on the estimates obtained from the estimated models, average treatment effects (ATEs) are computed to provide an intuitive and policy-relevant summary of subgroup-specific behavioral relationships. First, we assign each individual to a base level of an exogenous variable of interest (for instance, assigning all individuals as <25 years old), while retaining the values of all other exogenous variables, and calculate the probability of each individual selecting each outcome level. We then assign each individual to a treatment level (for instance, assigning all individuals as 25-34 years old) and again calculate the probability of each individual selecting each outcome level. The final ATE values are the differences in these probabilities averaged across the entire sample. For the continuous variables, the base level is taken as the 25th percentile value and the treatment is the 75th percentile value for each variable. Furthermore, for the household composition variable measuring the number of children in the household, the ATEs are calculated for a one-unit increase from zero to one child. Then, to further compare the resulting ATE estimates between the two datasets, the relative deviation of the LLM-based ATE from the THA-based ATE is calculated as the percentage difference between the two values relative to the absolute value of the THA-based ATE. Positive values indicate that the LLM-based ATE is numerically larger than the reference-based ATE, while negative values indicate that it is numerically smaller.

5. RESULTS

5.1 Ability to Identify Overall Trends in Attitudes, Preferences, and Behaviors

This section evaluates the extent to which the LLM-generated dataset reproduces the aggregate distributional patterns observed in the reference dataset for 14 mobility-related dimensions, as shown in Table 3. These include driver status (whether an individual currently drives), vehicle ownership⁴, typical mode use frequencies for four travel modes (private vehicle, public transportation, walking, and ride-hailing), typical telework (full-day remote work) frequency, package receiving frequency, opinions related to three transportation policies (investments in safety, investments in public transportation, and congestion pricing), and intentions to use three emerging technologies (electric vehicles, autonomous vehicles, and urban air mobility). The categories for each of these mobility-related dimensions, as well as the corresponding descriptive statistics, are presented in Table 3. Given the large sample size, Chi-square tests indicated significant distributional differences for all variables except current driver status. Thus, Cramer’s V was used as an additional measure of divergence (note that values below 0.10 indicate negligible difference, 0.10-0.20 weak, 0.20-0.40 moderate, 0.40-0.60 relatively strong, 0.60-0.80 strong, and above 0.80 very strong difference between the distributions; see Rea and Parker, 2014).

The LLM-generated data demonstrate strong performance only in replicating basic factual characteristics. In particular, the distribution of current driver status exhibits negligible difference (Cramér’s V = 0.021), indicating that the model can accurately reproduce simple, widely shared population attributes. Similarly, vehicle ownership also shows only moderate divergence from the true levels observed in the reference dataset, suggesting that LLM-generated data capture the aggregate distributions of these basic transportation behaviors reasonably well. However, even in these cases, there are a few notable differences. For instance, the LLM-generated data overestimate two-vehicle households, underestimate one-vehicle households, and consistently underrepresent households with more than two vehicles.

⁴ Vehicle ownership is defined as the number of motorized vehicles available in the household, including cars, trucks, SUVs, vans, and motorcycles, whether owned, leased, or a company vehicle, and excluding electric bicycles and electric scooters.

Table 3: Comparison of Key Variables between the Reference and LLM-generated Datasets with Chi-square and Cramer’s V Statistics

Variable	Reference dataset (%)	LLM-generated dataset (%)	Chi-square statistic	Cramér's V
<i>Driver status</i>				
No	14.0	12.5	7.3	0.021
Yes	86.0	87.5		
<i>Vehicle ownership</i>				
0	8.5	12.7	660.6	0.203
1	40.6	32.4		
2	34.1	46.8		
3	11.1	6.6		
4	4.0	0.3		
5 or more	1.7	1.2		
<i>Mode use frequency</i>				
<i>Private vehicle (drive alone)</i>				
This is not available to me	15.9	12.7	1332.7	0.288
This is available but I never use it	4.2	22.5		
I use it less than one day per month	3.3	1.0		
I use it 1-3 days per month	4.8	2.0		
I use it 1-2 days per week	11.9	7.5		
I use it 3 or more days per week	59.9	54.3		
<i>Public transportation</i>				
This is not available to me	31.2	26.2	1631.3	0.319
This is available but I never use it	36.8	64.0		
I use it less than one day per month	11.8	3.7		
I use it 1-3 days per month	7.7	3.7		
I use it 1-2 days per week	5.9	1.4		
I use it 3 or more days per week	6.6	1.0		
<i>Walk</i>				
This is not available to me	17.1	0.1	5487.1	0.585
This is available but I never use it	23.8	1.4		
I use it less than one day per month	14.7	1.6		
I use it 1-3 days per month	12.1	34.3		
I use it 1-2 days per week	11.3	28.5		
I use it 3 or more days per week	21.0	34.1		
<i>Ride-hailing</i>				
This is not available to me	25.2	3.4	3775.3	0.485
This is available but I never use it	37.2	81.0		
I use it less than one day per month	19.3	13.5		
I use it 1-3 days per month	10.6	2.0		
I use it 1-2 days per week	4.9	0.0		
I use it 3 or more days per week	2.8	0.1		
<i>Telework frequency</i>				
<i>Do all of your work for the day at home</i>				
Never	44.5	77.2	1615.2	0.422
Less than 1 day per month	10.4	13.1		
1-3 days per month	11.5	5.3		
1-2 days per week	15.2	4.3		
3 or more days per week	18.4	0.1		
<i>Package receiving frequency</i>				
<i>Receive packages (other than food or groceries) delivered by large providers (e.g., UPS, FedEx, Amazon)</i>				
Never	6.2	0.6	2727.9	0.412
Less than 1 day per month	26.5	1.5		
1-3 days per month	40.6	66.4		
1-2 days per week	19.6	20.2		
3 or more days per week	7.1	11.3		

Variable	Reference dataset (%)	LLM-generated dataset (%)	Chi-square statistic	Cramér's V
<i>Opinions regarding transportation policies</i>				
<i>The primary goal of transportation investments should be to reduce crashes and improve safety</i>				
Strongly disagree	2.2	0.0	5721.4	0.597
Somewhat disagree	4.9	0.0		
Neither agree nor disagree	21.6	0.3		
Somewhat agree	41.4	12.6		
Strongly agree	29.9	87.1		
<i>We should invest more in public transportation in urban areas</i>				
Strongly disagree	4.4	0.2	3064.9	0.437
Somewhat disagree	5.7	0.1		
Neither agree nor disagree	22.2	11.6		
Somewhat agree	34.8	76.5		
Strongly agree	32.9	11.6		
<i>Congestion pricing (charging drivers at busy times) is unfair</i>				
Strongly disagree	4.5	1.6	6681.2	0.645
Somewhat disagree	7.5	49.4		
Neither agree nor disagree	23.5	40.7		
Somewhat agree	30.4	8.0		
Strongly agree	34.1	0.3		
<i>Intention to use emerging transportation technologies</i>				
<i>Electric vehicles</i>				
I don't know what it is	4.1	0.0	4126.9	0.527
I will not use them	32.2	3.5		
I will use them	48.2	95.5		
I am not sure if I will use them	15.5	1.0		
<i>Autonomous vehicles</i>				
I don't know what it is	11.0	0.0	7463.5	0.722
I will not use them	44.5	12.6		
I will use them	27.9	0.2		
I am not sure if I will use them	16.6	87.2		
<i>Urban air mobility (flying shuttles or air taxis)</i>				
I don't know what it is	23.4	0.0	3460.8	0.582
I will not use them	34.2	42.9		
I will use them	24.5	0.1		
I am not sure if I will use them	17.9	57.0		

In terms of the data related to mode use frequencies (the third-row panel in Table 3), a notable discrepancy appears for driving a private vehicle alone. Hereafter, private vehicle use refers specifically to driving alone. Although only 4.2% of respondents in the reference data report that a private vehicle is available to them but they never use it, Gemini assigns 22.5% of respondents to this category. Correspondingly, the LLM-generated data underrepresent several occasional-use categories, compressing wider private vehicle use patterns into non-use. Similarly, LLM-generated data also underestimate public transportation usage levels, which could lead to underestimation of public transportation demand. Further, relatively strong divergences are found between the distributions for walking frequency, and ride hailing frequency. For instance, the LLM-generated data portray walking as nearly universally available, whereas 17.1% of THA respondents report that it is unavailable. Walking was also portrayed as more frequently used by LLM-generated data than it is in reality. Also, for ride hailing, Gemini selected “this is available to me, but I never use it” for 81% of respondents, compared with 37.2% in the reference dataset.

The results for the patterns of teleworking and package receiving are also informative. Telework patterns in the LLM-generated dataset show a strong divergence from the reference data,

indicating that the LLM does not accurately represent contemporary work arrangements. LLM-generated data heavily concentrate responses in the “never” category (77.2% compared to 44.5% in reference data), while dramatically underrepresenting all levels of regular teleworking. This suggests that the rapid growth and uptake of telework in recent years may not yet have penetrated the LLM’s training data strongly enough to be reflected accurately in its responses (U.S. Census Bureau, 2022). Thus, the results indicate that LLMs may struggle with such emerging or rapidly changing topics. On the contrary, in case of the use of delivery services, the model overestimates usage. While it correctly identifies the most prominent category and the extreme cases, it distorts the overall distribution. This may reflect the opposite problem from telework: delivery services are widely visible in online discourse since they capture target audience through online advertising. So, the LLMs may amplify their salience rather than reproducing their actual aggregate frequency.

The LLM-generated data again exhibit a tendency to over-concentrate responses in specific categories on questions related to opinions about transport policies (see the penultimate row panel of Table 3). For instance, for the question of more investment in public transportation, LLM-generated data place a disproportionately large share of respondents in the “somewhat agree” category, while underrepresenting both strong disagreement and agreement. This indicates that the model captures general support for public transportation investment but compresses the diversity of opinions into a narrow range of moderate agreement. This pattern of accurately identifying the majority response, but then overrepresenting it compared to the reference data is present across many of the behaviors considered here and aligns with the findings of Wu et al. (2025) regarding reduced heterogeneity in LLM responses. Furthermore, for the congestion-pricing question, the LLM fails to reproduce even the overall direction of public opinion. While the majority of respondents in the reference dataset somewhat or strongly agree that congestion pricing is unfair, the LLM-generated responses are concentrated in the “somewhat disagree” and “neither agree nor disagree” categories. Thus, caution is needed if LLM is to be used for studying less commonly studied policies or public opinions, particularly regarding new or evolving transportation topics.

Finally, the aggregate trends for the emerging transportation modes show some of the largest divergences between the datasets. The most consistent difference is that the LLM assumes near-universal familiarity with all three technologies, with almost no agents indicating that they do not know what EV, AV, or UAM are, despite a good number of respondents in the reference dataset indicating that they are unaware of these technologies. Moreover, the LLM-generated data strongly overestimate EV adoption, with nearly all respondents indicating willingness to use them. This pattern is consistent with the inflated expectations surrounding emerging technologies discussed in Section 2.1, where online discourse may present new technologies more optimistically than broader population sentiment. In contrast, for AVs and UAM, Gemini shifts responses toward uncertainty or non-adoption while substantially underestimating adoption intention. As discussed in Section 2.1, this may reflect the opposite issue: for technologies with very limited real-world deployment and scarce first-hand usage narratives, LLM-generated responses may emphasize uncertainty or non-adoption rather than accurately representing broader population sentiment.

Overall, the LLM-generated data can approximate dominant choices in some cases and reasonably reproduce distributions for simple, widely shared characteristics. However, it fails to capture the distributional diversity observed in the reference data, often over-concentrating responses in fewer categories. As a result, even at the aggregate level, caution is needed when relying on LLM-generated data, particularly in the case of less common or rapidly changing topics.

5.2 Ability to Balance Response Consistency and Sensitivity to Item Wordings

To assess whether LLM-generated responses can generate attitudinal profiles and represent these across-related survey items, we undertake a CFA based on the three latent constructs of (1) pro-transit orientation, (2) pro-environmental disposition, and (3) perception of online shopping (in)accessibility. We begin by examining the distributions of the latent-factor indicators in both datasets, presented in Figure 1. Notably, even at this aggregate level, the figure shows a clear and consistent difference between the reference and LLM-generated responses across all attitudinal indicators, mirroring the trends just discussed in Section 5.1. In the reference data, responses are distributed across the full Likert scale, with meaningful shares of disagreement, neutrality, and agreement. In contrast, LLM-generated responses are concentrated in a narrower set of categories, primarily “neither agree nor disagree” and “somewhat agree,” with very limited representation of strong opinions. Overall, Figure 1 suggests that the LLM-generated responses capture the broad direction of some attitudes, such as overall support for public transportation, but fail to represent the strength and variability of these attitudes. Thus, they portray a population that appears more moderate, less critical, and less polarized than the reference population.

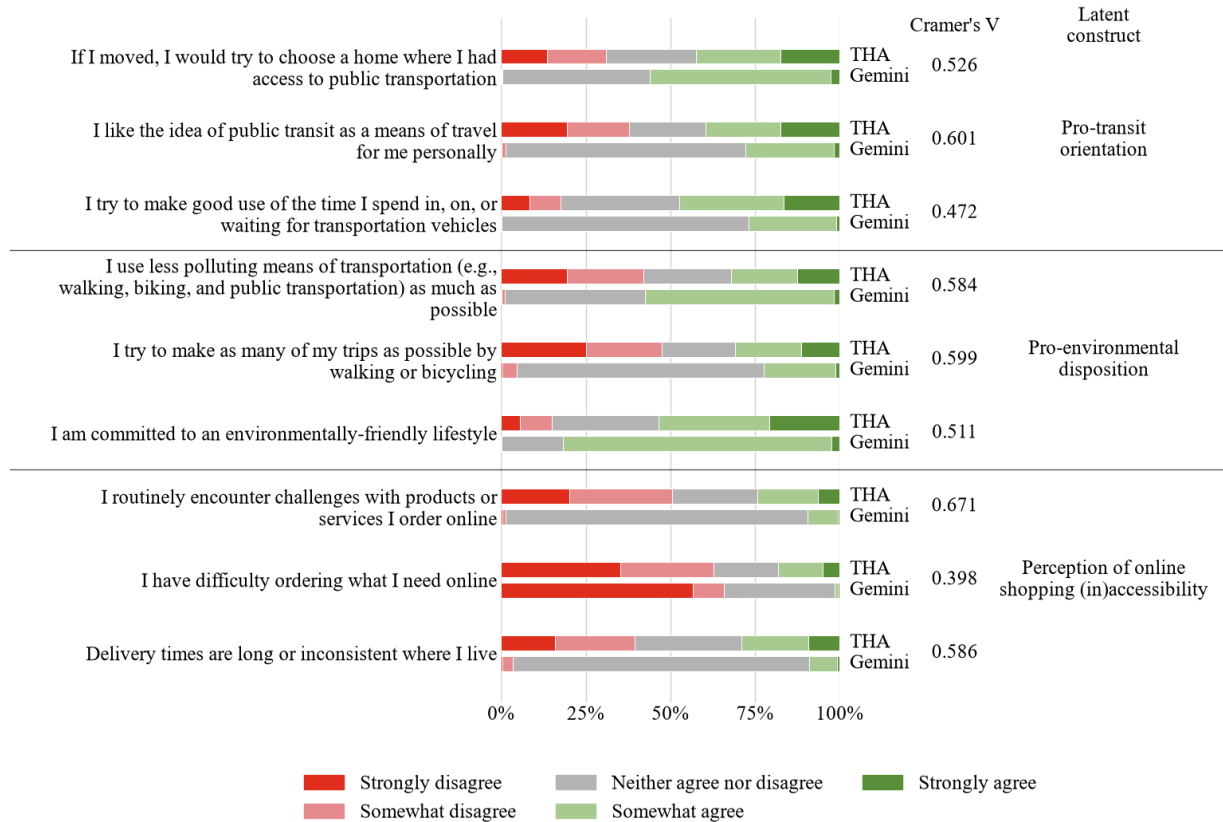
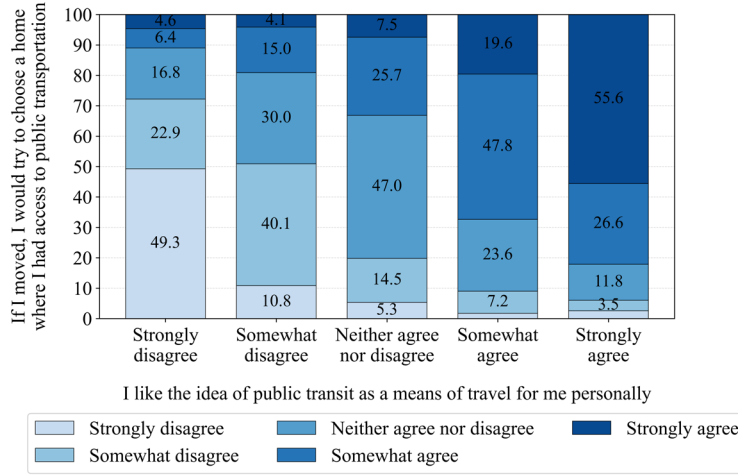
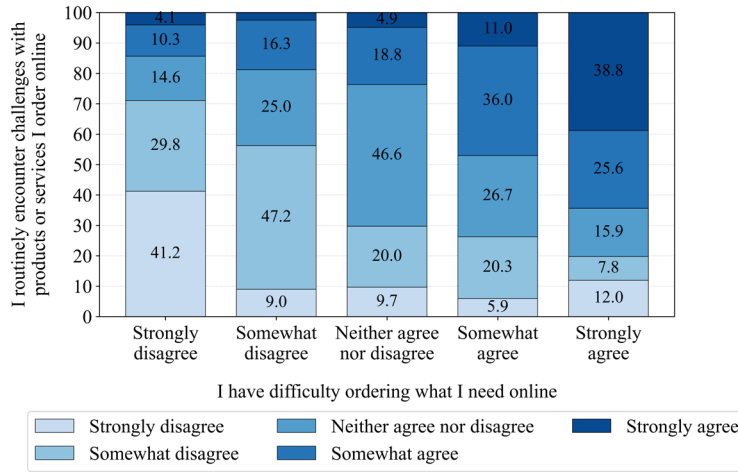


Figure 1: Distribution of Indicators of Latent Constructs across Datasets

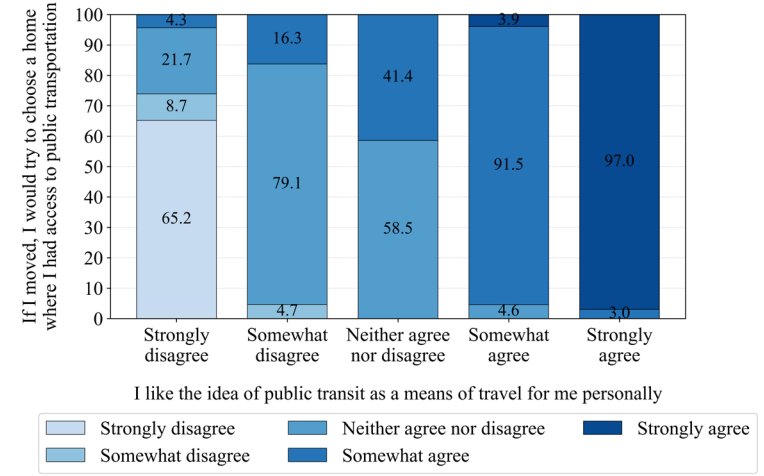
To examine whether the LLM-generated responses maintain consistency across related attitudinal indicators while still preserving the natural variation arising from differences in item wording, Figure 2 presents stacked bar charts for selected pairs of indicators from each of the latent constructs for pro-transit orientation and perception of online shopping (in)accessibility. Within each panel, each bar shows the distribution of responses to the indicator on the vertical axis among respondents selecting a given response to the indicator on the horizontal axis (thus, the sum of the percentages within each stacked bar is 100%).



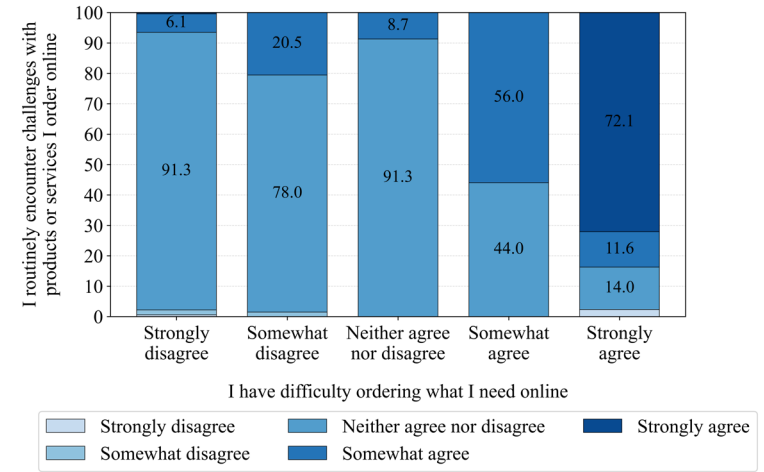
(a) Cross-tabulation of reference dataset responses to two indicators of pro-transit orientation



(c) Cross-tabulation of reference responses to two indicators of perception of online shopping (in)accessibility



(b) Cross-tabulation of LLM-generated responses to two indicators of pro-transit orientation



(d) Cross-tabulation of LLM-generated responses to two indicators of perception of online shopping (in)accessibility

Figure 2: Cross-tabulations of Responses to Attitudinal Indicators across Datasets

Panel (a) shows a clear ordered relationship between the two pro-transit indicators in the reference data. The modal (used here in a statistical sense) response to the statement about choosing a home with access to public transportation corresponds to the response given to the statement about liking public transportation for every horizontal-axis category. At the same time, substantial proportions select neighboring response categories, while progressively fewer respondents select categories that are further away on the response scale, producing a gradual transition across the bars. This pattern indicates that respondents express generally consistent pro-transit orientations without mechanically providing identical responses to the two differently worded indicators. A similar but somewhat more diffuse ordered pattern characterizes the two online-shopping indicators in panel (c): the modal response category of the vertical-axis corresponds to the horizontal-axis category for all the five stacked bars, but the responses are more broadly distributed across neighboring categories. For example, among those who strongly agree that they have difficulty ordering online, 38.8% strongly agree and 25.6% somewhat agree that they routinely encounter challenges with ordered products or services. This weaker concentration is behaviorally plausible because the items capture related but distinct experiences, and the broader dispersion indicates that the responses preserve a common underlying relationship while reflecting differences in item meaning.

The LLM-generated responses preserve some of the ordinal relationships between the pro-transit indicators but fail to reproduce the smoother variation observed in the reference data. Panel (b) shows an extremely concentrated relationship at the right end of the figure: 97.0% of those who strongly agree with liking public transportation provide the corresponding strongly agree response for choosing a home with public transportation access, while 91.5% of those who somewhat agree also provide the corresponding somewhat agree response. Such near-deterministic correspondence suggests that Gemini treats the two indicators as almost interchangeable when respondents express positive attitudes. The synthetic responses show some variation in the first three response categories, corresponding to those who disagree or are neutral about liking public transportation as a means of travel, but this variation is much more skewed than in reference data, with responses heavily concentrated in a single category within each bar. The discrepancy is more pronounced for the online-shopping (in)accessibility indicators in panel (d). For the first three response categories, Gemini overwhelmingly assigns a neutral response to the statement about routinely encountering challenges with online purchases. Responses become more differentiated only among those who agree that they have difficulty ordering what they need online, but the gradual dispersion across neighboring categories is still absent. This pattern suggests that Gemini may be overly sensitive to the semantic distinction between difficulty placing an online order and encountering challenges with the products or services ordered online, limiting its ability to represent their shared underlying construct consistently.

Taken together, the results indicate that the LLM can preserve the broad ordinal relationships between related indicators but does not reproduce the balance between attitudinal consistency and sensitivity to differences in wording observed in the reference responses. These issues are examined further by comparing the structural components of the latent variable models estimated separately using the reference and the LLM-generated data. The latent variable model estimation results are reported in Table B.1 and B.2 in Appendix B. Because the latent variables are scaled separately in the two models, the raw structural coefficients are not directly comparable. Therefore, the coefficients are normalized by the standard deviation of the corresponding latent construct to obtain scale-normalized structural effects. These scale-normalized/standardized

effects, reported in Table 4, are used to compare the relationships between respondent characteristics and latent attitudes across the two datasets.

Table 4: Comparison of Standardized Structural Effects in Reference- and LLM-Based Latent Construct Models

Latent variable structural equation model	Standardized structural effects					
	Pro-transit orientation		Pro-environmental disposition		Perception of online shopping (in)accessibility	
	THA	Gemini	THA	Gemini	THA	Gemini
Age (<25 years old)						
25-34 years old	0.085	-0.205	0.095	-0.167	-	0.241
35-44 years old	0.074	-0.310	0.093	-0.248	-	0.429
45-54 years old	-	-0.383	-	-0.322	-0.172	0.525
55-64 years old	-0.142	-0.426	-0.129	-0.428	-0.297	0.504
65 years old and above	-0.233	-0.470	-0.271	-0.479	-0.392	0.187
Education level (high school or less)						
Some college	-	0.239	-0.045	0.257	-0.085	-0.352
Bachelor or higher	0.129	0.410	0.077	0.474	-0.148	-0.493
Employment status (not employed)						
Employed full time	-	-0.112	0.038	-0.131	0.081	-0.399
Employed part time	-	-	-	0.058	0.085	-0.281
Race (not White or Black)						
White	-0.223	-0.264	-0.192	-0.330	-0.120	-0.094
Black	0.100	0.187	-	-	-0.124	0.185
Number of children in the household	-0.031	-	-0.016	-0.016	-	0.069
Number of adults in the household	0.029	0.036	-	0.014	-	-
Annual household income (< \$50,000)						
\$50,000-\$99,999	-0.050	0.105	-0.076	0.170	-0.169	-0.591
\$100,000 and above	-	0.335	-	0.405	-0.147	-0.880
Housing tenure type (own)						
Rent	0.081	0.417	-	0.222	-0.069	0.158
Other tenure type	-	0.110	-	-	-	0.525
Housing unit type (other housing type)						
Single family housing	-	-0.198	-	-0.081	-	-0.212
Apartment	0.250	0.515	0.218	0.307	-	-0.184

The structural effects in Table 4 indicate that the attitudes generated in the LLM dataset are more strongly associated (higher magnitudes of the standardized coefficients) with respondents' individual- and household-level demographic characteristics than those observed in the reference dataset. This results in relationships that are more monotonic than those in the reference data. For example, the LLM-generated data show increasingly negative age effects on pro-transit orientation and pro-environmental disposition constructs across all age groups. In the reference data, however, the relationships are less uniform, with positive effects among some younger and middle-aged groups and negative effects emerging mainly among older respondents. The discrepancies are even more pronounced for perception of online shopping (in)accessibility. The reference data show negative effects only among respondents aged 45 and above, which may be because older respondents order online less frequently and hence do not strongly identify with the perceived online-shopping (in)accessibility indicators. However, the LLM-generated data show positive effects for every age group, indicating a fundamentally different relationship between age and perceived online-shopping (in)accessibilities. As discussed in Section 4.2, this third construct is

more experience-based, which, in particular, the LLM-generated data are not able to relate appropriately with respondent age or other demographics. Further, the LLM-generated model, across all latent constructs, also distorts the effects of educational attainment, employment status, race, household composition, income, housing tenure and unit type, with many effects being exaggerated, introduced where none are observed in reference data, or reversed. For example, the reference data indicate that households with annual incomes of \$50,000–\$99,999 have less favorable pro-transit and pro-environmental orientations than lower-income households, while the LLM-generated data suggest the opposite.

Overall, the attitudes generated appear to be more strongly and systematically tied to observable respondent characteristics than those in the reference data. This is also reflected in the number of statistically significant structural effects: across the three latent constructs, only five effects of observed respondent characteristics turned up to be statistically insignificant in the LLM-generated model, compared with twenty-two in the reference model. This overreliance on individual- and household-level socio-demographic attributes may indicate that the model is not generating sufficiently rich and nuanced intrinsic attitudinal heterogeneity that can be carried through consistently across related survey items. This could be a contributing reason behind the compressed response variation observed in the indicator distributions and raises concerns about using LLM-generated data to represent latent preferences, beliefs, or attitudes in downstream behavioral models.

5.3 Ability to Distinguish Preferences and Behaviors among Population Subgroups

As discussed in Section 4.3, for our disaggregate analysis to examine the ability of LLM-based data to distinguish preferences and behaviors among population subgroups, we consider eight outcomes from the 14 outcomes considered in the aggregate trends analysis. These are vehicle ownership; mode use frequencies of private vehicle, public transportation, and walking (including walking to public transportation stops); teleworking and package receiving frequencies; and EV and AV adoption intentions. The full model estimation results obtained from the reference THA dataset and the LLM-generated data are available from the authors. However, while the model estimation results reveal key differences between the datasets, the coefficient estimates themselves do not indicate the magnitude of the response differences between the estimates from each dataset. Thus, to better illustrate how policy-relevant estimates would change based on the predictions made using each of the two models, we compute average treatment effects (ATEs) for exogenous variables in each model. To keep the discussion concise and streamlined, Table 5 reports the THA- and LLM-based ATEs, along with their relative deviations, for one selected category of interest and for select exogenous variables for each outcome. Specifically, the ATEs are calculated for owning three or more vehicles; being in the highest private vehicle use and walking frequency categories; using public transportation at all; teleworking and receiving packages at least once per week; and expressing an intention to adopt EVs and AVs (the ATEs for all response levels of each of the eight outcomes and for each exogenous variable are available from the authors).

Table 5: Average Treatment Effect across Models

	Vehicle ownership: 3 or more			Private vehicle use frequency: 3+ days/week			Public transportation use frequency: Uses it			Walking frequency: 3+ days/week		
	ATE		D (%)	ATE		D (%)	ATE		D (%)	ATE		D (%)
	THA	Gemini		THA	Gemini		THA	Gemini		THA	Gemini	
Individual characteristics												
Age (<25 years old)												
25-34 years old	-0.059	0.000	100.0	0.000	0.035	100.0	-0.042	-0.029	31.0	-0.058	-0.068	-17.2
35-44 years old	-0.095	0.023	124.2	0.000	0.044	100.0	-0.046	-0.041	10.9	-0.052	-0.121	-132.7
45-54 years old	-0.091	0.035	138.5	0.000	0.065	100.0	-0.081	-0.043	46.9	-0.083	-0.135	-62.7
55-64 years old	-0.056	0.016	128.6	0.043	0.045	4.7	-0.119	-0.044	63.0	-0.125	-0.168	-34.4
65 years old and above	-0.073	0.027	137.0	0.053	0.056	5.7	-0.170	-0.037	78.2	-0.162	-0.166	-2.5
Gender (male)												
Female	0.016	-0.020	-225.0	-0.023	0.000	100.0	-0.075	0.000	100.0	-0.072	0.087	220.8
Education level (high school or less)												
Some college	0.048	0.034	-29.2	0.051	0.020	-60.8	0.000	0.020	100.0	0.000	0.043	100.0
Bachelor's degree or higher	0.029	0.025	-13.8	0.088	0.000	-100.0	0.019	0.036	89.5	0.019	0.089	368.4
Employment status (not employed)												
Employed full time	0.057	0.102	78.9	0.214	0.331	54.7	0.033	0.000	-100.0	0.000	0.000	0.0
Employed part time	0.056	0.087	55.4	0.118	0.169	43.2	0.027	-0.020	-174.1	0.024	0.055	129.2
Race (not White or Black)												
White	0.000	0.033	100.0	0.063	0.040	-36.5	-0.045	-0.028	37.8	0.000	0.000	0.0
Black	-0.057	-0.026	54.4	0.000	0.000	0.0	0.000	0.028	100.0	0.000	0.000	0.0
Ethnicity (not Hispanic)												
Hispanic	0.000	0.000	0.0	0.000	0.000	0.0	0.000	0.011	100.0	0.000	0.000	0.0
Household characteristics												
No. of children in the household (zero)												
One	0.060	0.062	3.2	0.019	0.032	68.4	-0.009	0.000	100.0	0.000	0.030	100.0
Annual household income (<\$50,000)												
\$50,000-\$99,999	0.166	0.299	80.1	0.108	0.150	38.9	0.000	0.000	0.0	0.000	0.000	0.0
\$100,000 and above	0.240	0.430	79.2	0.145	0.200	37.9	0.025	0.011	-56.0	0.000	0.000	0.0
Housing tenure type (own)												
Rent	-0.115	-0.311	-170.4	-0.071	-0.176	-147.9	0.043	0.075	74.4	0.020	0.055	175.0
Other tenure type	-0.066	-0.225	-240.9	-0.155	-0.166	-7.1	0.063	0.027	-57.1	0.036	-0.031	-186.1
Residential neighborhood characteristics												
Road network density	-0.034	0.000	100.0	0.000	0.000	0.0	0.135	0.029	-78.5	0.057	0.037	-35.1

Table 5 (Continued)

	Teleworking frequency: At least once per week			Package receiving frequency: At least once per week			EV adoption intention: I will use it			AV adoption intention: I will use it		
	ATE THA	Gemini	D (%)	ATE THA	Gemini	D (%)	ATE THA	Gemini	D (%)	ATE THA	Gemini	D (%)
Individual characteristics												
Age (<25 years old)												
25-34 years old	0.000	0.010	100.0	0.000	0.000	0.0	0.000	0.000	0.0	0.041	0.000	-100.0
35-44 years old	0.000	0.000	0.0	-0.030	0.000	100.0	0.000	0.000	0.0	0.031	0.000	-100.0
45-54 years old	-0.047	0.000	100.0	-0.051	0.000	100.0	-0.068	0.000	100.0	0.000	-0.001	-100.0
55-64 years old	0.000	0.000	0.0	-0.085	-0.027	68.2	-0.151	-0.016	89.4	-0.086	-0.002	97.7
65 years old and above	-0.052	0.000	100.0	-0.087	-0.044	49.4	-0.243	0.000	100.0	-0.166	-0.002	98.8
Gender (male)												
Female	0.000	0.000	0.0	0.000	0.027	100.0	-0.104	0.000	100.0	-0.115	0.000	100.0
Education level (high school or less)												
Some college	0.036	0.013	-63.9	0.032	0.037	15.6	0.058	0.010	-82.8	0.000	0.001	100.0
Bachelor's degree or higher	0.145	0.020	-86.2	0.049	0.046	-6.1	0.133	0.017	-87.2	0.042	0.001	-97.6
Employment status (not employed)												
Employed full time	-0.065	-0.202	-210.8	0.035	0.307	777.1	0.000	0.000	0.0	0.000	0.001	100.0
Employed part time	0.000	0.000	0.0	0.000	0.139	100.0	0.000	0.000	0.0	0.000	0.001	100.0
Race (not White or Black)												
White	-0.056	0.000	100.0	0.024	0.036	50.0	-0.039	-0.012	69.2	-0.025	0.000	100.0
Black	0.000	0.000	0.0	0.000	0.000	0.0	0.074	0.000	-100.0	0.032	0.000	-100.0
Ethnicity (not Hispanic)												
Hispanic	0.000	0.000	0.0	-0.017	0.027	258.8	0.041	0.000	-100.0	0.030	0.000	-100.0
Household characteristics												
No. of children in the household (zero)												
One	0.000	0.000	0.0	0.028	0.048	71.4	0.019	0.000	-100.0	0.016	0.000	-100.0
Annual household income (<\$50,000)												
\$50,000-\$99,999	0.000	0.000		0.086	0.128	48.8	0.000	0.032	100.0	0.000	0.001	100.0
\$100,000 and above	0.077	0.016	-79.2	0.181	0.289	59.7	0.047	0.054	14.9	0.044	0.002	-95.5
Housing tenure type (own)												
Rent	-0.050	-0.007	86.0	-0.037	-0.081	-118.9	0.000	-0.021	-100.0	0.000	-0.001	-100.0
Other tenure type	-0.053	-0.015	71.7	-0.095	-0.133	-40.0	0.000	-0.026	-100.0	0.000	-0.001	-100.0
Residential neighborhood characteristics												
Road network density	0.038	0.000	-100.0	0.000	0.000	0.0	0.072	0.000	-100.0	0.038	0.000	-100.0

To illustrate the interpretation of the ATEs and their relative deviations, consider the vehicle ownership results for individuals aged 25-34. In the first numeric row under the vehicle ownership outcome in Table 5 for the THA-based model, the value of “-0.059” indicates that in a population of 1000 adults aged 25-34, approximately 59 fewer would own three or more vehicles compared to a population of 1000 adults aged 18-24. The corresponding value for the LLM-based model is “0.000,” indicating that the LLM-based model does not predict any difference in the vehicle ownership rate among these two groups. Finally, the percent deviation between the two ATE values is 100% (shown as D% in Table 5), indicating that the LLM model completely misses the treatment effect in the observed data. Also, because THA-based ATE serves as the denominator in the relative deviation calculation, the deviation is not defined when the THA-based ATE is equal to zero. Therefore, we assign a value of 0% when both ATE values are zero, indicating alignment between the two models. When the THA-based ATE is zero and the LLM-based ATE is non-zero, we assign a value of $\pm 100\%$, with the sign indicating the direction of the LLM-based effect, to indicate that the effect is present only in the LLM-based model. The remainder of this section discusses these results selectively, focusing on the relative levels of alignment and discrepancies between the reference and LLM-based models ATEs than on the specific empirical interpretations.

Overall, while there are important areas of alignment between the models based on the real THA data and the LLM-generated data, there are significant discrepancies. For instance, consider the age-related effects across the models. In some cases, the LLM-generated data capture the broad direction of some age-related patterns, particularly for outcomes with relatively clear and almost monotonic trends, such as public transportation use, walking, and package receiving frequency. Even for these outcomes, however, some effects are missing, particularly in the package receiving model. For outcomes with less monotonic age patterns, such as vehicle ownership and teleworking frequency, Gemini performs poorly, predicting effects in the opposite direction for vehicle ownership and literally no age-based heterogeneity for teleworking. For private vehicle use frequency, the reference model indicates a greater propensity for use among respondents aged 55 and above, while the LLM-based model predicts significant differences for nearly every age group from the base group. More importantly, for the EV and AV adoption intention models, the reference data show clear age-related declines in adoption intention, but Gemini does not capture these trends accurately and predicts mostly zero or negligible age effects. This suggests that Gemini performs particularly poorly in capturing age-based heterogeneity in preferences toward emerging transportation technologies. Thus, the LLM’s performance in capturing age effects is inconsistent across outcomes, reinforcing the need to assess the suitability of LLM-generated data separately for each behavioral domain.

Gender effects show alarmingly weak alignment, as the LLM-generated data fail to reproduce the ATEs correctly across nearly all outcomes. This may reflect two related issues. First, as discussed in Section 2.1, uneven gender representation in web-based training corpora may limit the model’s ability to learn gender-based behavioral differences that are observable in systematically collected survey data. Second, it reflects the concern about algorithmic fidelity raised in Section 2.3: even when gender is provided as a conditioning variable, the LLM does not preserve the gender-based behavioral differences observed in the THA models.

Compared to gender, the LLM-generated data show relatively better alignment for educational attainment and employment status, generally reproducing the direction of the effects observed in the reference data. The effect of part-time employment on public transportation use frequency is the only case in which the direction is reversed. However, the LLM-based models also miss some effects observed in the reference data (effects of a bachelor’s degree or higher on

frequent solo driving and of full-time employment on public transportation use) and produce some additional effects that are not statistically significant in the reference models (effects of some college education on public transportation use, walking, and AV adoption intention; of part-time employment on weekly package receiving and AV adoption intention; and of full-time employment on AV adoption intention). Aside from these, the relatively better alignment may reflect clearer differences in commuting patterns, work arrangements, technology use, and lifestyles across these groups, or stronger representation of these distinctions in the training data.

In case of race and ethnicity, the LLM-generated data exhibit a more systematic pattern of mobility differences than the reference data. Across the vehicle ownership, private vehicle and public transportation use frequency models, the LLM-generated ATEs portray White respondents as more automobile-oriented and Black and Hispanic respondents as more transit-oriented. Specifically, the LLM-generated models indicate that White respondents are more likely to own three or more vehicles and to use a private vehicle frequently, but less likely to use public transportation. They also indicate that Black respondents are less likely to own three or more vehicles and that both Black and Hispanic respondents are more likely to use public transportation. In the reference models, however, only a subset of these relationships is statistically significant. Thus, the synthetic data provide broader statistical support for a race/ethnicity pattern, while the reference data show a more mixed picture, even though some of the relationships that are significant in the THA data have larger effect magnitudes. Interestingly, the broader trend produced by the LLM is not implausible, as prior research shows that in the U.S., White households are generally more automobile-oriented, while Black and Hispanic populations rely more heavily on public transportation (Brown, 2017; Blumenberg et al., 2020; Wang and Renne, 2023). The concern, however, is that the LLM reproduces these generalized race-mobility associations more consistently than they appear in the reference data, potentially imposing a simplified racial pattern on outcomes for which the survey shows more limited or nuanced differences. This aligns with Bisbee et al. (2024), who show that even if LLM-generated responses appear plausible at an aggregate level, it may be substantially distorted when disaggregated by race.

For household-level attributes, the LLM-generated data show mixed performance. For example, although the effect of a child in the household is broadly consistent with reference data for several outcomes, some notable discrepancies remain. In particular, the reference data indicate that public transportation use decreases as the number of children in the household increases from zero to one, consistent with studies showing that children can increase household car reliance through escorting, scheduling, and household-task coordination demands (Oakil et al., 2014; Klein and Smart, 2019). In contrast, the LLM-based model identifies no significant effect of having a child on public transportation use, thereby missing a key effect with important social and policy implications. The LLM-based models also introduce a significant effect of having a child on walking where none is observed in the reference data and fail to reproduce the effects on EV and AV adoption intentions. These discrepancies may arise because the mobility implications of having a child are highly context-dependent, varying with the child's age, school or childcare arrangements, parental work schedules, vehicle availability, and the allocation of escorting responsibilities.

The results are more consistent for household income and housing tenure, for which the LLM-generated data generally follow the patterns observed in the THA data. Among the outcomes unrelated to emerging technologies, the only mismatch for these variables is the reversal of the positive effect of other housing tenure types on frequent walking. This relatively better alignment may be because household income and housing tenure are broad indicators of household resources

and socioeconomic position, with well-documented associations with vehicle ownership, mobility, and mode use (Sun and Zacharias, 2021; Wang and Renne, 2023). These broad and frequently documented relationships may therefore be more strongly represented in the LLM’s training data.

Finally, among the built-environment variables, road network density showed the greatest alignment in the estimated probit models and was therefore included in the ATE comparisons in Table 5. Yet, even for this variable, the LLM-based model misses or underestimates several significant ATEs with important policy implications. Specifically, Gemini misses the negative effect of road network density on vehicle ownership, underestimates its positive effects on public transportation use and frequent walking, and fails to reproduce its positive effect on teleworking.

Overall, the subgroup-level results indicate that the LLM-generated data do not perform uniformly across all behavioral domains; rather, they reproduce broad and commonly documented socioeconomic patterns more reliably than complex or context-dependent behavioral relationships. The largest discrepancies arise for effects that are nuanced, non-monotonic, or mediated by unobserved constraints. These findings reinforce that LLM-generated survey data should not be evaluated only using aggregate distributions, because even when aggregate patterns appear plausible, the conditional relationships needed for subgroup analysis and policy evaluation may be substantially distorted.

6. POLICY IMPLICATIONS

The ATE results discussed above show how reliance on LLM-generated data could affect policy interpretation. Four key implications emerge from this analysis, which are discussed in the following subsections.

6.1 Aggregate Alignment Masking Subgroup-level Distortions

The deviations between the ATE values show that even when aggregate trends are reasonably replicated by LLM-generated data, important subgroup heterogeneity can still be masked. This is most evident for vehicle ownership, private vehicle and public transportation use frequency. As discussed in Section 5.1, these outcomes were among the relatively better-performing cases in the aggregate comparison. However, the ATE results show that this aggregate alignment does not translate into reliable subgroup-level inference. In the vehicle ownership model, the LLM-generated data exaggerate the effects of employment, income, and housing tenure. Consequently, high vehicle ownership appears more strongly concentrated among employed individuals, and higher-income and owner-occupied households, and much less likely among renters and households in other tenure arrangements. Similar patterns appear in private vehicle and public transportation use models. For private vehicle use frequency, the LLM-generated data overpredict the positive effects of employment and income and substantially exaggerate the negative effect of renting. In the public transportation use model, the LLM-generated data miss the positive association between full-time employment and public transportation use observed in the reference data, reverse the positive effect of part-time employment, and overpredict public transportation use among renters. The first two models also underpredict the positive effect of educational attainment higher than high school on high vehicle ownership and frequent private vehicle use, while the public transportation model introduces a positive effect for some college where none is observed in the reference model and amplifies the positive effect of a bachelor’s degree or higher. Taken together, these patterns suggest that the LLM-generated data may encode broad and familiar narratives about travel behavior, such as employed individuals and those from higher-income

households being more auto-oriented, and renters being more transit-oriented. However, these narratives do not fully reflect the complex behavioral structure observed in the reference data.

From a policy perspective, this matters because such distortions could lead to misleading conclusions about where auto dependence and transit dependence are concentrated. If LLM-generated data were used for forecasting or policy evaluation, they could overstate the strength of the relationship between employment, income, and frequent solo driving, potentially biasing demand or revenue forecasts for congestion pricing and workplace parking policies. Also, they could make transit usage appear too narrowly concentrated among renters, reducing attention to transit needs among less educated, employed or higher-income travelers. Thus, even when aggregate distributions appear reasonably aligned, subgroup-level validation remains essential before LLM-generated data can be used for targeted policy design.

6.2 Misrepresentation of Mobility Constraints and Behaviors

A second policy concern is that the LLM-generated data can misrepresent subgroup-level variations in mobility constraints and behaviors, which are central to transportation planning. For instance, in the aggregate comparisons in Section 5.1, the LLM-generated data portrayed walking as almost universally available and more frequent than in reality. These results are even more pronounced for specific demographic groups, as highlighted by the ATE results. Specifically, the decline in walking with age is overstated in the LLM-generated data, while the negative gender effect on walking observed in the reference data is reversed. If used for policy analysis, these patterns could make walking appear more widespread and less constrained than it is in reality, while masking possible accessibility- and gender-related barriers. At the same time, overstating the age-related decline in walking could exaggerate age-based barriers and divert attention from other groups that may also require pedestrian facility improvements.

In contrast to walking, Section 5.1 shows that the LLM-generated data underrepresent regular teleworking. The ATE results reveal that this underrepresentation is accompanied by weak subgroup differentiation. Although both datasets show that higher education and higher income are associated with a greater probability of teleworking at least once per week, these effects are strongly underpredicted in the LLM-generated data. The LLM-generated data also miss the positive effect of road network density, which in the reference data may reflect more urbanized settings where larger employers and congestion-related commute burdens make remote work more attractive. By missing this, the LLM-generated data understate the spatial effect of teleworking. This could weaken the understanding of how telework affects commute demand and peak-period traffic, while also leading planners to overlook where telework could meaningfully reduce commuting pressure or reshape travel patterns.

6.3 Challenges in Representing Emerging Modes

A third concern is the limited ability of the LLM-generated data to address emerging technologies. For EVs, the LLM-generated data assume near-universal awareness and almost universal adoption intention, which can mask the practical, financial, and informational barriers that shape EV adoption. In contrast, for AVs, the LLM-generated data assume near-universal awareness but concentrate responses in uncertainty, which may overstate the barriers to adoption. Thus, in both cases, the LLM-generated data provide a distorted view of the constraints surrounding emerging mode adoption. These findings also connect to the misrepresentation of mobility constraints discussed in Section 6.2.

Moreover, the reference data show lower EV adoption intention among older adults and female respondents, but these effects are strongly underpredicted or missed in the LLM-generated data. For older adults, this may obscure the need for more age-friendly EV design and support, as they may be more cautious toward newer technologies or may require simpler user interfaces, easier charging systems, and clearer visual guidance. For female respondents, the reference data prompt the need to investigate gender-specific concerns, which could be related to perceived loss of control, safety concerns (especially if escorting children), and household travel responsibilities. By missing these effects, the LLM-generated data could erase subgroup-specific barriers that may be important for EV outreach, design, and infrastructure planning. The concern is even greater for AV adoption intention. In the reference data, AV adoption intention varies meaningfully by age, gender, education, race and ethnicity, household composition, income, road network density, and census division. In the LLM-generated data, however, most of these effects are reduced to zero or near zero, largely erasing this heterogeneity. Consequently, the LLM-generated data provide little guidance regarding which population groups are more receptive to AVs and which may require targeted outreach or policy interventions.

These findings support the discussion in Section 2.1 on emerging technologies and their representation in LLM training corpora. EVs and AVs are both emerging transportation technologies, but web-based training corpora may capture them at different stages of maturity. By the time of data generation, EVs had already become relatively prominent in public discourse, and early-adopter experiences and positive narratives may have been disproportionately represented online. Consequently, the LLM may overstate general EV readiness and mask barriers faced by specific groups. Compared with EVs, AVs remain less familiar in everyday use and may therefore be represented more strongly through uncertainty, safety concerns, and skepticism about readiness. This may explain why the LLM-generated data suggest little AV adoption and provide almost no useful subgroup information. This erasure of AV effects is especially consequential because AVs are still in an early phase of deployment, making subgroup-specific public responses important for planning, communication, and deployment strategies. Therefore, the weak performance of the LLM-generated data-based AV model raises concerns about relying on LLM-generated data for policy analysis in emerging technology contexts where real-world experience remains limited.

6.4 Subgroup Fidelity across Categories

As already noted, the LLM-generated data perform particularly poorly for the emerging technology outcomes. Still, even across the remaining outcomes, the results show that the LLM does not reproduce all subgroup relationships equally well.

Income is the most consistently captured explanatory variable in terms of significance and directional consistency across the models, although its effects are often exaggerated. Moreover, as discussed in Section 6.1, the effects of other factors that also play important roles in transportation planning and forecasting, such as educational attainment, employment status, and housing tenure type, appear to be distorted in the LLM-generated data. Gender effects are particularly poorly reproduced, with most relationships either missed or reversed. Thus, even where income effects are directionally correct, the distortion of these other variables leaves planners with an incomplete understanding of what drives the outcomes and limited leverage for policy intervention. For instance, the fact that full-time employed individuals have a greater propensity for public transportation use has important planning implications that are missed in the LLM-generated data-based results. These results point to important ways to reduce commute-based demands among workers who may be more willing to use services, which may be an effective way to reduce peak-

period demands to high-density employment centers. Thus, missing these results using the LLM-generated data would reduce the effectiveness of policy interventions targeting demand reductions for peak-period commutes, because the potential for shifting demand to public transportation would be underestimated. Similarly, the reference data show that road network density has a substantial positive effect on public transportation use. However, the LLM-generated data suggest only a minimal effect. If planners relied on these results, they might deprioritize infill development and street connectivity in transit-oriented development corridors, thereby weakening a core urban-design strategy for supporting public transportation use.

Adding to this challenge, there is no clear systematic pattern to these distortions. While it does appear that the LLM responds to the input demographic and geographic profiles (leading to significant effects based on these factors), they often do not match the reference data or expected theoretical relationships well. Some of the effects align strongly with stereotypical behaviors attributed to specific demographic groups, rather than teasing out the rich mixed and nuanced responses across demographic groups that exist in reality.

Overall, the study shows very mixed performance of LLM based on different explanatory variables. Because of the significant variation in the performance across the variables and outcomes, there is little evidence suggesting that the estimates can be trusted without further validation, which defeats the purpose of using synthetic data to extend knowledge into data-scarce contexts. Furthermore, the distortion of magnitudes makes even directionally correct effects unreliable for identifying where interventions will be most effective and how much behavior change to expect. Thus, the results indicate that the LLM-generated data are likely to be unsuitable for many of the transportation planning applications that are reliant on differentiation of behaviors and subgroup-level inference.

7. CONCLUSION

Survey data have long been a cornerstone of empirical social research, providing an important avenue to systematically measure how individuals think, behave, and interact within complex social systems. Such data serve as a critical component of transportation planning, providing the empirical foundation for understanding how people move and why they make the choices they do, enabling forecasts to support travel demand management and policymaking. As LLMs become progressively more capable of producing structured, human-like responses, the possibility that they could play a role in these processes becomes increasingly appealing. Still, given widespread uncertainties about the capabilities of these models, their ability to accurately and reliably represent human behaviors and preferences, and the rapidly changing nature of transportation behaviors and decision making that need to be well represented, these models need to be studied carefully. This study examined whether LLM-generated survey responses can reproduce the distributional, attitudinal, and subgroup-level structure of human responses in a broad transportation survey setting. Using the Transportation Heartbeat of America survey as the reference dataset, we generated one synthetic response for each human respondent using Gemini 2.5 Pro, conditioned on individual- and household-level demographic, and geographic characteristics. We then evaluated the LLM-generated data across three dimensions: the ability to identify overall trends in attitudes, preferences, and behaviors; the ability to consistently maintain and represent realistic attitudinal profiles across the survey; and the ability to distinguish preferences and behaviors across population subgroups.

The results show that the LLM-generated data performed reasonably well in identifying population-level trends for some simple outcomes, such as driver status and vehicle ownership.

This performance, however, did not extend to more behaviorally complex outcomes. Instead, the LLM-generated responses showed compressed variation, overstated dominant or moderate categories, and reflected simplified narratives about travel behavior, all of which can lead to systematic misrepresentations of mobility behaviors if LLM-generated data are used for policy analysis. Besides, cross-tabulations of attitudinal indicators showed that the LLM used in this paper either treats similarly worded indicators as nearly interchangeable or becomes overly sensitive to wording differences that human respondents appear to reconcile through a shared underlying construct. Furthermore, LLM-generated attitudes were found to be more strongly and systematically tied to observable demographic and household characteristics than those in the reference data, suggesting an overreliance on profile attributes and insufficient representation of unobserved attitudinal heterogeneity. Together, these findings raise concerns about using LLM-generated survey data to represent latent preferences, beliefs, or attitudes in downstream behavioral models. Finally, the analysis of population subgroup heterogeneity showed that the LLM-generated data did not consistently reproduce the socio-demographic and built-environment relationships observed in the reference data. The findings also suggest that these errors were not random; rather they reflect systematically distorted mobility patterns that could translate into policy-relevant misinterpretations.

Overall, the results indicate that conditioning an LLM on observable individual- and household-level socio-demographic, and geographic characteristics, is not sufficient to ensure algorithmic fidelity in transportation survey data generation. The model may recognize that these characteristics matter, but it does not reliably reproduce how they matter across different travel behavior outcomes. The current results suggest that significant caution should be employed when using LLMs for the generation of synthetic samples in transportation context. While the raw model outputs may appear human-like and aggregate assessments may even appear theoretically reasonable, current LLM-based models exhibit significant limitations and may systematically bias resulting estimates and analyses. Therefore, we recommend that such models and the synthetic samples they produce should not yet be treated as substitutes for empirically collected human survey data in travel behavior research without careful validation against ground truth. These findings do not imply that LLM-generated data have no role in transportation research. These models are continuing to evolve rapidly, and their capabilities are likely to continue to improve in the coming years. They may indeed be capable of providing important insights, particularly regarding behaviors that are more stable over long time-horizons and are broadly generalizable. Further, responses generated with current state-of-the-art LLMs may be useful for survey pretesting, identifying possible response patterns, developing hypotheses, or exploring how transportation topics are represented in online discourse.

There are several additional avenues for future research. First, this study evaluates one LLM, one prompting framework, and one nation-wide transportation survey. Future work should examine whether the observed patterns persist across different models, prompt designs, model settings, and survey contexts. Such comparisons would be particularly valuable given the topic and knowledge dependent performance of LLM observed in this study. Second, future work should examine how much contextual grounding is needed to improve behavioral fidelity. Richer prompts could incorporate local transportation context, such as public transportation and pedestrian infrastructure availability, EV and AV market penetration trends, or local employment environments that may shape telework feasibility. Third, future research can investigate the specific elements of the survey data generation process that contribute the most to response

compression and subgroup distortion, including prompt format, survey length, question ordering, and response-option structure.

ACKNOWLEDGEMENTS

This research was partially supported by the U.S. Department of Transportation through the Center for Understanding Future Travel Behavior and Demand (TBD) (Grant No. 69A3552344815 and No. 69A3552348320). The authors are grateful to Lisa Macias for her help in formatting this document.

REFERENCES

- Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. *Proceedings of the 40th International Conference on Machine Learning*, 337–371. <https://proceedings.mlr.press/v202/aher23a.html>
- Alsaleh, T., & Farooq, B. (2025, July 29). *Towards Locally Deployable Fine-Tuned Causal Large Language Models for Mode Choice Behaviour*. arXiv.Org. <https://arxiv.org/abs/2507.21432v2>
- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3), 337–351. <https://doi.org/10.1017/pan.2023.2>
- Baack, S. (2024). A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, 2199–2208. <https://doi.org/10.1145/3630106.3659033>
- Baeza-Yates, R. (2018). Bias on the Web. *Commun. ACM*, 61(6), 54–61. <https://doi.org/10.1145/3209581>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Bessette, D. L., Wilson, R. S., & Arvai, J. L. (2021). Do People Disagree with Themselves? Exploring the Internal Consistency of Complex, Unfamiliar, and Risky Decisions. *Journal of Risk Research*, 24(5), 593–605. <https://doi.org/10.1080/13669877.2019.1569107>
- Bhat, C. R., & Dubey, S. K. (2014). A New Estimation Approach to Integrate Latent Psychological Constructs in Choice Modeling. *Transportation Research Part B: Methodological*, 67, 68–85. <https://doi.org/10.1016/j.trb.2014.04.011>
- Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, 32(4), 401–416. <https://doi.org/10.1017/pan.2024.5>
- Blumenberg, E., Brown, A., & Schouten, A. (2020). Car-Deficit Households: Determinants and Implications for Household Travel in the U.S. *Transportation*, 47(3), 1103–1125. <https://doi.org/10.1007/s11116-018-9956-6>
- Boelaert, J., Coavoux, S., Ollion, É., Petev, I., & Präg, P. (2025). Machine Bias. How Do Generative Language Models Answer Opinion Polls?. *Sociological Methods & Research*, 54(3), 1156–1196. <https://doi.org/10.1177/00491241251330582>

- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems*, 29. https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html
- Bonnel, P., & Munizaga, M. A. (2018). Transport Survey Methods—In the Era of Big Data Facing New and Old Challenges. *Transportation Research Procedia, Transport Survey Methods in the Era of Big Data: Facing the Challenges*, 32, 1–15. <https://doi.org/10.1016/j.trpro.2018.10.001>
- Borup, M., Brown, N., Konrad, K., & Van Lente, H. (2006). The Sociology of Expectations in Science and Technology. *Technology Analysis & Strategic Management*, 18(3–4), 285–298. <https://doi.org/10.1080/09537320600777002>
- Brand, J., Israeli, A., & Ngwe, D. (2023). *Using LLMs for Market Research* (SSRN Scholarly Paper No. 4395751). Social Science Research Network. <https://doi.org/10.2139/ssrn.4395751>
- Brown, A. E. (2017). Car-Less or Car-Free? Socioeconomic and Mobility Differences among Zero-Car Households. *Transport Policy*, 60, 152–159. <https://doi.org/10.1016/j.tranpol.2017.09.016>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Byrne, R. M. J. (2007). *The Rational Imagination: How People Create Alternatives to Reality*. MIT Press.
- Byrne, R. M. J. (2017). Counterfactual Reasoning and Imagination. In *International Handbook of Thinking and Reasoning*. Routledge.
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.*, 15(3), 39:1-39:45. <https://doi.org/10.1145/3641289>
- Chapman, J., Fox, E. H., Bachman, W., Frank, L. D., Thomas, J., & Reyes, A. R. (2021). Smart Location Database Technical Documentation and User Guide Version 3.0. *Environmental Protection Agency EPA*, 2006.
- Chaudhuri, S., Purkar, P., Raghav, R., Mallick, S., Gupta, M., Jana, A., & Ghosh, S. (2025). TripCraft: A Benchmark for Spatio-Temporally Fine Grained Travel Planning. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 17035–17064). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.acl-long.834>
- Chen, C., He, Y., Wang, H., Chen, J., & Luo, Q. (2024). *DelayPTC-LLM: Metro Passenger Travel Choice Prediction under Train Delays with Large Language Models* (arXiv:2410.00052). arXiv. <https://doi.org/10.48550/arXiv.2410.00052>
- Chng, S., White, M. P., Abraham, C., & Skippon, S. (2019). Consideration of Environmental Factors in Reflections on Car Purchases: Attitudinal, Behavioural and Sociodemographic Predictors among a Large UK Sample. *Journal of Cleaner Production*, 230, 927–936. <https://doi.org/10.1016/j.jclepro.2019.05.179>

- Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni, A., Lintz, N., Pais, T. C., Jacobsson, H., Szpektor, I., Jiang, N.-J., ... Helmholz, W. (2025, July 7). Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv.Org. <https://arxiv.org/abs/2507.06261v6>
- Cramér, H. (1999). *Mathematical Methods of Statistics*. Princeton University Press.
- Dedehayir, O., & Steinert, M. (2016). The Hype Cycle Model: A Review and Future Directions. *Technological Forecasting and Social Change*, 108, 28–41. <https://doi.org/10.1016/j.techfore.2016.04.005>
- Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., Eichstaedt, J. C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D. C., Dweck, C. S., Gross, J. J., & Pennebaker, J. W. (2023). Using Large Language Models in Psychology. *Nature Reviews Psychology*, 2(11), 688–701. <https://doi.org/10.1038/s44159-023-00241-5>
- Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI Language Models Replace Human Participants? *Trends in Cognitive Sciences*, 27(7), 597–600. <https://doi.org/10.1016/j.tics.2023.04.008>
- Etezady, A., Shaw, F. A., Mokhtarian, P. L., & Circella, G. (2021). What Drives the Gap? Applying the Blinder–Oaxaca Decomposition Method to Examine Generational Differences in Transportation-Related Attitudes. *Transportation*, 48(2), 857–883. <https://doi.org/10.1007/s11116-020-10080-5>
- Frattoni, F., Bianchi, M., De Massis, A., & Sikimic, U. (2014). The Role of Early Adopters in the Diffusion of New Products: Differences between Platform and Nonplatform Innovations. *Journal of Product Innovation Management*, 31(3), 466–488. <https://doi.org/10.1111/jpim.12108>
- Geng, M., He, S., & Trotta, R. (2024). *Are Large Language Models Chameleons? An Attempt to Simulate Social Surveys* (arXiv:2405.19323). arXiv. <https://doi.org/10.48550/arXiv.2405.19323>
- Google Gemini. (2026, March 8). Gemini. <https://gemini.google.com>
- Hernandez, J. P. (2025, April 28). Reddit Statistics in 2025 and Tactics to Grow Your Brand. *Sprout Social*. <https://sproutsocial.com/insights/reddit-statistics/>
- Holmes, J., Liu, Z., Zhang, L., Ding, Y., Sio, T. T., McGee, L. A., Ashman, J. B., Li, X., Liu, T., Shen, J., & Liu, W. (2023). Evaluating Large Language Models on a Highly-Specialized Topic, Radiation Oncology Physics. *Frontiers in Oncology*, 13, 1219326. <https://doi.org/10.3389/fonc.2023.1219326>
- Horton, J. J. (2023). *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* (Working Paper No. 31122). National Bureau of Economic Research. <https://doi.org/10.3386/w31122>
- Hua, W., Fan, L., Li, L., Mei, K., Ji, J., Ge, Y., Hemphill, L., & Zhang, Y. (2024). *War and Peace (WarAgent): Large Language Model-Based Multi-Agent Simulation of World Wars* (arXiv:2311.17227). arXiv. <https://doi.org/10.48550/arXiv.2311.17227>
- Huang, K., Kockelman, K., & Gurumurthy, K. M. (2023). Innovations Impacting the Future of Transportation: An Overview of Connected, Automated, Shared, and Electric Technologies. *Transportation Letters*, 15(6), 490–509. <https://doi.org/10.1080/19427867.2022.2070091>

- Hundertmark, S. (2025, July 11). LLM Insight: Why Gemini “Notices” More and What “Knowledge Cutoff” Really Means! *Medium*. <https://sophiehundertmark.medium.com/llm-insight-why-gemini-notices-more-and-what-knowledge-cutoff-really-means-6cb44ba6ab09>
- Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). Large Language Models Struggle to Learn Long-Tail Knowledge. *Proceedings of the 40th International Conference on Machine Learning*, 15696–15707. <https://proceedings.mlr.press/v202/kandpal23a.html>
- Karmakar, P., Chaudhuri, S., Mallick, S., Gupta, M., Jana, A., & Ghosh, S. (2025, October 24). *TripTide: A Benchmark for Adaptive Travel Planning under Disruptions*. arXiv.Org. <https://arxiv.org/abs/2510.21329v1>
- Kenesei, Z., Ásványi, K., Kökény, L., Jászberényi, M., Miskolczi, M., Gyulavári, T., & Syahrivar, J. (2022). Trust and Perceived Risk: How Different Manifestations Affect the Adoption of Autonomous Vehicles. *Transportation Research Part A: Policy and Practice*, 164, 379–393. <https://doi.org/10.1016/j.tra.2022.08.022>
- Klein, N. J., & Smart, M. J. (2019). Life Events, Poverty, and Car Ownership in the United States: A Mobility Biography Approach. *Journal of Transport and Land Use*, 12(1), 395–418.
- Kozłowski, A. C., Kwon, H., & Evans, J. A. (2024, May 23). *In Silico Sociology: Forecasting COVID-19 Polarization with Large Language Models*. arXiv.Org. <https://arxiv.org/abs/2407.11190v1>
- Kroesen, M., Handy, S., & Chorus, C. (2017). Do Attitudes Cause Behavior or Vice Versa? An Alternative Conceptualization of the Attitude-Behavior Relationship in Travel Behavior Modeling. *Transportation Research Part A: Policy and Practice*, 101, 190–202. <https://doi.org/10.1016/j.tra.2017.05.013>
- Krosnick, J. A., Judd, C. M., & Wittenbrink, B. (2018). The Measurement of Attitudes. In *The Handbook of Attitudes, Volume 1: Basic Principles* (2nd ed.). Routledge.
- Lavieri, P. S., Garikapati, V. M., Bhat, C. R., & Pendyala, R. M. (2017). Investigation of Heterogeneity in Vehicle Ownership and Usage for the Millennial Generation. *Transportation Research Record*, 2664(1), 91–99. <https://doi.org/10.3141/2664-10>
- Lee, S., Peng, T.-Q., Goldberg, M. H., Rosenthal, S. A., Kotcher, J. E., Maibach, E. W., & Leiserowitz, A. (2024). Can Large Language Models Estimate Public Opinion about Global Warming? An Empirical Assessment of Algorithmic Fidelity and Bias. *PLOS Climate*, 3(8), e0000429. <https://doi.org/10.1371/journal.pclm.0000429>
- Li, D., Li, L., & Qiu, H. S. (2025a). *ChatGPT Is Not a Man but Das Man: Representativeness and Structural Consistency of Silicon Samples Generated by Large Language Models* (arXiv:2507.02919). arXiv. <https://doi.org/10.48550/arXiv.2507.02919>
- Li, H., Jiang, L., Hwang, J. D., Kim, H., Santy, S., Sorensen, T., Lin, B. Y., Dziri, N., Ren, X., & Choi, Y. (2024a). *CULTURE-GEN: Revealing Global Cultural Perception in Language Models through Natural Language Prompting* (arXiv:2404.10199). arXiv. <https://doi.org/10.48550/arXiv.2404.10199>
- Li, L., Li, J., Chen, C., Gui, F., Yang, H., Yu, C., Wang, Z., Cai, J., Zhou, J. A., Shen, B., Qian, A., Chen, W., Xue, Z., Sun, L., He, L., Chen, H., Ding, K., Du, Z., Mu, F., ... Dong, Y. (2024b). *Political-LLM: Large Language Models in Political Science* (arXiv:2412.06864). arXiv. <https://doi.org/10.48550/arXiv.2412.06864>
- Li, M., Chen, H., Wang, Y., Zhu, T., Zhang, W., Zhu, K., Wong, K.-F., & Wang, J. (2025b). *Understanding and Mitigating the Bias Inheritance in LLM-Based Data Augmentation on Downstream Tasks* (arXiv:2502.04419). arXiv. <https://doi.org/10.48550/arXiv.2502.04419>

- Li, X., Huang, F., Lv, J., Xiao, Z., Li, G., & Yue, Y. (2024c). *Be More Real: Travel Diary Generation Using LLM Agents and Individual Profiles* (arXiv:2407.18932). arXiv. <https://doi.org/10.48550/arXiv.2407.18932>
- Li, Y., Du, M., Song, R., Wang, X., & Wang, Y. (2024d). *A Survey on Fairness in Large Language Models* (arXiv:2308.10149). arXiv. <https://doi.org/10.48550/arXiv.2308.10149>
- Liu, T., Li, M., & Yin, Y. (2024). *Can Large Language Models Capture Human Travel Behavior? Evidence and Insights on Mode Choice* (SSRN Scholarly Paper No. 4937575). Social Science Research Network. <https://doi.org/10.2139/ssrn.4937575>
- Liu, T., Li, M., & Yin, Y. (2025b). *Aligning LLM with Human Travel Choices: A Persona-Based Embedding Learning Approach* (arXiv:2505.19003). arXiv. <https://doi.org/10.48550/arXiv.2505.19003>
- Liu, T., Yang, J., & Yin, Y. (2025c). Toward LLM-Agent-Based Modeling of Transportation Systems: A Conceptual Framework. *Artificial Intelligence for Transportation*, 1, 100001. <https://doi.org/10.1016/j.ait.2025.100001>
- Liu, Y., Cao, J., Liu, C., Ding, K., & Jin, L. (2025a). Datasets for Large Language Models: A Comprehensive Survey. *Artificial Intelligence Review*, 58(12), 403. <https://doi.org/10.1007/s10462-025-11403-7>
- Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X. (Alexis), Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., & Hooker, S. (2024). A Large-Scale Audit of Dataset Licensing and Attribution in AI. *Nature Machine Intelligence*, 6(8), 975–987. <https://doi.org/10.1038/s42256-024-00878-8>
- Ma, C. (2025). Evaluating Silicon Sampling: LLM Accuracy in Simulating Public Opinion on Facial Recognition Technology. *AHFE International Human Interaction and Emerging Technologies (IHiet 2025)*, 5. <https://doi.org/10.54941/ahfe1006738>
- Mo, B., Xu, H., Zhuang, D., Ma, R., Guo, X., & Zhao, J. (2023). *Large Language Models for Travel Behavior Prediction* (arXiv:2312.00819). arXiv. <https://doi.org/10.48550/arXiv.2312.00819>
- Morris, G. E., Leff, B., & Enns, P. K. (2026, June 2). *Can AI “Digital Twins” Replace Human Survey Respondents? | Verasight* [Report]. Verasight. <https://www.verasight.io/reports/can-ai-digital-twins-replace-human-respondents>
- Morvan, C., & O'Connor, A. (2017). *An Analysis of Leon Festinger's A Theory of Cognitive Dissonance*. Macat Library. <https://doi.org/10.4324/9781912282432>
- National Research Council. (2013). Nonresponse in Social Science Surveys: A Research Agenda. In *National Research Council. 2013. Nonresponse in Social Science Surveys: A Research Agenda*. Washington (p. 18293). <https://doi.org/10.17226/18293>
- Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2025). A Comprehensive Overview of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 16(5), 1–72. <https://doi.org/10.1145/3744746>
- Nishida, R., Ishigaki, T., & Onishi, M. (2025). Large Language Models Predict Transportation Mode Choice Behavior for a Variety of Alternative Sets. *Transportation Research Record*, 2679(12), 249–264. <https://doi.org/10.1177/03611981251352499>
- Oakil, A. T. M., Ettema, D., Arentze, T., & Timmermans, H. (2014). Changing Household Car Ownership Level and Life Cycle Events: An Action in Anticipation or an Action on Occurrence. *Transportation*, 41(4), 889–904. <https://doi.org/10.1007/s11116-013-9507-0>

- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8), 9.
- Raj, H., Gupta, V., Rosati, D., & Majumdar, S. (2025). *Improving Consistency in Large Language Models through Chain of Guidance* (arXiv:2502.15924). arXiv. <https://doi.org/10.48550/arXiv.2502.15924>
- Rea, L. M., & Parker, R. A. (2014). *Designing and Conducting Survey Research: A Comprehensive Guide*. John Wiley & Sons.
- Robbennolt, D., Beliveau, A., Batur, I., Pendyala, R. M., Mokhtarian, P. L., Shaw, A. F., Polzin, S. E., Conway, A. J., Chen, C., & Bhat, C. R. (2025). *Transportation Heartbeat of America Survey Wave 1: Findings and Recommendations*.
- Sameen, M., Zhang, X., & Zhao, X. (2025). *Synthesizing Attitudes, Predicting Actions (SAPA): Behavioral Theory-Guided LLMs for Ridesourcing Mode Choice Modeling* (arXiv:2509.18181). arXiv. <https://doi.org/10.48550/arXiv.2509.18181>
- Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose Opinions Do Language Models Reflect? *Proceedings of the 40th International Conference on Machine Learning*, 29971–30004. <https://proceedings.mlr.press/v202/santurkar23a.html>
- Shao, Z., Wu, J., Chen, W., & Wang, X. (2025). Personal Travel Solver: A Preference-Driven LLM-Solver System for Travel Planning. In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 27622–27642). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.acl-long.1339>
- Shaw, F. A., Malokin, A., Mokhtarian, P. L., & Circella, G. (2021). Who Doesn't Mind Waiting? Examining the Relationships between Waiting Attitudes and Person- and Travel-Related Attributes. *Transportation*, 48(1), 395–429. <https://doi.org/10.1007/s11116-019-10054-2>
- Shi, Y., & Herniman, J. (2023). The Role of Expectation in Innovation Evolution: Exploring Hype Cycles. *Technovation*, 119, 102459. <https://doi.org/10.1016/j.technovation.2022.102459>
- Simmons, G. (2023). Moral Mimicry: Large Language Models Produce Moral Rationalizations Tailored to Political Identity. In V. Padmakumar, G. Vallejo, & Y. Fu (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)* (pp. 282–297). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.acl-srw.40>
- Sullivan, G. M., & Feinn, R. (2012). *Using Effect Size—Or Why the P Value Is Not Enough*. *Journal of graduate medical education*, 4(3), 279–282. <https://doi.org/10.4300/JGME-D-12-00156.1>
- Sun, K., Xu, Y., Zha, H., Liu, Y., & Dong, X. L. (2024). Head-to-Tail: How Knowledgeable Are Large Language Models (LLMs)? A.K.A. Will LLMs Replace Knowledge Graphs? In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 311–325). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.naacl-long.18>
- Sun, Z., & Zacharias, J. (2021). Do Housing Tenure and Public Transport Provision Matter in Automobile Use in Bedroom Suburban Communities? Evidence from Beijing. *Journal of Housing and the Built Environment*, 36(1), 241–262. <https://doi.org/10.1007/s10901-020-09748-2>

- Tang, Y., Tuncel, D., Koerner, C., & Runkler, T. (2025). The Few-Shot Dilemma: Over-prompting Large Language Models. *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*, 134–141. <https://doi.org/10.1109/FLLM67465.2025.11391015>
- U.S. Census Bureau. (2022, September 15). *The Number of People Primarily Working From Home Tripled Between 2019 and 2021* [Press Release]. U.S. Census Bureau. Census.Gov. <https://www.census.gov/newsroom/press-releases/2022/people-working-from-home.html>
- U.S. Census Bureau. (2026). *American Community Survey 2024 (5-Year Estimates)*. Census.Gov. <https://www.census.gov/programs-surveys/acs.html>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Łukasz, & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- Villarreal-Zegarra, D., & Bellido-Boza, L. (2026). *Generation of Synthetic Data in Health Surveys Using Large Language Models* (p. 2026.01.27.26345015). medRxiv. <https://doi.org/10.64898/2026.01.27.26345015>
- Vogelsmeier, L. V. D. E., Oliveira, E., Misiejuk, K., López-Pernas, S., & Saqr, M. (2025). Delving into the Psychology of Machines: Exploring the Structure of Self-Regulated Learning via LLM-Generated Survey Responses. *Computers in Human Behavior*, 173, 108769. <https://doi.org/10.1016/j.chb.2025.108769>
- Wang, J. Y., Sukiennik, N., Li, T., Su, W., Hao, Q., Xu, J., Huang, Z., Xu, F., & Li, Y. (2024). *A Survey on Human-Centric LLMs* (arXiv:2411.14491). arXiv. <https://doi.org/10.48550/arXiv.2411.14491>
- Wang, L., Duan, P., He, Z., Lyu, C., Chen, X., Zheng, N., Yao, L., & Ma, Z. (2025). AI-Driven Day-to-Day Route Choice. *Transportation Research Part C: Emerging Technologies*, 180, 105307. <https://doi.org/10.1016/j.trc.2025.105307>
- Wang, X., & Renne, J. L. (2023). Socioeconomics of Urban Travel in the U.S.: Evidence from the 2017 NHTS. *Transportation Research Part D: Transport and Environment*, 116, 103622. <https://doi.org/10.1016/j.trd.2023.103622>
- Wang, X., Shaw, F. A., Mokhtarian, P. L., & Watkins, K. E. (2023a). Response Willingness in Consecutive Travel Surveys: An Investigation Based on the National Household Travel Survey Using a Sample Selection Model. *Transportation*, 50(6), 2339–2373. <https://doi.org/10.1007/s11116-022-10312-w>
- Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X., & Liu, Q. (2023b). *Aligning Large Language Models with Human: A Survey* (arXiv:2307.12966). arXiv. <https://doi.org/10.48550/arXiv.2307.12966>
- Wu, Z., Peng, R., Ito, T., Onizuka, M., & Xiao, C. (2025, June 24). *LLM-Based Social Simulations Require a Boundary*. arXiv.Org. <https://arxiv.org/abs/2506.19806v2>
- Yan, Y., Liu, T., & Yin, Y. (2026). Valuing Time in Silicon: Can Large Language Models Replicate Human Value of Travel Time. *Travel Behaviour and Society*, 44, 101245. <https://doi.org/10.1016/j.tbs.2026.101245>
- Ye, W., Ou, M., Li, T., Chen, Y., Ma, X., Yanggong, Y., Wu, S., Fu, J., Chen, G., Wang, H., & Zhao, J. (2023). *Assessing Hidden Risks of LLMs: An Empirical Study on Robustness, Consistency, and Credibility* (arXiv:2305.10235). arXiv. <https://doi.org/10.48550/arXiv.2305.10235>

- Yousefzadeh Barri, E., Farber, S., Jahanshahi, H., & Beyazit, E. (2022). Understanding Transit Ridership in an Equity Context Through a Comparison of Statistical and Machine Learning Algorithms. *Journal of Transport Geography*, 105, 103482. <https://doi.org/10.1016/j.jtrangeo.2022.103482>
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Dong, Z., Hou, Y., Zhang, B., Min, Y., Zhang, J., Liu, P., Wang, X., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., ... Wen, J.-R. (2026). A Survey of Large Language Models. *Frontiers of Computer Science*, 20(12), 2012627. <https://doi.org/10.1007/s11704-026-60308-3>
- Zheng, Y., Wang, S., & Zhao, J. (2021). Equality of Opportunity in Travel Behavior Prediction with Deep Neural Networks and Discrete Choice Models. *Transportation Research Part C: Emerging Technologies*, 132, 103410. <https://doi.org/10.1016/j.trc.2021.103410>

APPENDICES

APPENDIX A

A.1 Mathematical Formulation of the Confirmatory Factor Analysis Model

For the mathematical formulation of the latent construct model, let l be the index for the latent constructs ($l = 1, 2, \dots, L$) and label each latent construct as z_l^* . In the current context $L = 3$. Then, the $(L \times 1)$ vector of latent constructs may be written as $\mathbf{z}^* = (z_1^*, z_2^*, \dots, z_L^*)'$. Each of these latent constructs is identified using a series of ordinal indicators (shown in Figure 1 in Section 5.2), labeled as n ($n = 1, 2, \dots, N$). Then, let J_n be the number of categories for the n th ordinal indicator ($J_n \geq 2$) and let j_n be the corresponding index ($j_n = 1, 2, \dots, J_n$). Then, each indicator may be written based on an underlying latent variable $\tilde{\tau}_n^*$ whose horizontal partitioning leads to the observed selection. These underlying variables may be written as a function of the three latent constructs as:

$$\tilde{\tau}_n^* = \tilde{\mathbf{d}}_n' \mathbf{z}^* + \tilde{\varepsilon}_n \quad (6)$$

with $\tilde{\psi}_{n,a_n-1} < \tilde{\tau}_n^* < \tilde{\psi}_{n,a_n}$, assuming that the individual chooses the a_n^{th} ordinal category. $\tilde{\mathbf{d}}_n$ is an $(L \times 1)$ vector of loadings of each latent construct on the ordinal indicator, the $\tilde{\psi}$ terms represent thresholds to be estimated, and $\tilde{\varepsilon}_n$ represents the standard normal random error. For each indicator, the thresholds must be ordered $\tilde{\psi}_{n,0} < \tilde{\psi}_{n,1} < \tilde{\psi}_{n,2} \dots < \tilde{\psi}_{n,J_n-1} < \tilde{\psi}_{n,J_n}$, with $\tilde{\psi}_{n,0} = -\infty$, $\tilde{\psi}_{n,1} = 0$, and $\tilde{\psi}_{n,J_n} = +\infty$. Stack these thresholds into the vectors $\tilde{\boldsymbol{\psi}}_n = (\tilde{\psi}_{n,2}, \tilde{\psi}_{n,3}, \dots, \tilde{\psi}_{n,J_n-1})$ and matrix $\tilde{\boldsymbol{\Psi}} = (\tilde{\boldsymbol{\psi}}_1', \tilde{\boldsymbol{\psi}}_2', \dots, \tilde{\boldsymbol{\psi}}_N)'$. Then, also stack the N underlying continuous variables $\tilde{\tau}_n^*$ into an $(N \times 1)$ vector $\tilde{\boldsymbol{\tau}}^*$ and error terms $\tilde{\varepsilon}_n$ into an $(N \times 1)$ vector $\tilde{\boldsymbol{\varepsilon}}$. Define the $(N \times L)$ matrix $\tilde{\mathbf{d}} = (\tilde{\mathbf{d}}_1, \tilde{\mathbf{d}}_2, \dots, \tilde{\mathbf{d}}_N)$, and stack the N lower thresholds $\tilde{\psi}_{n,a_n-1}$ into an $(N \times 1)$ vector $\tilde{\boldsymbol{\psi}}_{low}$ and the N upper thresholds $\tilde{\psi}_{n,a_n}$ into another vector $\tilde{\boldsymbol{\psi}}_{up}$. Then the equation for the ordinal indicators may be written in matrix form:

$$\tilde{\boldsymbol{\tau}}^* = \tilde{\mathbf{d}} \mathbf{z}^* + \tilde{\boldsymbol{\varepsilon}}, \quad \tilde{\boldsymbol{\psi}}_{low} < \tilde{\boldsymbol{\tau}}^* < \tilde{\boldsymbol{\psi}}_{up} \quad (7)$$

Finally, each latent construct may be written as a function of exogenous variables as:

$$z_l^* = \boldsymbol{\alpha}_l' \mathbf{w} + \eta_l \quad (8)$$

where $\mathbf{w}$ is a $(D \times 1)$ vector of observed variables, $\boldsymbol{\alpha}_l$ is a corresponding $(D \times 1)$ vector of coefficients, and η_l is a random error term that is assumed to be standard normally distributed. These components may again be stacked across the latent constructs into the $(L \times D)$ matrix $\mathbf{A} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_L)'$, and the $(L \times 1)$ vectors $\mathbf{z}^* = (z_1^*, z_2^*, \dots, z_L^*)'$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_L)'$. In order to accommodate unobserved factors that may also jointly influence multiple latent variables, we allow a multivariate normal correlation structure for $\boldsymbol{\eta}$. $\boldsymbol{\eta} \sim MVN_L[\mathbf{0}_L, \boldsymbol{\Gamma}]$, where $\mathbf{0}_L$ is an $(L \times 1)$ column vector of zeros, and $\boldsymbol{\Gamma}$ is an $(L \times L)$ correlation matrix. Then, the equation for the latent constructs may be written in matrix form as:

$$\mathbf{z}^* = \boldsymbol{\alpha}\boldsymbol{w} + \boldsymbol{\eta} \quad (9)$$

Thus, the estimation of the CFA for the latent constructs takes the form of a maximum likelihood estimation, identifying the vector $\tilde{\boldsymbol{d}}$ corresponding to the loadings of the latent constructs on the indicators while simultaneously the multivariate regression of the latent constructs on the covariates allows for the identification of the coefficients in the vector $\boldsymbol{\alpha}$.

APPENDIX B

Table B.1: CFA Results for Attitudinal Latent Constructs for Reference Dataset

Latent variable structural equation model	Pro-transit orientation		Pro-environmental disposition		Perception of online shopping (in)accessibility	
	Coeff	t-stat	Coeff	t-stat	Coeff	t-stat
Age (<25 years old)						
25-34 years old	0.118	2.888	0.132	3.253	-	-
35-44 years old	0.103	2.685	0.129	3.385	-	-
45-54 years old	-	-	-	-	-0.226	-5.651
55-64 years old	-0.198	-4.781	-0.179	-4.409	-0.390	-9.106
65 years old and above	-0.325	-8.342	-0.375	-9.718	-0.515	-12.141
Education level (high school or less)						
Some college	-	-	-0.062	-2.379	-0.112	-3.106
Bachelor or higher	0.180	6.693	0.107	3.611	-0.195	-5.334
Employment status (not employed)						
Employed full time	-	-	0.053	2.271	0.107	3.066
Employed part time	-	-	-	-	0.112	2.249
Race (not White or Black)						
White	-0.311	-9.978	-0.266	-9.496	-0.158	-4.386
Black	0.140	4.072	-	-	-0.163	-3.498
Number of children in the household	-0.043	-3.398	-0.022	-1.713	-	-
Number of adults in the household	0.041	4.173	-	-	-	-
Annual household income (< \$50,000)						
\$50,000-\$99,999	-0.070	-2.516	-0.105	-3.786	-0.222	-6.202
\$100,000 and above	-	-	-	-	-0.193	-4.871
Housing tenure type (own)						
Rent	0.113	4.309	-	-	-0.091	-2.883
Other tenure type	-	-	-	-	-	-
Housing unit type (other housing type)						
Single family housing	-	-	-	-	-	-
Apartment	0.349	9.687	0.302	9.481	-	-
Latent variable measurement equation model	Loading	t-stat	Loading	t-stat	Loading	t-stat
If I moved, I would try to choose a home where I had access to public transportation	0.967	75.806	-	-	-	-
I like the idea of public transit as a means of travel for me personally	1.001	74.017	-	-	-	-
I try to make good use of the time I spend in, on, or waiting for transportation vehicles	0.617	52.216	-	-	-	-
I use less polluting means of transportation (e.g., walking, biking, and public transportation) as much as possible.	-	-	1.019	78.898	-	-
I try to make as many of my trips as possible by walking or bicycling	-	-	0.980	73.701	-	-
I am committed to an environmentally-friendly lifestyle	-	-	0.530	44.341	-	-
I routinely encounter challenges with products or services I order online	-	-	-	-	0.749	47.925
I have difficulty ordering what I need online	-	-	-	-	0.725	45.584
Delivery times are long or inconsistent where I live	-	-	-	-	0.659	43.071
Correlations between latent variables			Coeff	t-stat	Coeff	t-stat
Pro-transit orientation	-	-	0.768	88.464	0.156	9.920
Pro-environmental disposition	-	-	-	-	0.230	15.004
Perception of online shopping (in)accessibility	-	-	-	-	-	-

Table B.2: CFA Results for Attitudinal Latent Constructs for LLM-generated Dataset

Latent variable structural equation model	Pro-transit orientation		Pro-environmental disposition		Perception of online shopping (in)accessibility	
	Coeff	t-stat	Coeff	t-stat	Coeff	t-stat
Age (<25 years old)						
25-34 years old	-0.317	-5.414	-0.244	-4.404	0.363	4.520
35-44 years old	-0.480	-8.300	-0.363	-6.589	0.646	8.040
45-54 years old	-0.594	-10.057	-0.470	-8.394	0.790	9.658
55-64 years old	-0.660	-10.876	-0.626	-10.928	0.758	9.210
65 years old and above	-0.729	-12.142	-0.700	-12.362	0.282	3.593
Education level (high school or less)						
Some college	0.371	10.379	0.376	11.089	-0.529	-10.168
Bachelor or higher	0.636	17.217	0.692	19.703	-0.741	-13.146
Employment status (not employed)						
Employed full time	-0.174	-5.222	-0.192	-5.881	-0.600	-11.746
Employed part time	-	-	0.085	2.315	-0.422	-6.029
Race (not White or Black)						
White	-0.410	-11.912	-0.482	-16.390	-0.142	-2.850
Black	0.290	7.923	-	-	0.278	4.305
Number of children in the household	-	-	-0.023	-2.193	0.104	5.213
Number of adults in the household	0.056	4.120	0.021	1.649	-	-
Annual household income (< \$50,000)						
\$50,000-\$99,999	0.162	4.562	0.248	7.363	-0.889	-16.374
\$100,000 and above	0.519	12.846	0.592	15.414	-1.324	-19.550
Housing tenure type (own)						
Rent	0.646	17.037	0.325	9.425	0.237	4.541
Other tenure type	0.171	3.322	-	-	0.790	8.555
Housing unit type (other housing type)						
Single family housing	-0.307	-7.261	-0.119	-2.993	-0.319	-5.432
Apartment	0.799	15.449	0.448	9.279	-0.277	-3.914
Latent variable measurement equation model	Loading	t-stat	Loading	t-stat	Loading	t-stat
If I moved, I would try to choose a home where I had access to public transportation	0.340	68.641	-	-	-	-
I like the idea of public transit as a means of travel for me personally	0.305	67.290	-	-	-	-
I try to make good use of the time I spend in, on, or waiting for transportation vehicles	0.215	49.467	-	-	-	-
I use less polluting means of transportation (e.g., walking, biking, and public transportation) as much as possible.	-	-	0.408	78.045	-	-
I try to make as many of my trips as possible by walking or bicycling	-	-	0.300	58.795	-	-
I am committed to an environmentally-friendly lifestyle	-	-	0.242	58.149	-	-
I routinely encounter challenges with products or services I order online	-	-	-	-	0.104	20.060
I have difficulty ordering what I need online	-	-	-	-	0.421	36.220
Delivery times are long or inconsistent where I live	-	-	-	-	0.137	23.172
Correlations between latent variables			Coeff	t-stat	Coeff	t-stat
Pro-transit orientation	-	-	0.889	94.222	-0.147	-5.839
Pro-environmental disposition	-	-	-	-	-0.181	-7.501
Perception of online shopping (in)accessibility	-	-	-	-	-	-