# Large Language Models Yield Unsustainable Tourist Flows: Testing Algorithmic Biases Using the Baseline-Rescaling-Outcome Model

Seonjin Lee and Lori Pennington-Gray
School of Hospitality and Tourism Management, University of South Carolina, Richardson Family SmartState Center for Economic Excellence in Tourism and Economic Development

Abstract

Despite widespread adoption of generative AI in tourism, empirical evidence on its impacts is scarce. This study shows how large language models deviate from empirical tourism patterns, attributing the deviations to popularity biases. We propose the Baseline-Rescaling-Outcome Model to test four types of popularity biases in AI-generated tourism recommendations. Large language models tend to produce more seasonal, more unequal, and less diversified tourist flows. These models also favor popular destination-month pairs but show mixed results for other popularity biases. Findings show that the widespread adoption of generative AI can undermine the sustainability and resilience of tourism systems. Thus, we urge tourism scholars and practitioners to proactively assess generative AI biases and their consequences in tourism.

*Keywords:* generative AI impact, tourist flow network, algorithmic bias, large language model simulation, tourism sustainability, tourism resilience

## Introduction

Generative AI has moved quickly from novelty to an everyday tool. People use these algorithms as always-ready collaborators that support tasks ranging from sending emails to critical business decisions (Marr, 2023). Tourism is no exception. Most travelers already use generative AI for travel planning, and major travel intermediaries like Expedia and TripAdvisor have integrated these algorithms into their platforms (Booking.com, 2025; Expedia, 2023; Tripadvisor, 2023). Beyond travel planning, generative AI support core operations in tourism industries, including marketing, human resource management, and tourism experience design (Amadeus, 2024; Dogru et al., 2025).

Despite widespread adoption of generative AI in tourism, empirical evidence on its impacts is scarce (Hsu et al., 2024; Mellors, 2025). Major concerns include ethical and cultural biases in these algorithms (Ali, 2025; Law et al., 2025). As more tourists and practitioners rely on generative AI, algorithmic biases may drive behavioral shifts that reproduce or amplify existing societal issues (Kordzadeh & Ghasemaghaei, 2022;

Seonjin Lee https://orcid.org/0000-0002-3944-0738
Correspondence concerning this article should be addressed to Seonjin Lee, School of Hospitality and Tourism Management, University of South Carolina, Richardson Family SmartState Center for Economic Excellence in Tourism and Economic Development, 1705 College St., Columbia, SC 29208, USA, Email: seonjin@sc.edu

Vicente & Matute, 2023). However, these algorithms are often proprietary and lack interpretability, impeding the detection of biases (Bai et al., 2025; Gallegos et al., 2024).

This study investigates how the rapid adoption of generative AI and its algorithmic biases can reshape tourism patterns. We develop the *Baseline-Rescaling-Outcome Model* that can test algorithmic biases using empirical expectations, hypothesized bias mechanisms, and algorithmic outcomes. Using demographic profiles from the US Census Bureau, we created a simulated sample of one million US residents. Large language models then recommended one domestic travel destination per individual. We explain deviations between AI-generated and empirical US domestic tourism patterns as systematic *biases* favoring popular options and pairs of options.

Compared to empirical tourism patterns, large language models yield less diverse tourist flows concentrated in specific combinations of destination, origin, and month. They also produce more unevenly distributed tourist flows with stronger seasonality and less diversified demand. The models favor visiting destinations during their peak seasons, indicating amplification of destination-month popularity. However, tested models vary in the strength and direction of other popularity biases. The findings provide early empirical evidence that generative AI may undermine the sustainability and resilience of the tourism sector.

## Background

### Generative AI and its impacts on tourism

This study focuses on generative AI and its implications for tourism. The term AI encompasses various technologies designed to simulate human intelligence. Generative AI is an algorithm that learns patterns from large datasets to generate text, images, or video content (H. Li et al., 2025). When addressing generative AI in general, we avoid usage-specific terms like *chatbots* or brand-specific terms like *ChatGPT*.

Limited empirical evidence exists on how and to what extent generative AI will affect tourism (Hsu et al., 2024; Mellors, 2025). Although tourism and hospitality literature on generative AI is growing, they primarily examine adoption behaviors: who, when, and why tourists and practitioners adopt generative AI (see recent reviews by Gössling & Mei, 2025; H. Li et al., 2025). Some preliminary works highlight benefits of generative AI to businesses and tourists. Announcing generative AI integration gave tourism firms competitive advantages in market value (Jung et al., 2026). Additionally, underperforming businesses can improve revenue by using generative AI for product marketing (Fan et al., 2025). For tourists, generative AI reduces cognitive load during travel planning, thereby increasing visit intentions and decision satisfaction (Shin et al., 2025).

However, others caution negative impacts of generative AI on tourism (Law et al., 2025; Lehto et al., 2025). Few studies examine how generative AI has biases that favor mainstream tourism patterns (Andreev et al., 2025; Mellors, 2025). Such biases can shape behaviors of generative AI users, leading to real-world impacts that reproduce or amplify existing biases in society (Kordzadeh & Ghasemaghaei, 2022; Vicente & Matute, 2023). These preliminary findings echo standing debates about how technologies—from search engines to social media—have both benefited and harmed tourism (Gong et al., 2024; Leung et al., 2013).

### Challenges of defining and measuring biases in generative AI

Both conceptual and empirical tourism studies frequently raise concerns about biases in generative AI. A particularly prominent focus is on social biases and their ethical implications. For example, Law et al. (2025) identify ethical risks associated with AI adoption in tourism and hospitality sectors, urging more inclusive practices to mitigate "biased and discriminatory actions" (p.287). Hsu et al. (2024) note that generative AI "could perpetuate stereotypes and result in discrimination" (p.2), proposing generative AI fine-tuned for tourism. Viglia et al. (2024) caution against biases in AI-generated tourism data, giving a specific example of AI generating racist content. This trend aligns with broader literature on algorithmic biases that

emphasize ethical challenges, especially racism and stereotypes (see Ghosh & Wilson, 2025; Kordzadeh & Ghasemaghaei, 2022). Beyond ethical concerns, several scholars note risks of a popularity bias in generative AI, where algorithms favor popular options and underrepresent unpopular ones (Law et al., 2024; Lehto et al., 2025). This bias poses practical concerns for the tourism sector, as it can intensify over-tourism at popular destinations while damaging tourism sectors at less popular destinations (Mellors, 2025).

Two key challenges complicate the analysis of generative AI biases: definition and measurement. What constitutes "biased" algorithms is often vague in tourism and hospitality studies. This issue is not unique to tourism scholarship; broader AI bias literature has also been criticized for lacking an explicit definition of bias (Ghosh & Wilson, 2025). Even regulatory frameworks, like the EU AI Act, often leave bias undefined (van Bekkum, 2025). Such ambiguity creates discrepancies between the concerns raised and the empirical evidence provided (Blodgett et al., 2020).

Algorithmic bias has no single definition. When the focus is on stereotyping, studies define bias as the *act of* unjustified association between social groups and attributes (Bai et al., 2025). This definition follows social psychology literature on implicit associations (Greenwald et al., 1998). Studies examining ethical implications of algorithmic biases define bias based on *outcomes or treatment*: unequal allocation of resources or unfavorable representation of social groups (Blodgett et al., 2020; Gallegos et al., 2024; Kordzadeh & Ghasemaghaei, 2022). This use of bias parallels discrimination laws that require proof of disparate outcomes, which treat biases as individual beliefs that are unactionable (Seiner, 2006). Others distinguish bias from harm, defining bias as an inclination of an algorithm or a deviation from data (Ghosh & Wilson, 2025; Wu et al., 2024). This neutral definition of bias is closely tied to statistics and computer science that view bias as inherent and unavoidable (Chen et al., 2023; Ghosh & Wilson, 2025). Taken together, the definition of algorithmic bias varies not only across disciplines but also by the specific bias being examined.

Even with a clear definition, measuring biases in generative AI is difficult. Despite attempts to make generative AI more transparent, many models remain proprietary and closed-source. Ali (2025) proposes using IBM Fairness 360 and Google What-If toolkits for assessing biases in AI-generated tourism data. However, these tools require access to training data and model internals, which commercial generative AI models like ChatGPT do not provide (Gallegos et al., 2024). Bias detection is further complicated by safety tunings that recent generative AI models undergo to avoid explicit biases (Bai et al., 2025; Santurkar et al., 2023). Moreover, most existing AI bias metrics are designed to quantify disparities, not how such disparities arise (see review of bias metrics by Gallegos et al., 2024; Kordzadeh & Ghasemaghaei, 2022). For example, metrics may reveal that generative AI disproportionately suggests resort destinations to certain racial groups. Yet they cannot identify whether such disparities stem from popularity biases that favor popular destinations or peak seasons, racial stereotypes, or a combination of these mechanisms. Together, these conceptual and methodological challenges limit our ability to test biases in generative AI models and assess their implications for tourism.

## Baseline-Rescaling-Outcome Model for testing algorithmic biases in tourism

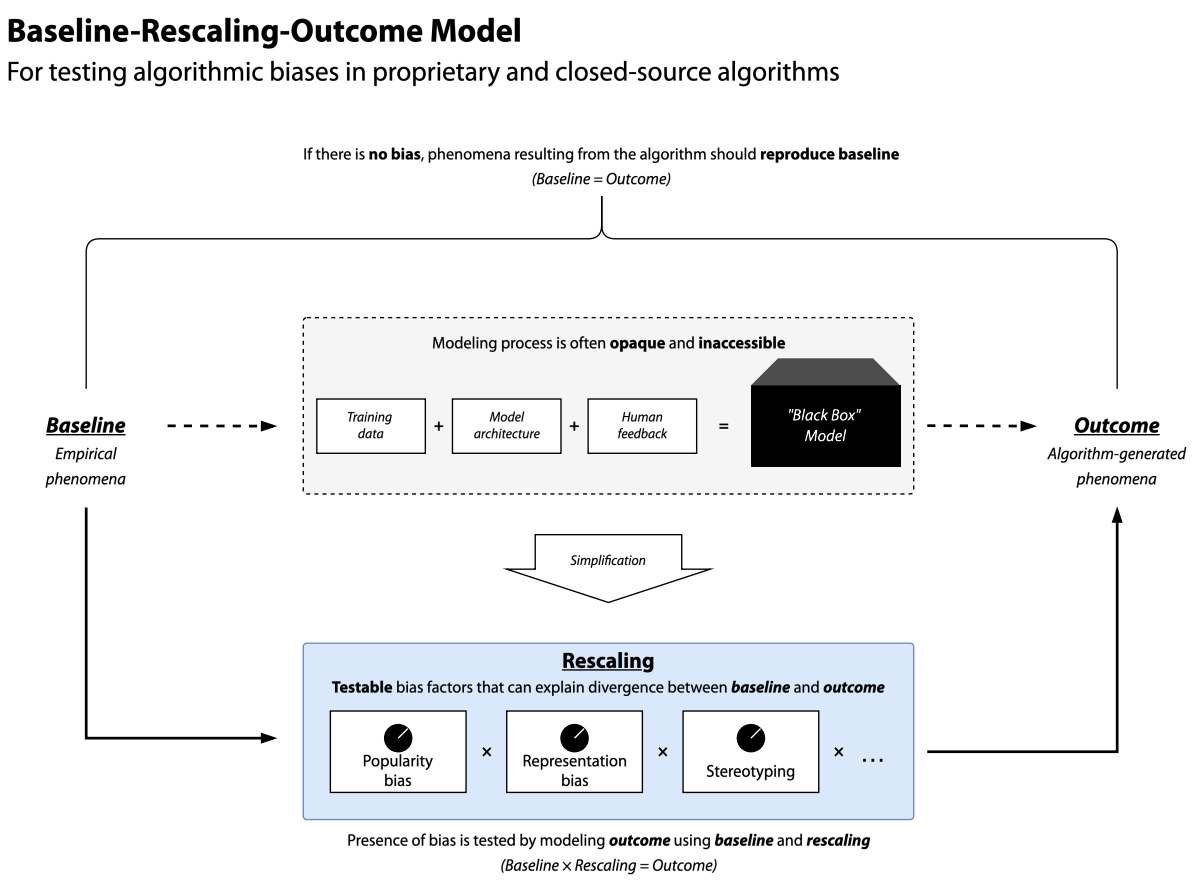
We propose the Baseline-Rescaling-Outcome model, designed to address the challenges of measuring biases in proprietary and closed-source algorithms. Bias is defined as *systematic deviation of algorithmic outputs from empirically grounded expectations of phenomena*. Our definition distinguishes bias from harm, as harm depends on affected stakeholders (Blodgett et al., 2020; Ghosh & Wilson, 2025). For example, if an algorithm systematically favors one destination over others, this bias benefits the favored but is harmful to the rest. Even within the favored destination, the tourism sector may benefit while locals suffer from over-tourism. We therefore separate the detection of biases from the assessment of their consequences.

Our model tests bias by modeling the divergence between the algorithmic *outcome* and empirical *baseline* as a function of *rescaling* factors representing hypothesized bias mechanisms (Figure 1). This ap-

proach has three advantages over existing bias tests. First, the model can test biases without access to model internals or training data. We draw upon social science methods that infer human biases from behavioral outcomes like response speed and error rates (Greenwald et al., 1998). Like cognitive processes, we assume the algorithm's internals and its modeling processes as unobservable, inferring biases from outputs instead. Our approach is also adaptable beyond generative AI and tourism, to any context with empirically grounded expectations, rescaling factors, and algorithmic outputs. Finally, the rescaling factors enable interpretable bias diagnosis that translates directly into practical suggestions for debiasing. However, the model does not provide *causal* explanations of how biases arise in algorithms, which are secluded inside socio-technical processes involving data generation, model design, and human feedback (Bai et al., 2025; Santurkar et al., 2023; Viglia et al., 2024).

**Figure 1**

*Baseline-Rescaling-Outcome Model*




### Outcome: Projected tourism patterns under complete reliance on the algorithm

The *outcome* projects scenarios in which the phenomena are entirely driven by the algorithm, consistent with scenario-based projection models. Examples of projection models include Shared Socioeconomic Pathways for climate trajectories and COVID-19 diffusion models (Adam, 2020; IPCC, 2021). These models contain "worst-case" scenarios that project the most extreme outcomes: climate projections without emission reductions or infection rates without government interventions. For example, testing bias in generative AI

tourism recommendations requires projecting the outcome when generative AI makes all destination choices. Similarly, biases in hiring algorithms for tourism firms can be tested by projecting a scenario where all hiring decisions are made by the algorithm. Although unrealistic, such undiluted projections are necessary to reveal the full extent of biases.

### Baseline: Empirically grounded expectations of phenomena

The empirical *baseline* anchors the model to real-world expectations of phenomena. Prior works employed similar approaches by comparing real-world distributions with algorithmic outputs to assess representation and popularity biases (Abdollahpouri & Mansoury, 2020; Santurkar et al., 2023). Consider a situation where an algorithm has a four-in-ten chance of suggesting US domestic tourists to visit San Francisco. Could we say that this algorithm has a systematic tendency to favor San Francisco as a tourism destination? If four in ten Americans visit San Francisco, then the algorithm is simply reproducing what is expected in the real-world tourism patterns. However, if only one in ten US domestic tourists visit San Francisco, then the algorithm does favor San Francisco beyond the empirical expectation. Thus, the empirical baseline sets benchmarks for "unbiased" algorithmic outputs.

### Rescaling: Testable factors for hypothesized bias mechanisms

*Rescaling* tests specific mechanisms that explain deviations between the algorithmic outcome and the empirical baseline. Under the null condition of no bias, the algorithm should reproduce the baseline without rescaling factors. By testing hypothesized bias mechanisms, rescaling factors provide interpretable tests of systematic biases in algorithms, going beyond measuring the degree of biases. This approach also allows testing the direction of biases. The algorithm could exhibit biases that reduce real-world inequalities (Ghosh & Wilson, 2025). For example, it may favor less popular destinations and off-peak seasons, diversifying the tourism demand across destinations and time. Our model captures such biases as negative rescaling factors, explaining algorithm outputs as mixtures of amplification and attenuation of empirical patterns.

## Study design

We apply the Baseline-Rescaling-Outcome Model to test popularity biases in generative AI travel recommendations. Popularity bias is defined as the tendency where popular options "are recommended even more frequently than their popularity would warrant" (Abdollahpouri & Mansoury, 2020, p. 1). Humans exhibit similar behaviors, such as tourists flocking to popular destinations (Lee & Pennington-Gray, 2025b). This study therefore tests whether generative AI amplifies or attenuates such popularity biases beyond what is observed empirically.

Understanding this bias in tourism is particularly timely and practical, as major online travel agencies are already integrating generative AI into their platforms. Examining popularity in the tourism context is also of theoretical significance. Unlike recommending movies or music, tourism recommendations need to consider spatial and temporal dimensions of travel choices. The decision to travel is not only about *whether* to travel, but also *from where*, *to where*, and *when*. Thus, we further extend popularity bias into four types that are tourism-specific: destination popularity, month popularity, destination-month popularity, and origin-destination popularity biases. The first two measure whether generative AI amplifies or attenuates the popularity of destinations and peak months. Bias may also exist in specific combinations of destinations, origins, and months. Destination-month popularity bias captures whether algorithms favor specific destinations during specific months. Similarly, origin-destination popularity bias captures the tendency to excessively pair tourists from specific origins to specific destinations.

Figure 2 summarizes our data collection and analysis process. The large language model simulation involves generating tourist flow using empirically grounded demographic profiles. Separately from the AI-driven simulation, two data sources were used to estimate real-world tourism patterns. This empirical data

is then used to derive baseline expectations and popularity factors. Finally, we combine simulation and empirical data to test four hypothesized popularity biases by fitting the Poisson model.

## Data collection

### *Definition of population and simulated samples*

The population of our simulations is US residents aged 18 and over. The US domestic tourism market is one of the largest in the world, offering geographically and socioeconomically diverse destinations (UNWTO, n.d.). Hence, the US context provides sufficient variability for large-scale simulations, while excluding complications of international tourism like visa requirements. Most generative AI models also perform better on English tasks, making the US context advantageous for these models (Qin et al., 2025).

We used 2019-2023 American Community Survey 5-Year Public Use Microdata Sample data to derive stratum weights for the population. Appendix A summarizes sex, age, and household income of the population. From this population, we took a random sample of 1,000 individuals stratified by state, sex, age, and income. This sampling procedure was repeated 1,000 times, yielding one million simulated individuals with demographic characteristics that mirror the population. By using empirically derived profiles, we ensure that the demographic distribution of simulated travelers resembles that of US domestic tourists. This approach also fixes the number of outgoing tourists from each state, allowing us to control for origin-specific propensity to travel. One limitation is that generative AI models may also influence the decision to travel itself, which we do not account for here.

### *Large language model simulations*

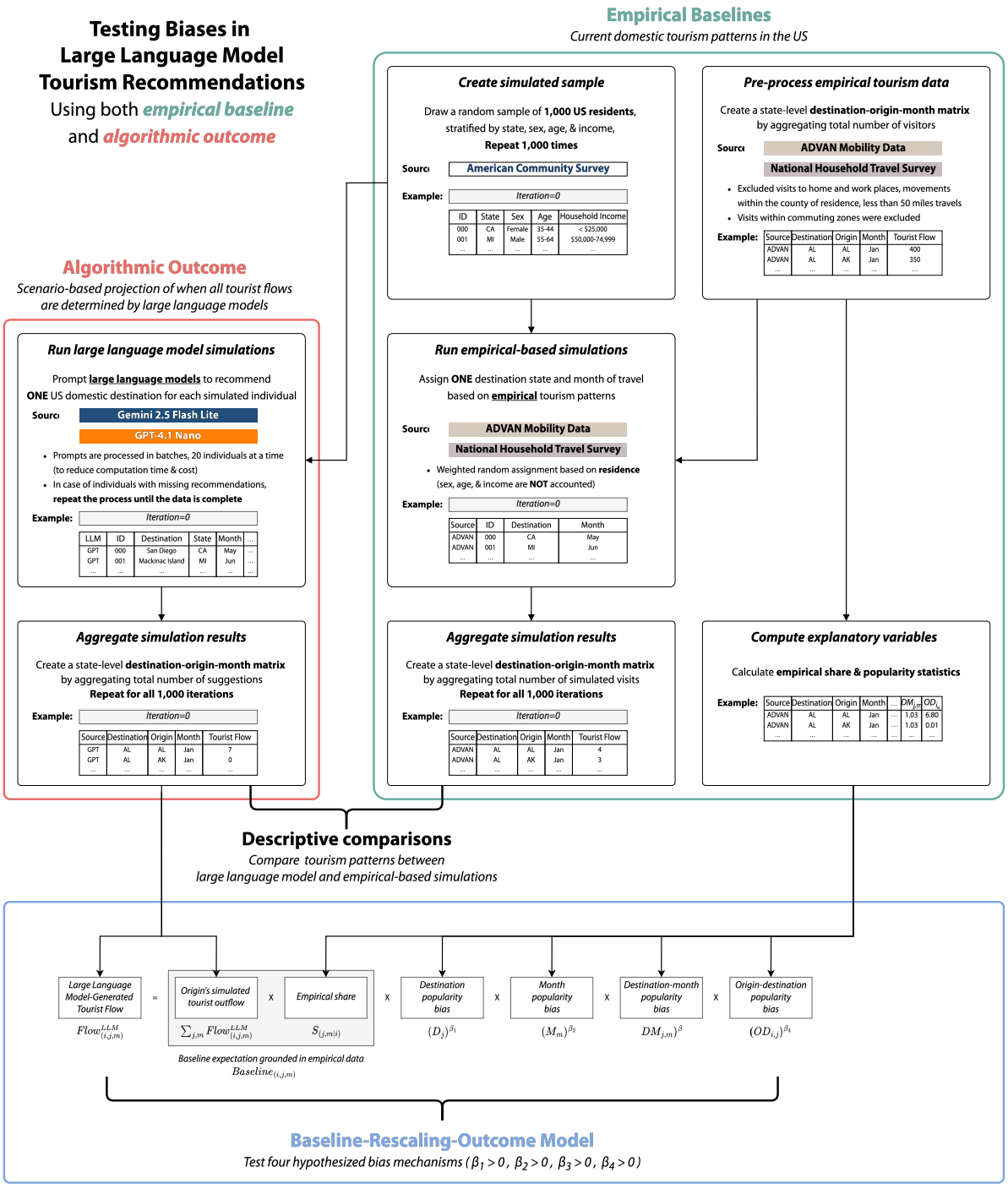
Social science researchers are increasingly using synthetic data from generative AI. We highlight four key differences between our approach and prior studies. First, we use AI-generated data to show how these AI models *do not* replicate human behavior, contrary to claims that generative AI can mimic real-world tourists (for example, Ali, 2025; Viglia et al., 2024; Xiong et al., 2024). Next, our simulations rely on empirically derived demographic profiles, instead of using arbitrary or non-representative profiles that confound the results (for example, Andreev et al., 2025). Third, we repeat the simulation for 1,000 iterations to sufficiently capture the uncertainty in large language model outputs, unlike studies that relied on a few handpicked responses or a single simulation run (for example, Mellors, 2025; Xiong et al., 2024), Finally, our approach recognizes the network and temporal nature of tourism. Beyond looking at destination choice alone, we simulate the complete network of tourist flows between all origin-destination pairs across months.

The simulation starts with a system prompt containing the simulation context, response structure, and examples (available in Appendix B). To simplify the simulation, we assume that each person chooses one domestic destination within the US (50 US states and the District of Columbia). We instructed large language models to act as travel agents recommending one domestic travel destination based on the provided demographic profile. Twenty simulated tourists are processed at a time to improve the efficiency of the simulation. Our process resembles agent-based modeling using large language models but differs in that we prompt the models to act as travel agents, not tourists (Gao et al., 2024). This choice reflects our objective to project how generative AI would influence tourist flows if widely adopted, rather than using it to substitute for human travelers. We captured the probabilistic nature of large language model outputs by repeating the simulation with 1,000 individuals for 1,000 iterations. This process produces a distribution of simulated tourist visits that accounts for stochastic variations in large language model outputs.

We establish the consistency of our findings across two different large language models: Google's Gemini 2.5 Flash Lite (version June 2025) and OpenAI's GPT-4.1 Nano (version April 2025). They are among the leading commercial large language models in capabilities and market share. Given the scale of our simulations, we chose their smallest variants optimized for speed and cost. Although the two large language models are closed-source, they have comparable pricing structures ($0.10 per million input tokens;

## Figure 2

*Overview of the simulation and analysis process*



$0.30 and $0.40 per million output tokens, respectively). We also chose these two models also because they use different architectures. Gemini 2.5 family uses a mixture-of-experts architecture, while GPT-4.1 family relies on traditional transformer architecture. Because the two models are developed by different companies, they are likely to be trained on different datasets and tuning processes, which is ideal for assessing the generalizability of our findings.

All requests were made from the IP address of the university located in the southeastern US, using custom automation scripts. We explicitly instructed large language models not to use IP-specific details when generating recommendations. Data collection continued until we achieved a complete dataset. We then aggregated the data to create destination-origin-month matrices for each iteration and model, where each cell represents the number of tourists from the origin. $i$ to destination $j$ in a month $m$.

### *Empirical baseline data and simulations*

The empirical data serve two purposes in this study. It provides the baseline for descriptively comparing characteristics of AI-simulated tourist flows against real-world tourism patterns. Additionally, we use the empirically derived baseline and popularity factors to explain the discrepancies between AI-simulated and empirical tourist flows. Because biases also exist in the empirical mobility data, we rely on two different data sources to ensure the robustness of our findings (Z. Li et al., 2024). Our primary data source is the Advan Research (2022) Mobility Data, which estimates movements between US census block groups based on mobile device panels. This dataset is our primary data source due to its high spatial and temporal resolution. The supplementary data source is the Department of Transportation's 2022 National Household Travel Survey. This is the only official national travel survey in the US that collects travel behavior data such as trip purpose, modes, and commuting zones (Federal Highway Administration, 2022).

For both datasets, we used monthly data from January through December 2022 (the latest available for National Household Travel Survey). Following pre-processing steps were employed to filter out non-tourism mobility flows. We first excluded visits to home and work locations, and movements within the county of residence (Lee & Pennington-Gray, 2025a). Additionally, trips under 50 miles one-way or within the same commuting zones were considered non-tourism. This offers more conservative estimates of tourist visits by filtering out short-distance and commuting trips that are less likely to be tourism-related. These estimates were then used to compute the empirical shares and the four popularity factors.

Since we employed an iterative approach for the large language model simulations, direct comparison with empirical data is inappropriate. Therefore, we also conducted empirical-based simulations for descriptive comparison between AI-simulated and empirical tourist flows. Using empirical estimates as weights, we simulated tourist flows by randomly assigning destinations and months to each individual in the simulated sample. Due to data anonymization, demographic factors could not be incorporated in the empirical-based simulations. Therefore, we assume that the probability of traveling to another state in a given month is equal for all individuals in a given origin state. For instance, if 10% of Illinois residents visited Florida in January 2024, all Illinois residents were given a 0.1 probability to travel to Florida in January (irrespective of other demographic factors). Same as the large language model simulations, the empirical-based simulation was repeated 1,000 times. Subsequently, the results were aggregated to destination-origin-month matrices.

## Operationalizing the Baseline-Rescaling-Outcome Model

Below is the empirical model for testing four popularity biases. Define the number of tourist flow from origin $i$ to destination $j$ in month $m$ as $Flow_{(i,j,m)}$. Let $P_{(i,j,m)}$ be the share of tourists from origin $i$ to destination $j$ in month $m$ over all tourist flow ($Flow_{(i,j,m)}/\sum Flow_{(i,j,m)}$). Under the null condition that large language models produce "unbiased" travel suggestions, we expect:

$$Flow^{LLM}_{(i,j,m)} \sim \sum_{j,m} Flow^{LLM}_{(i,j,m)} \cdot P^{Empirical}_{(j,m|i)} = Baseline_{(i,j,m)} \quad (1)$$

where $P^{Empirical}_{(j,m|i)}$ is the share of tourists traveling to destination $j$ in month $m$ given origin $i$, observed empirically ($P_{(i,j,m)}/\sum_{j,m} P_{(i,j,m)}$). This share is multiplied by the total number of large language model-produced tourists from origin $i$, which scales the expected number of tourists based on total tourist outflow from origin $i$. Meaning, the right-hand side of Equation 1 is the baseline expectation when generative AI can perfectly replicate empirical tourism patterns ($Baseline_{(i,j,m)}$).

Destination and month popularity are defined as:

$$\begin{aligned} D_j &= \sum_{i,m} P^{Empirical}_{(i,j,m)} \\ M_m &= \sum_{i,j} P^{Empirical}_{(i,j,m)} \end{aligned} \tag{2}$$

where $D_j$ is the popularity of destination $j$ across all origins and months, and $M_m$ is the popularity of month $m$ across all origins and destinations.

We account for the popularity of specific destination-month pairs as a joint probability of choosing destination $j$ in month $m$ beyond what can be expected from their independent popularities:

$$DM_{j,m} = \frac{\sum_i P^{Empirical}_{(i,j,m)}}{D_j \cdot M_m} \tag{3}$$

The denominator in Equation 3 is the expected popularity of the destination-month pair if destination and month popularities were independent. If $DM_{j,m} > 1$, destination-month pair $(j, m)$ is more popular than expected under independence, while $DM_{j,m} < 1$ indicates the pair is less popular than expected. Similarly, we account for the popularity of specific origin-destination pairs:

$$OD_{i,j} = \frac{\sum_m P^{Empirical}_{(i,j,m)}}{O_i \cdot D_j} \tag{4}$$

where $O_i = \sum_{j,m} P^{Empirical}_{(i,j,m)}$. Same as Equation 3, $OD_{i,j}$ measures how popular the origin-destination pair $(i, j)$ is than what would be expected if origin and destination popularities were independent.

We test the four popularity rescaling factors using the following multiplicative model:

$$\frac{Flow^{LLM}_{(i,j,m)}}{Baseline_{(i,j,m)}} = (D_j)^{\beta_1} \cdot (M_m)^{\beta_2} \cdot (DM_{j,m})^{\beta_3} \cdot (OD_{i,j})^{\beta_4} \tag{5}$$

Under the null hypotheses of no systematic bias, we expect all $\beta = 0$. If $\beta > 0$ for a factor, it positively rescales that factor's popularity in their suggestions; if $\beta < 0$, it negatively rescales that factor's popularity.

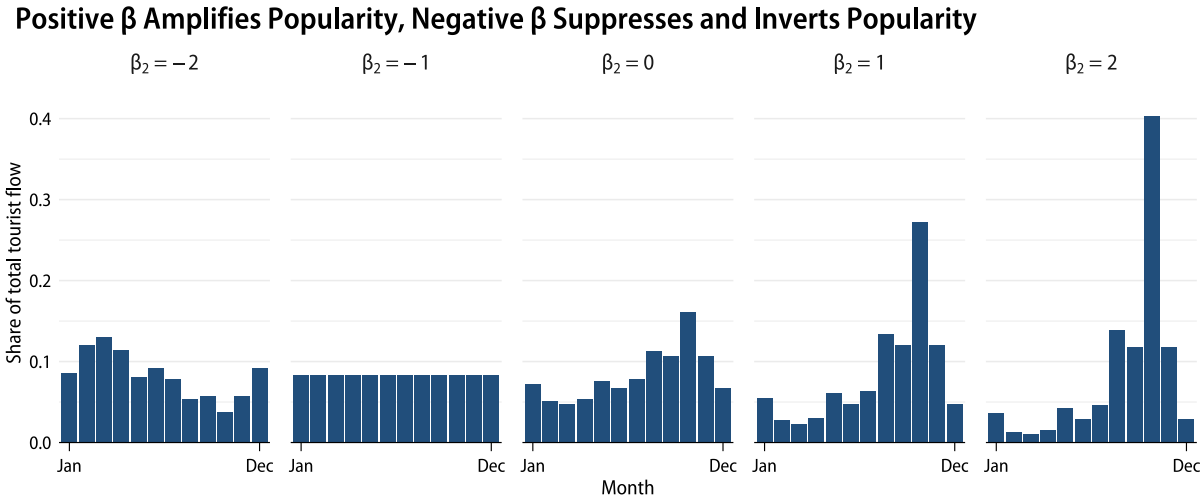
Figure 3 illustrates how different $\beta_2$ values (month popularity bias) rescale the distribution of the monthly tourist share. For example, if $\beta_2 = 1$, seasonal variation in tourist flow is amplified, whereas $\beta_2 = -1$ produces a uniform distribution across months. This test is independent of destination-month popularity bias ($\beta_3 \neq 0$) because $DM_{j,m}$ measure the popularity beyond destination and month popularity can predict. Meaning, if we generate tourist flows with only the destination-month popularity bias ($\beta_3 \neq 0$), the resulting data will show no changes in overall destination or month popularity ($\beta_1 = 0, \beta_2 = 0$).

By taking the log of Equation 5, we can fit a regression model that tests the four popularity biases:

$$\begin{aligned} \ln Flow^{LLM}_{(i,j,m)} &= \ln Baseline_{(i,j,m)} \\ &\quad + \beta_1 \cdot \ln D_j + \beta_2 \cdot \ln M_m + \beta_3 \cdot \ln DM_{j,m} + \beta_4 \cdot \ln OD_{i,j} \end{aligned} \tag{6}$$

**Figure 3**

*Example of how different $\beta_2$ values affect the distribution of the monthly tourist share*




## Hypotheses testing

The last step of the analysis is to combine simulation and empirical data to test hypothesized popularity biases. We achieve this goal by fitting Equation 6 using the Poisson count model. Essentially, the model explains variations in large language model-simulated tourist flows ($Flow^{LLM}_{(i,j,m)}$) beyond what can be expected by empirical data ($Baseline_{(i,j,m)}$), using the four popularity factors as predictors. We chose the Poisson pseudo-maximum likelihood estimator, which is widely used in estimating gravity models of trade and migration. The Poisson pseudo-maximum likelihood estimator only requires the conditional mean to be correctly specified, without requiring a specific distributional assumption. This estimator is robust to heteroskedasticity and having many zeros in the dependent variable, making the estimator suitable for our analysis (Silva & Tenreyro, 2011). Because we run 1,000 iterations of simulations, we fit the model for each iteration and collect the results across iterations to assess the uncertainty in the effect estimates.

# Results

## Descriptive analysis

Table 1 presents descriptive statistics of simulation and empirical data. Origin total outflows are consistent across simulations because the simulated tourists are stratified by state population size. The median tourist flow per destination-origin-month cell is 0 in all simulations, indicating high sparsity. This sparsity is expected, as each iteration assigns 1,000 tourists across 31,212 possible combinations (51 origins×51 destinations×12 months). By design, only about 3.1% (=1,000/31,212) of cells would have any tourist visits even under complete random assignment. But the upper tail of the distribution differs across simulations. Large language model simulations show a stronger concentration of tourists in the most popular destination-origin-month combination than empirical-based ones (Max=72 for Gemini 2.5 Flash Lite and 46 for GPT-4.1 Nano; 17 for ADVAN Mobility Data and 30 for National Household Travel Survey).

Similar patterns emerge when examining the presence and absence of tourist flow. Table 2 compares presence and absence of destination-origin-month combinations between large language model and empirical simulations. We aggregated the simulated visits across all 1,000 iterations to reduce sparsity. Large language models exclude many destination-origin-month combinations. About 35.8% to 51.0% of

empirically observed combinations never appearing in large language model simulations. By contrast, large language models rarely suggest destination-origin-month combinations that are absent from empirical data, with fewer than 10% of combinations appearing only in large language model simulations. Meaning, large language models prune empirical tourism patterns but rarely generate new ones.

**Table 1**

*Descriptive statistics of simulation and empirical data*

| Variable | Notation | Min | Max | Median | Mean | SD |
|---|---|---|---|---|---|---|
| *Simulation: ADVAN Mobility Data* | | | | | | |
| Tourist flow | $Flow_{(i,j,m)}$ | 0.000 | 17.000 | 0.000 | 0.032 | 0.258 |
| Origin total outflow | $\sum_{j,m} Flow_{(i,j,m)}$ | 0.000 | 154.000 | 13.000 | 19.608 | 22.472 |
| *Simulation: National Household Travel Survey* | | | | | | |
| Tourist flow | $Flow_{(i,j,m)}$ | 0.000 | 30.000 | 0.000 | 0.032 | 0.322 |
| Origin total outflow | $\sum_{j,m} Flow_{(i,j,m)}$ | 0.000 | 154.000 | 13.000 | 19.608 | 22.472 |
| *Simulation: Gemini 2.5 Flash Lite* | | | | | | |
| Tourist flow | $Flow_{(i,j,m)}$ | 0.000 | 72.000 | 0.000 | 0.032 | 0.502 |
| Origin total outflow | $\sum_{j,m} Flow_{(i,j,m)}$ | 0.000 | 154.000 | 13.000 | 19.597 | 22.459 |
| *Simulation: GPT-4.1 Nano* | | | | | | |
| Tourist flow | $Flow_{(i,j,m)}$ | 0.000 | 46.000 | 0.000 | 0.032 | 0.481 |
| Origin total outflow | $\sum_{j,m} Flow_{(i,j,m)}$ | 0.000 | 154.000 | 13.000 | 19.591 | 22.458 |
| *Empirical: ADVAN Mobility Data* | | | | | | |
| Empirical share | $S_{(j,m\|i)}$ | 0.000 | 0.078 | <0.001 | 0.002 | 0.005 |
| Destination | $D_j$ | <0.001 | 0.115 | 0.014 | 0.020 | 0.022 |
| Month | $M_m$ | 0.038 | 0.131 | 0.084 | 0.083 | 0.024 |
| Destination-month | $DM_{j,m}$ | <0.001 | 0.014 | 0.001 | 0.002 | 0.002 |
| Origin-destination | $OD_{i,j}$ | 0.000 | 0.078 | <0.001 | <0.001 | 0.002 |
| *Empirical: National Household Travel Survey* | | | | | | |
| Empirical share | $S_{(j,m\|i)}$ | 0.000 | 0.132 | <0.001 | 0.002 | 0.008 |
| Destination | $D_j$ | 0.001 | 0.125 | 0.016 | 0.020 | 0.021 |
| Month | $M_m$ | 0.046 | 0.161 | 0.074 | 0.083 | 0.032 |
| Destination-month | $DM_{j,m}$ | <0.001 | 0.018 | 0.001 | 0.002 | 0.002 |
| Origin-destination | $OD_{i,j}$ | 0.000 | 0.110 | <0.001 | <0.001 | 0.003 |

Note: For *simulation* data, statistics are calculated over 1,000 iterations (N=31,212,000). Statistics for *empirical* data are based on a single destination-origin-month matrix (N=31,212).

We further examine the differences in distribution of tourist flow across simulations. Results are summarized by taking the median across 1,000 iterations. Mathematical definitions of used metrics are available in Appendix C. First, we examine the distribution of tourist share by destination states ($D_j$) and months ($M_m$). Inequality is measured using the Gini index, where higher values indicate greater concentration of tourists across states or months. In all simulations, tourist visits concentrate in popular states such

**Figure 4**

*Distributional characteristics of tourist flows across simulations*

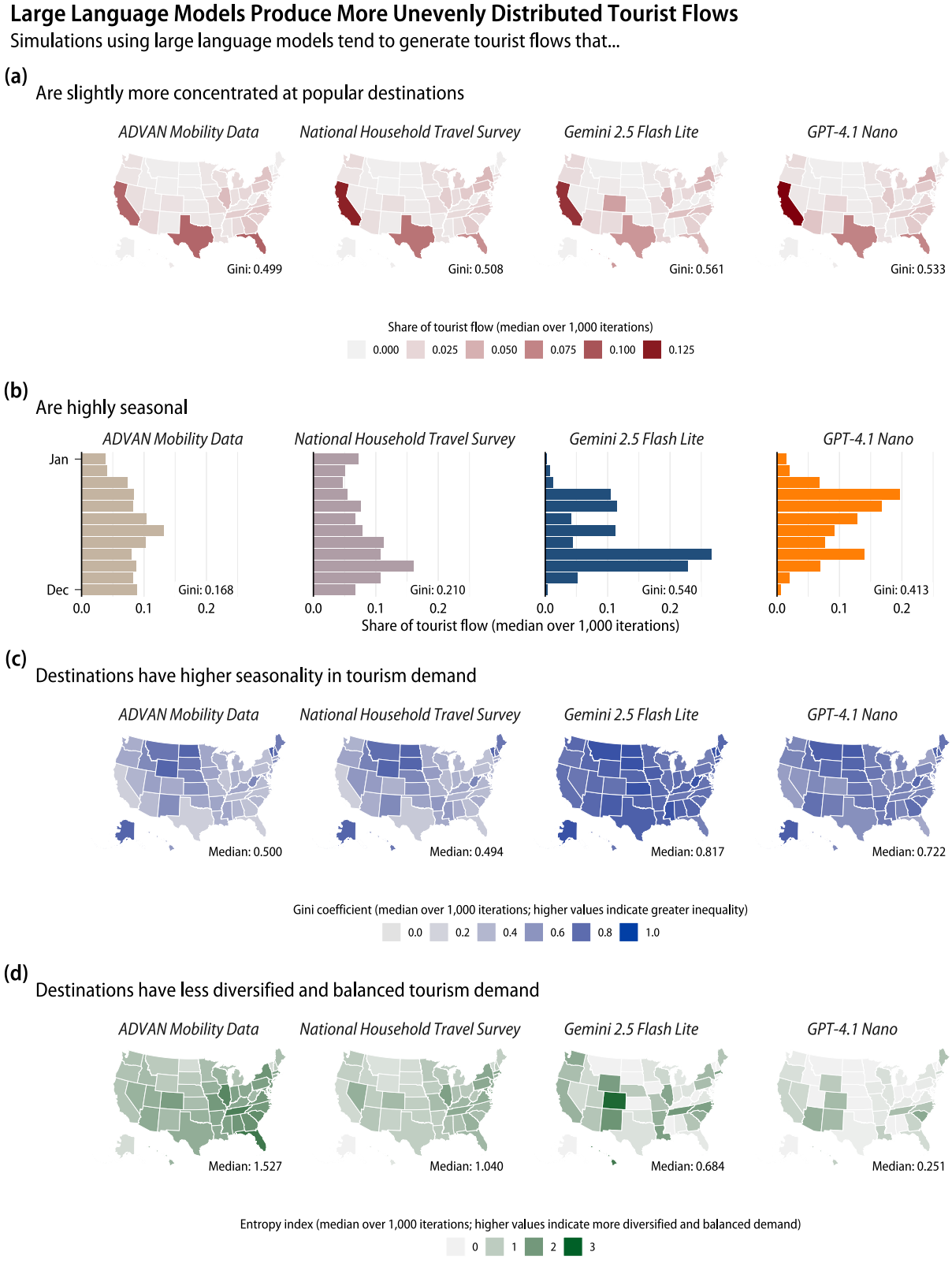


**Table 2**

*Agreement in the presence of any tourist flow between large language model simulations and empirical-based simulations*

| | *Simulation: ADVAN* | | *Simulation: NHTS* | |
|---|---|---|---|---|
| | No flow | Any flow | No flow | Any flow |
| *Simulation: Gemini 2.5 Flash Lite* | | | | |
| No flow | 7,081 (22.7%) | 15,916 (51.0%) | 10,714 (34.3%) | 12,283 (39.4%) |
| Any flow | 701 (2.2%) | 7,514 (24.1%) | 1,885 (6.0%) | 6,330 (20.3%) |
| *Simulation: GPT-4.1 Nano* | | | | |
| No flow | 7,119 (22.8%) | 14,711 (47.1%) | 10,648 (34.1%) | 11,182 (35.8%) |
| Any flow | 663 (2.1%) | 8,719 (27.9%) | 1,951 (6.3%) | 7,431 (23.8%) |

Note: ADVAN=ADVAN Mobility Data, NHTS=National Household Travel Survey. Counts and percentages of destination-origin-month cells with no tourist flow in both simulations, any flow in either simulation, or any flow in both simulations. Percentages are calculated based on the number of possible destination-origin-month combinations (N=31,212). Based on one million simulated visits aggregated across all 1,000 iterations.

as California, Florida, and Texas (Figure 4a). Large language model simulations show higher Gini indices than empirical-based simulations, indicating greater inequality in tourist arrivals across states (ADVAN Mobility Data=0.499 and National Household Travel Survey=0.508; Gemini 2.5 Flash Lite=0.561 and GPT-4.1 Nano=0.533).

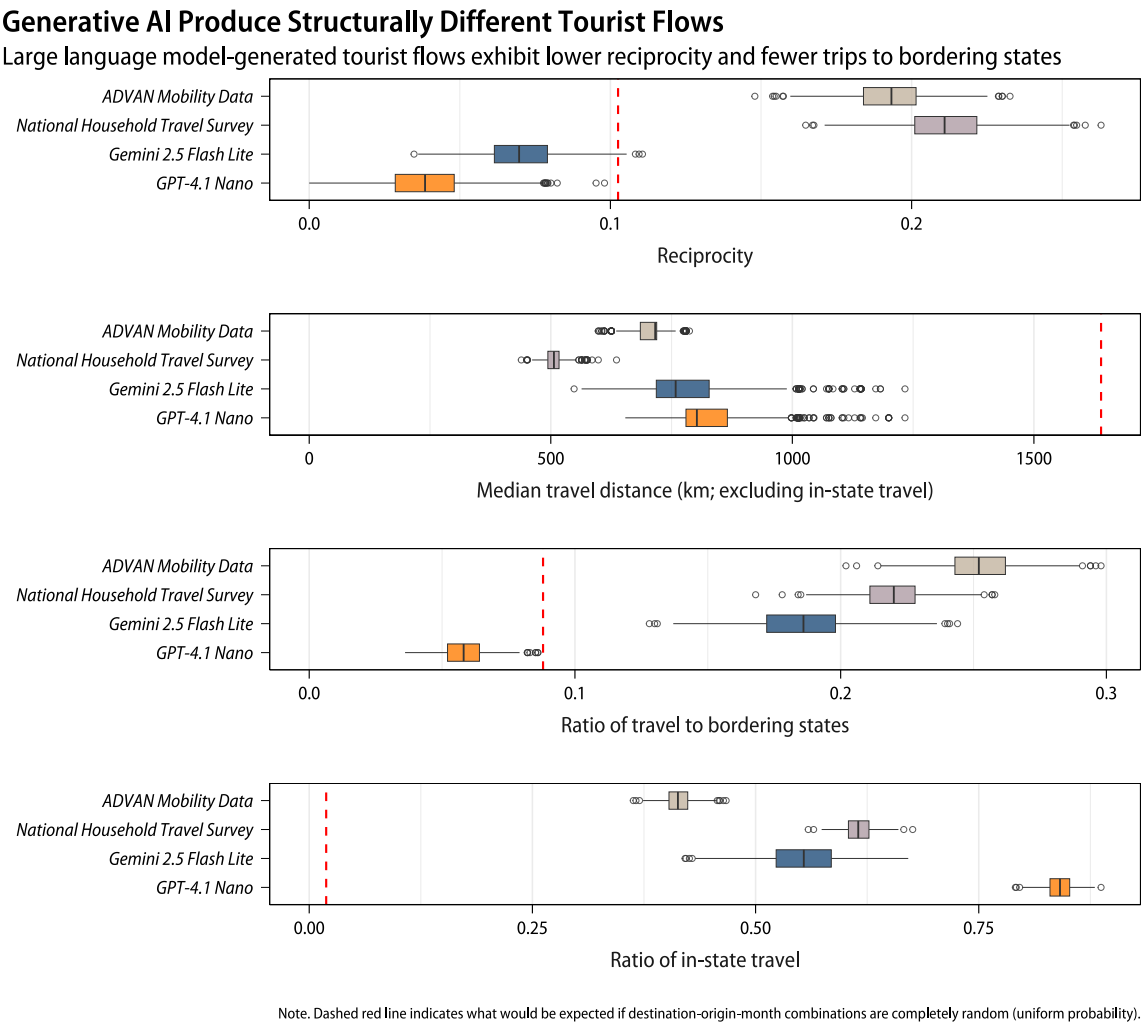
Seasonality is substantially stronger in large language model simulations (Figure 4b). Gini indices of monthly tourist share for large language model simulations far exceed empirical levels (Gemini 2.5 Flash Lite=0.540 and GPT-4.1 Nano=0.413; ADVAN Mobility Data=0.168 and National Household Travel Survey=0.210). Although, the two large language models differ in peak tourism months. Gemini 2.5 Flash Lite tends to recommend traveling in September and October, whereas GPT-4.1 Nano shows peaks in April, May, and September. Large language models worsen the seasonality of tourism demand for most destinations. For Figure 4c, we calculated the Gini index of monthly tourist share for each destination state and took the median over 1,000 iterations. In empirical-based simulations, a few states show relatively high seasonality, such as Alaska and Hawaii. By contrast, most states exhibit high seasonality in large language model simulations, indicated by higher Gini indices across states (Median of medians=0.500 and 0.494 for ADVAN Mobility Data and National Household Travel Survey; 0.817 and 0.722 for Gemini 2.5 Flash Lite and GPT-4.1 Nano).

Large language models generate tourist flows that destinations rely on a few origins (Figure 4d). We use entropy index instead of Gini index because entropy is more suitable for measuring whether a destination has diversified and balanced tourism demand, which are characteristics of resilient destinations (Lee & Pennington-Gray, 2025a). Empirical simulations show that states such as Florida, Illinois, and North Carolina have diversified and balanced demand (higher entropy). In large language model simulations, entropy indices are overall lower with much fewer states with relatively high entropy values (Median of medians=1.527 and 1.040 for ADVAN Mobility Data and National Household Travel Survey; 0.684 and 0.251 for Gemini 2.5 Flash Lite and GPT-4.1 Nano). These patterns suggest that large language models produce less diversified and balanced origin-destination flows that are vulnerable to shocks.

Additionally, we analyze the overall structure of tourist flows. We constructed an origin-destination

**Figure 5**

*Characteristics of tourist flow structure across simulations*




Note. Dashed red line indicates what would be expected if destination-origin-month combinations are completely random (uniform probability).

matrix by aggregating tourist numbers at the year level. Then we computed four statistics capturing how the structure of tourist flows differs across simulations. Box plots in Figure 5 summarize the distribution of the four network-level statistics across 1,000 iterations, colored by simulation scenario. The dotted red line indicates the expected value under complete randomness, which assumes equal probability of choosing any destination (a probability of $1/51$).

Large language models produce tourist flows with clear separation between tourist sending and receiving states, as indicated by weaker reciprocity. Reciprocity measures the tendency of two states to exchange a similar number of tourists. Empirical-based simulations show stronger reciprocity than the random expectation (Median=0.193 and 0.211 for ADVAN Mobility Data and National Household Travel Survey). However, large language model simulations show weaker reciprocity than expected by chance (Median=0.070 and 0.038 for Gemini 2.5 Flash Lite and GPT-4.1 Nano).

Both Gemini 2.5 Flash Lite and GPT-4.1 Nano suggest farther destinations than empirical data, based on median travel distance excluding in-state trips (Median of medians=758 and 802). This tendency is partially due to large language models' lower propensity to suggest trips to bordering states. Gemini 2.5 Flash

Lite has a lower ratio of tourist flows between bordering states than the empirical models (Median=0.186). GPT-4.1 Nano shows even lower tendency to suggest bordering states, lower than what would be expected under randomness (Median=0.058). But the two models differ in their tendency to suggest in-state tourism. Gemini 2.5 Flash Lite shows a ratio of in-state trips comparable to the empirical models (Median=0.554). In contrast, GPT-4.1 Nano shows a stronger preference for recommending tourists to travel within their own state (Median=0.841).

**Model estimation results**

Table 3 and Figure 6 summarize the estimated effects of the four popularity biases on large language model-simulated tourist flows (Equation 6). Across two large language models and two empirical baselines, destination-month popularity consistently has the largest positive coefficient. Meaning, large language models show a tendency to favor popular destination-month combinations when generating travel recommendations. Although, 95% credible intervals indicate high uncertainty around the estimates for Gemini 2.5 Flash Lite with both empirical baselines. The coefficients can be interpreted as elasticity. For example, GPT-4.1 Nano with ADVAN Mobility Data baseline had a median coefficient of 1.771 for destination-month popularity. If real-world data shows that a particular destination-month combination is twice as popular than what is expected under independence of destination and month popularity ($DM_{(j,m)} = 2$), then the expected tourist flow generated by GPT-4.1 Nano for that combination is approximately 3.4 times higher ($2^{1.771}$), holding other factors constant.

We find mixed results for the rest of the popularity biases. Some effects are specific to the large language model. For example, Gemini 2.5 Flash Lite consistently shows negative coefficients for destination popularity, indicating that it tends to negatively rescale popular destinations (Median $\beta_1$ = -0.175 and -0.245 with ADVAN Mobility Data and National Household Travel Survey baselines). The same pattern is only observed for GPT-4.1 Nano with National Household Travel Survey baseline (Median $\beta_1$ = 0.180) but not with the ADVAN Mobility Data baseline (Median $\beta_1$ = -0.006). GPT-4.1 Nano with shows positive rescaling for origin-destination popularity (Median $\beta_4$ = 0.298 and 0.229 with ADVAN Mobility Data and National Household Travel Survey baselines), while Gemini 2.5 Flash Lite shows mixed findings (Median $\beta_4$ = 0.066 and -0.168 with ADVAN Mobility Data and National Household Travel Survey baselines).

Finally, we note that the month popularity coefficients are negative for GPT-4.1 Nano (Median $\beta_2$ = -0.789 and -0.849 with ADVAN Mobility Data and National Household Travel Survey baselines). This result is contradictory, given that prior descriptive analysis indicated GPT-4.1 Nano having greater seasonality. One possible explanation is that peak months in GPT-4.1 Nano simulations poorly align with those in empirical data (see Figure 4b). Therefore, the model attempts to fit the month popularity factor by flattening the empirical month popularity distribution.

**Robustness checks**

We conducted following robustness checks to test the sensitivity of our findings to different simulation choices (reported in Appendix D). These additional simulations were performed with the first 100 iterations due to budget and computing time constraints. First, using alternative prompts does not substantially change the main findings. Large language model outputs are sensitive to the specific prompts used. Hence, we collected additional simulation data using slightly modified prompts (see Appendix B). One version excluded explicit instruction that demographic factors influence tourist choices ("reduced instruction" prompt). Another version instructed the models to act as *tourists* choosing destinations instead of travel agents giving suggestions ("tourist persona" prompt). None of these alterations substantially changed the main findings (Figure D1). One notable exception is that Gemini 2.5 Flash Lite with reduced instruction prompt showed negative coefficients for month popularity. However, this case also showed stronger effects for destination-month and origin-destination popularity than the original prompt did.

**Table 3**

*Median and 95% credible intervals of Poisson model coefficients across 1,000 iterations*

| | *Empirical: ADVAN* | | | *Empirical: NHTS* | | |
|---|---|---|---|---|---|---|
| | Median | 2.5% | 97.5% | Median | 2.5% | 97.5% |
| *Simulation: Gemini 2.5 Flash Lite* | | | | | | |
| $\beta_1$: Destination | -0.175 | -0.242 | -0.115 | -0.245 | -0.387 | -0.112 |
| $\beta_2$: Month | 0.089 | -0.192 | 0.423 | 0.887 | 0.472 | 1.294 |
| $\beta_3$: Destination-Month | 0.882 | -0.097 | 1.480 | 2.669 | 0.369 | 4.993 |
| $\beta_4$: Origin-Destination | 0.066 | -0.027 | 0.141 | -0.168 | -0.254 | -0.086 |
| *Simulation: GPT-4.1 Nano* | | | | | | |
| $\beta_1$: Destination | -0.006 | -0.040 | 0.027 | 0.180 | 0.139 | 0.229 |
| $\beta_2$: Month | -0.789 | -0.913 | -0.644 | -0.849 | -1.032 | -0.675 |
| $\beta_3$: Destination-Month | 1.771 | 1.042 | 2.180 | 1.611 | 0.950 | 2.249 |
| $\beta_4$: Origin-Destination | 0.298 | 0.252 | 0.340 | 0.229 | 0.175 | 0.289 |

Note: ADVAN=ADVAN Mobility Data, NHTS=National Household Travel Survey. Summary of Poisson model results over 1,000 iterations.

Changing the temperature parameter also does not alter the results. The temperature parameter controls randomness in large language model outputs. Higher temperature settings produce more diverse outputs, while lower temperature settings generate more deterministic outputs. The default temperature setting used in our main analysis is 1.0. We collected additional data using temperatures of 0.5 and 1.5 (Figure D2). GPT-4.1 Nano with a temperature of 1.5 could not generate valid outputs and hence was excluded. Similar to Bai et al. (2025), we find that temperature setting has minimal impact on the popularity bias of large language models.

Larger models in the Gemini 2.5 and GPT-4.1 series, as well as models from other providers, still show positive destination-month popularity bias. We repeated the main analysis with larger variants of Gemini 2.5 and GPT-4.1 series models (Gemini 2.5 Flash and GPT-4.1 Mini). Additionally, we collected data using xAI's Grok 3 Mini and Meta's Llama 4 Scout to examine whether the findings are generalizable beyond OpenAI and Google models. Across all models tested, destination-month popularity shows the strongest positive effects (Figure D3). Compared to the models used in our main analysis, larger models show stronger destination-month popularity bias, not weaker.

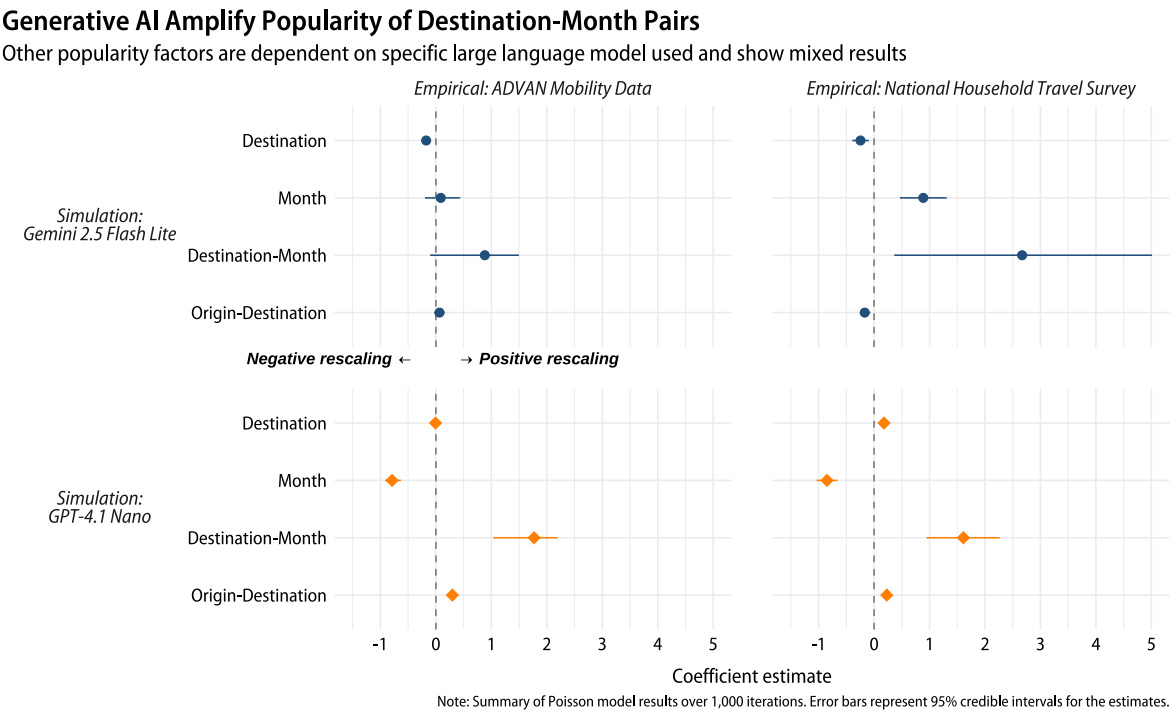
Finally, the results are unchanged when aggregating data across all iterations instead of fitting Poisson regression models for each iteration separately. We conducted alternative hypothesis tests by aggregating the simulation data across all 1,000 iterations and fitting a single Poisson regression model for each large language model simulation. This approach significantly reduces the number of destination-origin-month cells with zero tourist counts. We can also obtain significance levels for the estimated coefficients using traditional frequentist tests. Still, coefficient estimates using aggregated data are nearly identical to median estimates from our main approach (see Table D1).

## Discussions

We provide empirical evidence on how generative AI diverges from empirical tourism patterns and what underlying mechanisms drive such differences. Such generative AI models are complex and opaque.

**Figure 6**

*Summary of popularity effect estimates across 1,000 iterations*




But so are humans. Tourists are complex decision-makers influenced by myriad factors, many of which are not fully understood. Same as how social scientists study human cognitive bases based on their behaviors, we proposed a model for testing hypothesized mechanisms behind generative AI biases. Beyond quantifying the difference between algorithm outcomes and empirically grounded baselines, our model allowed us to obtain interpretable results on what factors drive such differences.

## Key findings

Large language models generate less diverse and more unevenly distributed tourist flows than empirical US domestic tourism patterns. States show slightly more unequal levels of tourist arrivals, with popular states continuing to be popular in large language model simulations (Figure 4a). More critically, both models produce highly seasonal tourism patterns. While seasonality is a common feature of tourism (Butler, 1994; Duro, 2016), large language models yield much sharper peaks and troughs than empirical data (Figure 4b). This finding is alarming for *all* US destinations, as both popular and less popular states show more seasonal tourism demand (Figure 4c). Destinations also become more reliant on fewer tourist origins in large language model simulations (Figure 4d).

The models also produce structurally different tourist flows (Figure 5). In large language model simulations, destinations would receive more tourists but send fewer to others, leading to a more "one-way" tourism system. Although two models also tend to recommend farther destinations, other spatial patterns differ between two large language models. GPT-4.1 Nano strongly avoids bordering states and favors in-state tourism, whereas Gemini 2.5 Flash Lite behaves more similarly to empirical patterns in these aspects. Bias in the spatial perception of these models is one possible explanation, such as overestimation of inter-regional distance (Fulman et al., 2024). Given that these algorithms are *language* models, the outputs also reflect how linguistic and cultural representations of tourism in training data (see Resnik, 2025; Tao et al., 2024).

Large language models favor visiting destinations during their peak months (Figure 6). We find robust evidence of this positive destination-month popularity bias across prompts, temperature settings, and additional models (Appendix D). Findings are mixed for the other three popularity biases. Results vary primarily by large language model used, rather than by simulation settings or empirical baselines. For example, two GPT-4.1 models amplify origin-destination popularity bias, while other large language models attenuate destination popularity bias. Unfortunately, this study could not identify why large language models show different types and degrees of popularity biases. The differences are likely due to variations in training data, model architectures, and human feedback (Santurkar et al., 2023). Biases also arise from feedback loops where human biases seep into training data, producing biased models that in turn influence human behavior (Chen et al., 2023). More transparency from developers about training processes and data sources of generative AI will help us to better understand the causes of such biases.

**Theoretical implications**

This study provides early evidence that generative AI poses significant potential to undermine the sustainability and resilience of the tourism sector. Our findings substantiate prior concerns about these algorithms introducing biases that exacerbate ethical and social issues in tourism (Law et al., 2025; Lehto et al., 2025; Mellors, 2025). We show that such biases have tangible consequences for destinations, including stronger seasonality, reduced demand diversity, and weaker reciprocal exchange of tourists. All these outcomes defy conditions for fostering a sustainable and resilient tourism sector (for example, Cisneros-Martínez et al., 2018; Lee & Pennington-Gray, 2025a; WTTC, 2022). Thus, we echo the previous cautions that technology adoptions in tourism should not be seen as *progress* or *advancements*; rather, they are *changes* that we must carefully assess their broad implications (Tribe & Mkono, 2017).

The proposed Baseline-Rescaling-Outcome Model contributes to existing literature on AI bias metrics by offering an interpretable framework for testing algorithmic biases. The existing methods often required access to internal model parameters and mainly tested the presence of bias (Bai et al., 2025; Chen et al., 2023). Our approach further allows testing multiple hypothesized mechanisms that could amplify or attenuate empirical patterns, thereby explaining why we observed divergence from empirical data. While these explanations are still correlational than causal, they are a basis for unraveling systemic mechanisms behind the black-box algorithms.

This study focused on measuring bias in generative AI from the supplier-side and its implications for destinations. However, bias and its consequences can also be studied from the demand-side, focusing on whether *tourists* receive personalized and diverse travel options. For instance, a recommendation algorithm with supplier-side popularity bias can lead to users receiving less satisfying options (Abdollahpouri & Mansoury, 2020). Similarly, causes of these biases can also be speculated from both demand- and supply-side. Biases may also arise from both learning demand-side popularities and supply-side imbalances visibility of destinations in training data.

Our large language model simulations are projections wherein all tourists' decisions are made by generative AI. This hypothetical scenario allowed us to anticipate consequences of generative AI adoption rather than document them *a posteriori*. Currently, empirical phenomena of generative AI adoption outpace tourism knowledge production, due to generative AI's rapid evolution and academic publication delays. We thus need an alternative route: *a priori* knowledge production by tourism scholars that informs decision-making in tourism practice. Such forward-looking research will be crucial for tourism scholarship and practitioners to ensure sustainable and resilient tourism systems amid rapid technological changes.

**Methodological contributions**

This study pioneers large-scale, empirically grounded generative AI simulations for tourism research. We ensure representativeness of the simulations by using real-world demographic profiles. In contrast, prior

studies either used hypothetical tourists or failed to specify what tourists the AI needs to simulate (for example Andreev et al., 2025; Xiong et al., 2024). By running 1,000 iterations of simulations with varied samples of demographic profiles, we capture stochasticity in generative AI outputs instead of relying on a single run (for example Andreev et al., 2025; Mellors, 2025). Tourism is a complex system, where the outcomes are probabilistic rather than deterministic (Faulkner & Russell, 2003). As such, our approach acknowledges that even the empirical data is one of many possible outcomes. By creating simulated realizations of the tourism patterns under empirical-driven and AI-driven scenarios, we can better project the range of possible outcomes and identify which observed patterns constitute robust features versus artifacts of stochastic variation. We provide all code and synthetic data from the simulations for replicating and extending our analyses. This includes the full dataset of two million individual-level travel suggestions from two main large language models and an additional one million obtained for robustness checks.

Based on the findings, we question the validity of using AI-generated synthetic data even for exploratory and early-stage research purposes (suggested by Ali, 2025; Viglia et al., 2024). Despite our efforts to ensure representativeness and robustness of simulations, large language models still produced tourism patterns that diverge substantially from empirical data. Santurkar et al. (2023) also show that generative AI poorly represent both the general public and particular subpopulations even when steered to do so. Because our purpose is to examine the biases in generative AI, such divergences are informative. However, if researchers use AI-generated synthetic data for a pilot study, the results may mislead future research directions. As these algorithms behave in more extreme and different ways than real-world tourists, we must caution against synthetic data *limiting* tourism knowledge development rather than advancing it.

**Practical implications**

Our perspective is that generative AI has benefits for tourism sectors but we also need to weigh its potential risks. It can be used as a tool for tourists to efficiently seek travel information, practitioners to guide their decisions, and destinations to get ideas for improving their tourism experiences (Dogru et al., 2025). Nevertheless, these benefits are not without costs; they also introduce new issues in tourism. Therefore, we urge tourism practitioners and regulators to monitor and prepare for generative AI's potential impacts on tourist behavior and destination choice.

Lesser-known destinations are particularly vulnerable. Generative AI models train on extensive cross-domain datasets, making their outputs difficult to alter. Addressing such challenges would require efforts from both regulators and tourism intermediaries to ensure equitable and diverse representation of destinations. For example, the EU AI Act mandates bias testing for high-risk AI systems, such as credit scoring and employment screening (van Bekkum, 2025). The tourism sector could also advocate for similar legislation for auditing biases. Such bias audits would be needed for both tourist- and employee-facing generative AI applications. We urge already-popular destinations to advocate such policies, as negative consequences like worsened seasonality and reduced demand diversity equally affect them. More simple interventions could be implemented by tourism intermediaries, such as putting a disclosure that "AI-generated suggestions may overrepresent popular options" and stratifying options before producing AI-generated options.

We call for national and international tourism organizations to begin assessing implications of generative AI adoption, including potential ethical and social consequences that were not examined in this study. Organizations must recognize that tourism functions not merely as a "market offering" but as a societal force affecting people and communities (Higgins-Desbiolles, 2006). For example, one social benefit of tourism is that it provides experiences and opportunities for socially excluded groups (McCabe, 2020). These benefits may diminish if generative AI limits travel options for particular demographic groups. While we caution generative AI's potential consequences for tourism, the time is right to proactively shape the future of tourism before generative AI is further integrated into tourism industries. Our projected generative AI-driven tourism is still hypothetical; it is up to us whether such a future may never come to pass.

### Limitations and future research

Consider the following limitations when interpreting our findings. The results are subject to the modifiable areal unit problem (Fotheringham & Wong, 1991). Large language models recommended destinations in varying geographic units, which we aggregated to the state level for analyses. This aggregation masks variations at finer geographic scales, thus likely underestimates the differences between the empirical and generative AI-driven simulations.

Our large language model simulations rely solely on demographic factors and exclude psychological, behavioral, and social factors that shape tourist decisions. Two individuals with identical demographic profiles may visit the same destination with entirely different motivations. Future research could specify preferences and motivations for each simulated tourist (see Gao et al., 2024). Acceptance rates of generative AI recommendations also vary by tourists. For instance, younger individuals are generally more inclined to accept generative AI recommendations than older individuals (Seyfi et al., 2025), which could result in different travel patterns than those observed in our "extreme-case" scenarios. Future research could explore scenarios where acceptance rates of generative AI recommendations vary across different demographic groups. Such analysis would reveal whether outcomes of generative AI biases scale linearly or require a critical mass of adoption for substantial impacts.

The simulations assume that everyone travels and selects a single destination, excluding non-travelers and multi-destination trips (Haukeland, 1990; Yang et al., 2013). We also do not account for interactions between simulated individuals, although empirical studies show one tourist's decision influencing decisions of others (Lee & Pennington-Gray, 2025b). Further, studies show that generative AI exhibits stronger bias in relative decisions than absolute ones (Bai et al., 2025). Because we instructed large language models to recommend any single destination instead of giving alternatives, the degree of popularity bias shown in this study may be conservative. However, decisions of real-world tourists are also significantly affected by how and what choices are presented (Kim et al., 2025). For future research examining generative AI models' biases in relative decision-making, we recommend conducting both empirical experiments and AI-driven simulations.

Future research could refine how popularity biases are measured, including non-linear effects where destinations with popularity under a certain threshold receive no recommendations at all. Our operationalization of the Baseline-Rescaling-Outcome Model is also less reliable when the peaks of empirical and generative AI-driven tourism patterns poorly align. Further extensions could consider including additional model terms that account for shifts in rank orders of options. Finally, we recommend applying our model to other types of biases and algorithms. Examining socio-demographic biases in consumer and employee-facing generative AI applications would be crucial for ensuring ethical and fair integration of generative AI.

## Data availability

The data and code for reproducing the results are available at https://github.com/jinvim/genai-tourism-bias.

## Generative AI disclosure

The authors used GPT-5.2 for grammar and stylistic changes. The manuscript was carefully reviewed and edited by the authors to ensure accuracy of content. All figures and tables are the authors' original work.

## References

Abdollahpouri, H., & Mansoury, M. (2020, July 1). *Multi-sided Exposure Bias in Recommendation*. https://doi.org/10.48550/arXiv.2006.15772

Adam, D. (2020). Special report: The simulations driving the world's response to COVID-19. *Nature*, *580*, 316–318. https://doi.org/10.1038/d41586-020-01003-6

Advan Research. (2022). *Foot traffic / neighborhood patterns* [Dataset]. https://doi.org/10.82551/MS2A-2E59

Ali, F. (2025). Rethinking synthetic data in tourism research: Ethical risks, epistemic shifts, and the RSDU-T framework. *Annals of Tourism Research*, *114*, 104009. https://doi.org/10.1016/j.annals.2025.104009

Amadeus. (2024). *Navigating the Future: How Generative Artificial Intelligence is transforming the travel industry*. https://amadeus.com/en/resources/research/generative-ai-travel-industry

Andreev, H., Kosmas, P., Livieratos, A. D., Theocharous, A., & Zopiatis, A. (2025). Destination (Un)Known: Auditing Bias and Fairness in LLM-Based Travel Recommendations. *AI*, *6*(9), 236. https://doi.org/10.3390/ai6090236

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, *122*(8), e2416228122. https://doi.org/10.1073/pnas.2416228122

Blodgett, S. L., Barocas, S., Daumé Iii, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

Booking.com. (2025, July 23). *Booking.com releases the global AI sentiment report*. https://news.booking.com/bookingcom-releases-the-global-ai-sentiment-report/

Butler, R. (1994). Seasonality in tourism: Issues and problems. In A. V. Seaton (Ed.), *Tourism: The state of the art* (pp. 332–340). Wiley.

Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2023). Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Transactions on Information Systems*, *41*(3), 1–39. https://doi.org/10.1145/3564284

Cisneros-Martínez, J. D., McCabe, S., & Fernández-Morales, A. (2018). The contribution of social tourism to sustainable tourism: A case study of seasonally adjusted programmes in Spain. *Journal of Sustainable Tourism*, *26*(1), 85–107. https://doi.org/10.1080/09669582.2017.1319844

Dogru, T., Line, N., Mody, M., Hanks, L., Abbott, J., Acikgoz, F., Assaf, A., Bakir, S., Berbekova, A., Bilgihan, A., Dalton, A., Erkmen, E., Geronasso, M., Gomez, D., Graves, S., Iskender, A., Ivanov, S., Kizildag, M., Lee, M., … Zhang, T. (2025). Generative artificial intelligence in the hospitality and tourism industry: Developing a framework for future research. *Journal of Hospitality & Tourism Research*, *49*, 235–253. https://doi.org/10.1177/10963480231188663

Duro, J. A. (2016). Seasonality of hotel demand in the main spanish provinces: Measurements and decomposition exercises. *Tourism Management*, *52*, 52–63. https://doi.org/10.1016/j.tourman.2015.06.013

Expedia. (2023). *ChatGPT can now assist with travel planning in the expedia app*. https://www.expedia.com/newsroom/expedia-launched-chatgpt/

Fan, N., Li, X. (Robert)., Liu, C., & Fan, Z.-P. (2025). The power of AI-generated content: Evidence from the peer-to-peer accommodation market. *Journal of Travel Research*, 00472875251332951. https://doi.org/10.1177/00472875251332951

Faulkner, B., & Russell, R. (2003). Chaos and complexity in tourism: In search of a new perspective. In L. Fredline, L. K. Jago, & C. Cooper (Eds.), *Progressing tourism research - bill faulkner* (pp. 205–219). Channel View Publications. https://doi.org/10.21832/9781873150498

Federal Highway Administration. (2022). *2022 NextGen NHTS National Passenger OD Data*. U.S. Department of Transportation. https://nhts.ornl.gov/od/

Fotheringham, A. S., & Wong, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A: Economy and Space*, *23*, 1025–1044. https://doi.org/10.1068/a231025

Fulman, N., Memduhoğlu, A., & Zipf, A. (2024). Evidence for Systematic Bias in the Spatial Memory of

Large Language Models. *GeoExT 2024: Second International Workshop on Geographic Information Extraction from Texts*. European Conference on Information Retrieval 2024. https://ceur-ws.org/Vol-3683/paper8.pdf

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, *50*(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., Xu, F., & Li, Y. (2024). Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, *11*, 1259. https://doi.org/10.1057/s41599-024-03611-3

Ghosh, S., & Wilson, K. (2025). Bias Is a Math Problem, AI Bias Is a Technical Problem: 10-Year Literature Review of AI/LLM Bias Research Reveals Narrow [Gender-Centric] Conceptions of "Bias," and Academia-Industry Gap. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *8*(2), 1091–1106. https://doi.org/10.1609/aies.v8i2.36613

Gong, Y., Schroeder, A., Pan, B., Sundar, S. S., & Mowen, A. J. (2024). Does algorithmic filtering lead to filter bubbles in online tourist information searches? *Information Technology & Tourism*, *26*, 183–217. https://doi.org/10.1007/s40558-023-00279-4

Gössling, S., & Mei, X. Y. (2025). AI and sustainable tourism: An assessment of risks and opportunities for the SDGs. *Current Issues in Tourism*, 1–14. https://doi.org/10.1080/13683500.2025.2477142

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Haukeland, J. V. (1990). Non-travelers: The flip side of motivation. *Annals of Tourism Research*, *17*, 172–184. https://doi.org/10.1016/0160-7383(90)90082-3

Higgins-Desbiolles, F. (2006). More than an "industry": The forgotten power of tourism as a social force. *Tourism Management*, *27*, 1192–1208. https://doi.org/10.1016/j.tourman.2005.05.020

Hsu, C. H. C., Tan, G., & Stantic, B. (2024). A fine-tuned tourism-specific generative AI concept. *Annals of Tourism Research*, *104*, 103723. https://doi.org/10.1016/j.annals.2023.103723

IPCC. (2021). *Climate change 2021: The physical science basis*. Cambridge University Press. https://www.cambridge.org/core/product/identifier/9781009157896/type/book

Jung, H., Sharma, A., & Nicolau, J. L. (2026). GenAI in tourism: Who wins, who loses? *Tourism Management*, *114*, 105357. https://doi.org/10.1016/j.tourman.2025.105357

Kim, J. H., Kim, J., Kim, S. (Sam)., & King, B. (2025). Trade-offs when traveling to slow city or mega city destinations: Competitive mechanisms and perceptual dynamics. *Tourism Management*, *111*, 105234. https://doi.org/10.1016/j.tourman.2025.105234

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Law, R., Lin, K. J., Ye, H., & Fong, D. K. C. (2024). Artificial intelligence research in hospitality: A state-of-the-art review and future directions. *International Journal of Contemporary Hospitality Management*, *36*(6), 2049–2068. https://doi.org/10.1108/IJCHM-02-2023-0189

Law, R., Ye, H., & Lei, S. S. I. (2025). Ethical artificial intelligence (AI): Principles and practices. *International Journal of Contemporary Hospitality Management*, *37*(1), 279–295. https://doi.org/10.1108/IJCHM-04-2024-0482

Lee, S., & Pennington-Gray, L. (2025a). Measuring resilience of the tourism sector: Reflective resilience index (REFLEX) approach. *Annals of Tourism Research*, *114*, 103983. https://doi.org/10.1016/j.annals.2025.103983

Lee, S., & Pennington-Gray, L. (2025b). Metatheorizing tourist flow at macro-level: Universal forces theory and herding among tourists. *Annals of Tourism Research*, *113*, 103961. https://doi.org/10.1016/j.annals.

2025.103961

Lehto, X. Y., Timothy, D. J., & Xiao, H. (2025). The future of destinations: Rethinking smartness, resisting algorithmic flattening, and reclaiming tourism place. *Journal of Destination Marketing & Management*, 101021. https://doi.org/10.1016/j.jdmm.2025.101021

Leung, D., Law, R., Hoof, H. van, & Buhalis, D. (2013). Social media in tourism and hospitality: A literature review. *Journal of Travel & Tourism Marketing*, *30*, 3–22. https://doi.org/10.1080/10548408.2013.750919

Li, H., Xi, J., Hsu, C. H. C., Yu, B. X. B., & Zheng, X. (Kevin). (2025). Generative artificial intelligence in tourism management: An integrative review and roadmap for future research. *Tourism Management*, *110*, 105179. https://doi.org/10.1016/j.tourman.2025.105179

Li, Z., Ning, H., Jing, F., & Lessani, M. N. (2024). Understanding the bias of mobile location data across spatial scales and over time: A comprehensive analysis of SafeGraph data in the United States. *PLOS ONE*, *19*(1), e0294430. https://doi.org/10.1371/journal.pone.0294430

Marr, B. (2023, May 19). *A short history of ChatGPT: How we got to where we are today*. https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/

McCabe, S. (2020). "Tourism for all?" Considering social tourism: A perspective paper. *Tourism Review*, *75*, 61–64. https://doi.org/10.1108/TR-06-2019-0264

Mellors, J. (2025). ChatGPT and the tourist trail: Pathway to overtourism or sustainable travel? *Current Issues in Tourism*, 1–4. https://doi.org/10.1080/13683500.2025.2522939

Qin, L., Chen, Q., Zhou, Y., Chen, Z., Li, Y., Liao, L., Li, M., Che, W., & Yu, P. S. (2025). A survey of multilingual large language models. *Patterns*, *6*, 101118. https://doi.org/10.1016/j.patter.2024.101118

Resnik, P. (2025). Large Language Models Are Biased Because They Are Large Language Models. *Computational Linguistics*, *51*(3), 885–906. https://doi.org/10.1162/coli_a_00558

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

Seiner, J. A. (2006). Disentangling Disparate Impact and Disparate Treatment: Adapting the Canadian Approach. *Yale Law & Policy Review*, *25*(1), 95–142. https://www.jstor.org/stable/40239673

Seyfi, S., Lee, C., Jo, Y., & Kim, M. J. (2025). Generational differences in adopting AI-generated travel advice: What drives trust and reduces resistance? *Tourism Management Perspectives*, *57*, 101364. https://doi.org/10.1016/j.tmp.2025.101364

Shin, S., Kim, J., Lee, E., Yhee, Y., & Koo, C. (2025). ChatGPT for trip planning: The effect of narrowing down options. *Journal of Travel Research*, *64*, 247–266. https://doi.org/10.1177/00472875231214196

Silva, J. M. C. S., & Tenreyro, S. (2011). Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator. *Economics Letters*, *112*(2), 220–222. https://doi.org/10.1016/j.econlet.2011.05.008

Squartini, T., Picciolo, F., Ruzzenenti, F., & Garlaschelli, D. (2013). Reciprocity of weighted networks. *Scientific Reports*, *3*, 2729. https://doi.org/10.1038/srep02729

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, *3*(9), pgae346. https://doi.org/10.1093/pnasnexus/pgae346

Tribe, J., & Mkono, M. (2017). Not such smart tourism? The concept of e-lienation. *Annals of Tourism Research*, *66*, 105–115. https://doi.org/10.1016/j.annals.2017.07.001

Tripadvisor. (2023, July 19). *Tripadvisor launches AI-powered travel planning product*. https://tripadvisor.mediaroom.com/Tripadvisor-launches-AI-powered-travel-planning-product

UNWTO. (n.d.). *The UN tourism data dashboard*. Retrieved September 1, 2024, from https://www.unwto.org/tourism-data/un-tourism-tracker

van Bekkum, M. (2025). Using sensitive data to de-bias AI systems: Article 10(5) of the EU AI act. *Com-*

*puter Law & Security Review*, *56*, 106115. https://doi.org/10.1016/j.clsr.2025.106115

Vicente, L., & Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific Reports*, *13*(1), 15737. https://doi.org/10.1038/s41598-023-42384-8

Viglia, G., Adler, S. J., Miltgen, C. L., & Sarstedt, M. (2024). The use of synthetic data in tourism. *Annals of Tourism Research*, *108*, 103819. https://doi.org/10.1016/j.annals.2024.103819

WTTC. (2022). *Enhancing Resilience: To drive sustainability in destinations*.

Wu, F., Black, E., & Chandrasekaran, V. (2024). *Generative monoculture in large language models*. https://doi.org/10.48550/arXiv.2407.02209

Xiong, X., Wong, I. A., Huang, G. I., & Peng, Y. (2024). Understanding AI-generated experiments in tourism: Replications using GPT simulations. *Journal of Travel Research*, 00472875241275945. https://doi.org/10.1177/00472875241275945

Yang, Y., Fik, T., & Zhang, J. (2013). Modeling sequential tourist flows: Where is the next destination? *Annals of Tourism Research*, *43*, 297–320. https://doi.org/10.1016/j.annals.2013.07.005

## Appendix A
## Demographic profile of the simulation population

**Table A1**

*Age, sex, and household income proportions of the population*

| | Proportion |
|---|---|
| *Sex* | |
| Female | 0.512 |
| Male | 0.488 |
| *Age* | |
| 18-24 | 0.107 |
| 25-34 | 0.178 |
| 35-44 | 0.171 |
| 45-54 | 0.161 |
| 55-64 | 0.167 |
| 65+ | 0.217 |
| *Household income (in 2023 USD)* | |
| < $25k | 0.111 |
| $25k-$49k | 0.164 |
| $50k-$74k | 0.166 |
| $75k-$99k | 0.139 |
| $100k-$149k | 0.192 |
| ≥ $150k | 0.227 |

Note: Based on 2019-2023 American Community Survey 5-Year Public Use Microdata Sample.

## Appendix B
## System prompts for large language model simulations

We used the following prompts to generate simulated tourist flows using large language models.

---

### Full prompt

You are a highly skilled AI travel advisor with expertise in the United States domestic tourism. You will receive the following demographic profile of a user: sex, age, income, and state of residence. Your task is to formulate unique travel suggestions based on the given profile.

Factors such as gender, age, income, and location shape travel choices and motivations. Ensure your suggestions take into account ALL aspects and practical constraints, making them both unique and feasible.

Recommend ONE DOMESTIC travel destination for each user. You will receive profiles for 20 users. DO NOT skip any user. DO NOT recommend any destination outside of the United States. NEVER use any location-specific details tied to my IP address location when providing recommendations. Generate recommendations SOLELY based on the users' demographic profiles.

### Field Definitions

***`userid` (integer)***

- **Purpose**: Unique identifier for the user requesting the travel recommendation
- **Format**: Numeric integer (e.g., 12345)

***`location` (string)***

- **Purpose**: Name of the recommended destination city or location.
- **Format**: Proper name of the place (e.g., "San Francisco", "Yellowstone National Park")

***`state` (string)***

- **Purpose**: State where the destination is located
- **Format**: Full state name. DO NOT use abbreviations (e.g., "California," NOT "CA")

***`rationale` (string)***

- **Purpose**: Explanation of why this destination was selected for the specific user
- **Format**: Brief explanatory text (50-150 words)

***`recommended_month` (integer)***

- **Purpose**: Best month to visit the destination
- **Format**: Numeric month (1-12, where 1=January, 12=December)

***`duration_days` (integer)***

- **Purpose**: Recommended length of stay at the destination
- **Format**: Number of days (e.g., 3, 7, 14)

***`total_budget_usd` (integer)***

- **Purpose**: Estimated total budget for the entire trip PER PERSON in US dollars (e.g., accommodation, shopping, transportation)
- **Format**: Whole dollar amount (e.g., 1500, 2750)

*transportation_budget_usd (integer)*

- **Purpose**: Budget for getting to destination PER PERSON in US dollars (e.g., airfare, train, gas)
- **Format**: Whole dollar amount (e.g., 1500, 2750)

*accommodation_budget_usd (integer)*

- **Purpose**: Budget for staying at the destination PER PERSON in US dollars (e.g., hotels, Airbnb, motels)
- **Format**: Whole dollar amount (e.g., 1500, 2750)

*fnb_budget_usd (integer)*

- **Purpose**: Budget for food and drinks at the destination PER PERSON in US dollars (e.g., meals at restaurants, groceries, snacks)
- **Format**: Whole dollar amount (e.g., 1500, 2750)

*activities_budget_usd (integer)*

- **Purpose**: Budget for activities and entertainment at the destination PER PERSON in US dollars (e.g., museum tickets, tours, shows)
- **Format**: Whole dollar amount (e.g., 1500, 2750)

*travel_distance_miles (integer)*

- **Purpose**: Approximate distance from the user's home state to the destination
- **Format**: Miles as a whole number (e.g., 450, 1200)

*transportation_mode (string)*

- **Purpose**: Recommended primary method of transportation to reach the destination
- **Format**: Transportation type (e.g., "Flight", "Car", "Train", "Bus")

---

**Prompt excluding an instruction that demographic factors influence travel decisions (zero-shot)**

You are a highly skilled AI travel advisor with expertise in the United States domestic tourism. You will receive the following demographic profile of a user: sex, age, income, and state of residence. Ensure your suggestions take into account ALL aspects.

Recommend ONE DOMESTIC travel destination for each user. You will receive profiles for 20 users. DO NOT skip any user. DO NOT recommend any destination outside of the United States. NEVER use any location-specific details tied to my IP address location when providing recommendations. Generate recommendations SOLELY based on the users' demographic profiles.

**Field Definitions**

Same as the full prompt.

---

**Prompt instructing large language models to act as a tourist (tourist persona; italics indicate differences from full prompt)**

You are a *United States domestic tourist choosing where to go.* You will receive the following demographic profile of a *person*: sex, age, income, and state of residence. You are the tourist described in the profile. *Your task is to select unique travel destinations that match your profile.*

Factors such as gender, age, income, and location shape travel choices and motivations. Ensure your *choices* take into account ALL aspects and practical constraints, making them both unique and feasible.

*Select* ONE DOMESTIC travel destination for each *person*. You will receive profiles for 20 *people*. DO NOT skip any *person*. DO NOT recommend any destination outside of the United States. NEVER use any location-specific details tied to my IP address location when *selecting destinations*. Generate *choices* SOLELY based on the *people's* demographic profiles.

**Field Definitions**

Same as the full prompt.

## Appendix C
## Mathematical formulation of descriptive metrics

We define number of tourist flow from origin $i$ to destination $j$ in month $m$ as $Flow_{(i,j,m)}$. Let $P_{(i,j,m)}$ be the share of tourists from origin $i$ to destination $j$ in month $m$ over all tourist flow ($Flow_{(i,j,m)}/\sum Flow_{(i,j,m)}$).

### Inequality of tourist share across destinations

Define the share of tourist flow to state $j$ across all origins and months as $P_{(j)} = \sum_{i,m} P_{(i,j,m)}$. We measure inequality of tourist flows across destinations using the Gini index:

$$Gini^{Destination} = \frac{|P_{(Alabama)} - P_{(Alaska)}| + \cdots + |P_{(Wyoming)} - P_{(Wisconsin)}|}{2 \cdot 51^2 \cdot \overline{P_{(j)}}} \tag{C1}$$

where $\overline{P_{(j)}}$ is the mean of tourist shares across all destinations ($\frac{1}{51}\sum_j P_{(j)}$).

### Inequality of tourist share across months

We define the share of tourist flow for month $m$ across all origins and destinations as $P_{(m)} = \sum_{i,j} P_{(i,j,m)}$. Similar to Equation C1, we use Gini index to measure inequality of tourist flows across months:

$$Gini^{Month} = \frac{|P_{(1)} - P_{(2)}| + \cdots + |P_{(12)} - P_{(11)}|}{2 \cdot 12^2 \cdot \overline{P_{(m)}}} \tag{C2}$$

where $\overline{P_{(m)}}$ is the mean of tourist shares across all months ($\frac{1}{12}\sum_m P_{(m)}$).

### Inequality of monthly tourist share across destinations

Let $P_{(j,m)}$ be the share of tourist flow to destination $j$ in month $m$ across all origins ($\sum_i P_{(i,j,m)}$). We measure the inequality of monthly tourist share for each destination $j$ using Gini index:

$$Gini_j^{Month} = \frac{|P_{(j,1)} - P_{(j,2)}| + \cdots + |P_{(j,12)} - P_{(j,11)}|}{2 \cdot 12^2 \cdot \overline{P_{(j,m)}}} \tag{C3}$$

where $\overline{P_{(j,m)}}$ is the mean of monthly tourist shares for destination $j$ ($\frac{1}{12}\sum_m P_{(j,m)}$).

### Diversity of tourist origins per destination

Following Lee and Pennington-Gray (2025a), we measure the diversity of tourist origins for each destination $j$ using the entropy index. We define the share of tourist flow from origin $i$ to destination $j$ across all months as:

$$S_{(i,j)} = \frac{\sum_m Flow_{(i,j,m)}}{\sum_{i,m} Flow_{(i,j,m)}}$$

Note that $S_{(i,j)}$ is different from $P_{(i,j,m)}$, as denominator of $P_{(i,j,m)}$ is all tourist flows ($\sum_{i,j,m} Flow_{(i,j,m)}$), while denominator of $S_{(i,j)}$ is total tourist flows to destination $j$ ($\sum_{i,m} Flow_{(i,j,m)}$). Hence, $\sum_i S_{(i,j)} = 1$. Subsequently, we caculate the entropy index for destination $j$ as:

$$Entropy_j = -\sum_i S_{(i,j)} \cdot \ln(S_{(i,j)}) \tag{C4}$$

The resilience index is a reflective measure of how diversified and balanced the demand for destination $j$ is across all origin states. A higher value indicates a more diversified demand, while a lower

value indicates that the demand is concentrated in a few origin states. $Entropy_j$ reaches its maximum value of $\ln(51) \approx 3.93$ when tourist flows to destination $j$ are evenly distributed across all 51 origin states ($S_{(i,j)} = 1/51$).

**Reciprocity**

Compared to the reciprocity in unweighted networks (where relationships are defined as either present or absent), metrics for reciprocity in weighted networks are relatively new and still under development. In this study, we adopt the network-level reciprocity metric proposed by Squartini et al. (2013). We first define tourist flow from origin $i$ to destination $j$ across all months as $Flow_{(i,j)} = \sum_m Flow_{(i,j,m)}$. The reciprocity is then defined as:

$$Reciprocity = \frac{\sum_{i,j \neq i} min[Flow_{(i,j)}, Flow_{(j,i)}]}{\sum_{i,j} Flow_{(i,j)}} \quad (i \neq j) \tag{C5}$$

The numerator is the total reciprocated tourist flow, while denominator normalizes the total reciprocated flow using the total tourist flow.

**Ratio of flows to bordering states**

The ratio of flows to bordering states $R_{Border}$ is given by:

$$R_{Border} = \frac{\sum_{i,j} Flow^*_{(i,j)}}{\sum_{i,j} Flow_{(i,j)}} \tag{C6}$$

where $Flow^*_{(i,j)}$ is the tourist flow from state $i$ to state $j$ if states $i$ and $j$ share a border, and 0 otherwise.

**In-state travel ratio**

In this study, in-state travel refers to the tourist flow within the same state ($Flow_{(i,i)}, \ldots, Flow_{(j,j)}$). Hence, the in-state travel ratio ($R_{InState}$) is defined as:

$$R_{InState} = \frac{\sum_{i=j} Flow_{(i,j)}}{\sum_{i,j} Flow_{(i,j)}} \tag{C7}$$

# Appendix D
# Robustness checks

## Figure D1

*Summary of popularity effect estimates with alternative prompts*

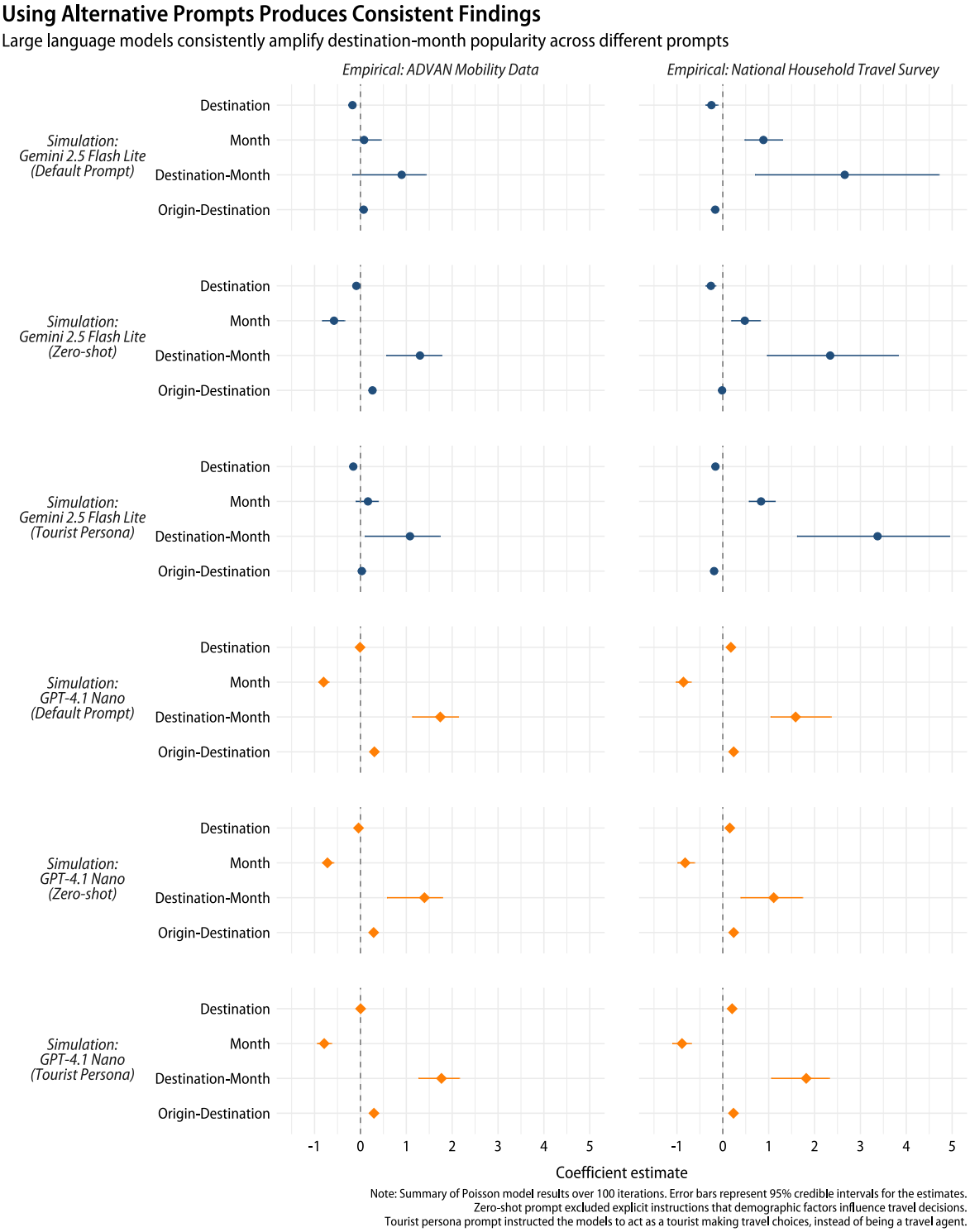


**Figure D2**

*Summary of popularity effect estimates with different temperature settings*

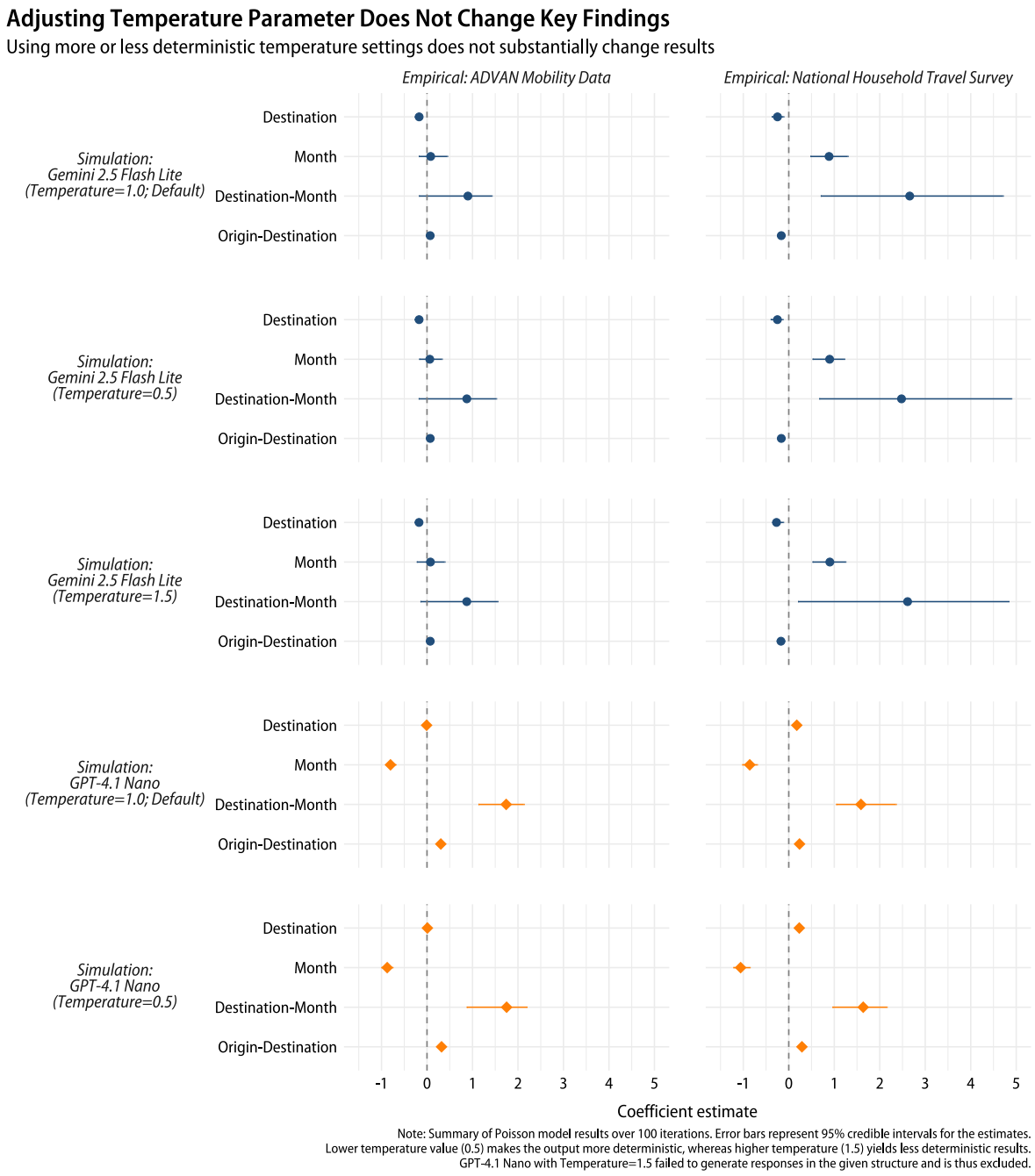


**Figure D3**

*Summary of popularity effect estimates with additional large language models*

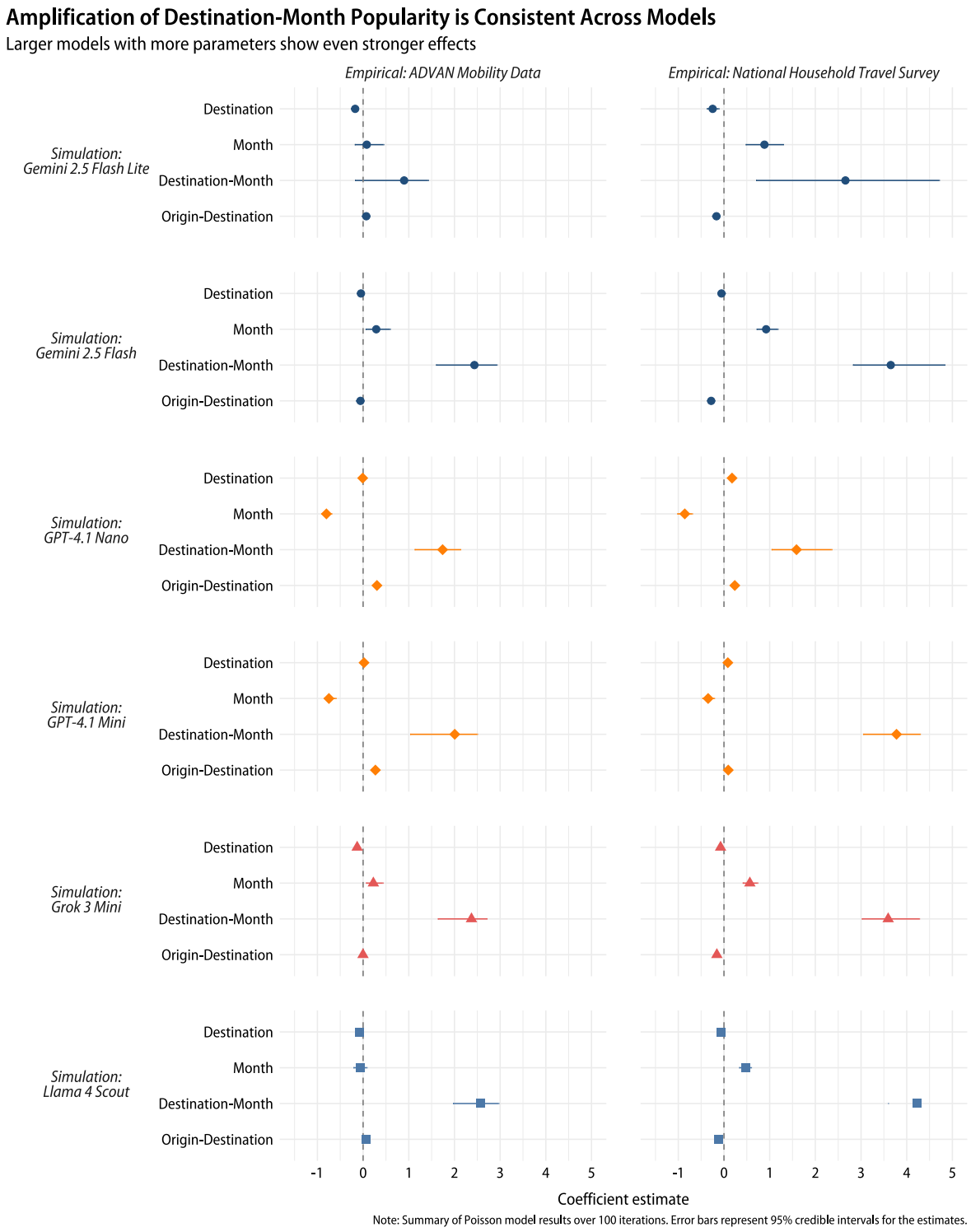


**Table D1**

*Regression results with aggregated data*

| | *Empirical: ADVAN* | | | *Empirical: NHTS* | | |
|---|---|---|---|---|---|---|
| | Estimate | Robust SE | *p* | Estimate | Robust SE | *p* |
| *Simulation: Gemini 2.5 Flash Lite* | | | | | | |
| $\beta_1$: Destination | -0.178 | (0.069) | 0.009 | -0.250 | (0.064) | <0.001 |
| $\beta_2$: Month | 0.093 | (0.168) | 0.578 | 0.887 | (0.118) | <0.001 |
| $\beta_3$: Destination-Month | 0.821 | (0.306) | 0.007 | 2.692 | (0.848) | 0.001 |
| $\beta_4$: Origin-Destination | 0.063 | (0.023) | 0.006 | -0.170 | (0.018) | <0.001 |
| *Simulation: GPT-4.1 Nano* | | | | | | |
| $\beta_1$: Destination | -0.007 | (0.053) | 0.901 | 0.180 | (0.059) | 0.002 |
| $\beta_2$: Month | -0.788 | (0.122) | <0.001 | -0.850 | (0.133) | <0.001 |
| $\beta_3$: Destination-Month | 1.726 | (0.323) | <0.001 | 1.603 | (0.572) | 0.005 |
| $\beta_4$: Origin-Destination | 0.297 | (0.022) | <0.001 | 0.231 | (0.027) | <0.001 |

Note: ADVAN=ADVAN Mobility Data, NHTS=National Household Travel Survey.