Large Language Models Yield Unsustainable Tourist Flows

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

This study quantifies the impact of generative AI on tourist flow using scenario-based projection models. We simulate one million US domestic tourists using Gemini 2.5 Flash and GPT 4.1 Nano. Their tourism patterns are compared to the null model and empirical-based simulations. Large language models generate tourist flows that are more seasonal, more unequal, and less reciprocal than empirical-based simulations. Geographic patterns, like mean travel distance and preference for neighboring states and intra-state destinations, vary by model. Findings show that the widespread adoption of generative AI can undermine the sustainability and resilience of tourism systems. We urge tourism scholars and practitioners to proactively assess the consequences of adopting generative AI in tourism.

Keywords: tourist flow network, social network analysis, generative AI impact, large language models, network simulation, scenario-based projection models

Introduction

A day would come when we no longer need to plan our trips. Or at least, that is what companies like OpenAI, Google, and Anthropic envision—easily "outsourcing" travel decisions to AI agents (Dawes, 2024). From writing code to composing music, generative AI is transforming how we do many things (Marr, 2023). Tourism is no exception in this regard. The majority of travelers are already using generative AI to plan their trips (Booking.com, 2025). In response, major online travel agencies like Expedia and TripAdvisor are integrating generative AI into their platforms (Expedia, 2023; Tripadvisor, 2023).

As tourism researchers, we pose a critical question: *What consequences await destinations if more tourists rely on generative AI to guide their travel decisions?* Some tourism and hospitality experts warn that generative AI can suppress destinations' uniqueness and undermine the sustainability of tourism (Dogru et al., 2025; Lehto et al., 2025). Yet, we lack evidence to substantiate such concerns (Gössling & Mei, 2025; Mellors, 2025). This knowledge gap leaves tourism destinations and businesses uncertain about how to prepare for generative AI's potential impacts, or to determine whether such preparation is even necessary.

In this study, we provide empirical insights into how this rapid adoption of generative AI may reshape tourism patterns. We do so using scenario-based projections simulating decisions of one million US

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

domestic tourists. Scenarios include tourists making random choices, adhering to current patterns, and relying exclusively on large language models. The last scenario offers a "worst-case" benchmark for assessing whether the concerns about generative AI's impacts are valid.

We show that compared to empirical-based simulations, large language models tend to concentrate visits during peak seasons and favor popular destinations. Large language models also produce tourism patterns where popular states primarily receive tourists while reciprocating fewer. However, these models show different tendencies regarding geographic patterns of tourist flows, like mean travel distance and preference for neighboring states and intra-state travel. We conclude that generative AI has the potential to undermine the sustainability and resilience of the tourism sector. While intuitive, this study offers valuable evidence quantifying how generative AI may reshape tourism patterns. These findings show the need for tourism scholars and practitioners to assess and prepare for generative AI's impact on the tourism system.

Background

Generative AI and its impacts on tourism

AI has become a catch-all term that encompasses various technologies aimed at simulating human intelligence. Adding to the confusion, tourism literature often discusses AI alongside other technologies like robotics, augmented reality, and virtual reality (Gössling & Mei, 2025). This study specifically focuses on generative AI and its implications for tourism. Trained to recognize patterns in vast data, these models can produce content such as text, images, or videos (see Epstein et al., 2023). We avoid *chatbots* because they represent just one way generative AI is used. Likewise, we do not use brand-specific terms like *ChatGPT* when discussing generative AI in general.

Despite widespread adoption of generative AI across industries, empirical studies on its impact on tourism are scarce (Hsu et al., 2024). Tourism and hospitality sector has seen rapid integration of generative AI tools, sparking scholarly interest. Studies primarily looked at who, when, and why tourists and practitioners adopt generative AI (see recent reviews by Gössling & Mei, 2025; H. Li et al., 2025). This gap calls for more research providing evidence for projecting how generative AI will shift tourism's trajectory (Law et al., 2025; Mellors, 2025).

Preliminary works explore both positive and negative implications of generative AI in tourism. One perspective is that generative AI brings benefits to tourists and businesses. For tourists, generative AI reduces cognitive load during travel planning, thereby increasing visit intentions and decision satisfaction (Shin et al., 2025). Businesses can also benefit by using generative AI to assist in marketing and content creation (Fan et al., 2025). However, generative AI adoption in tourism also raises concerns about its negative impacts. Lehto et al. (2025) warned against "algorithmic flattening" that erases destination uniqueness. Generative AI may undermine the sustainability of tourism by exacerbating over-tourism (Mellors, 2025). Some conceptual works have examined both positive and negative aspects of generative AI adoption in tourism, concerning value co-creation and co-destruction (e.g., Dogru et al., 2025; Grundner & Neuhofer, 2021).

How AI biases shape tourism

Regardless of whether the study focuses on positive or negative impacts, bias in generative AI is a common concern [refs]. This mirrors previous discussions about biases in other technologies and their impacts on tourism, such as search engines, social media, or smart technologies (Gong et al., 2024; Leung et al., 2013; Pan et al., 2021). Consequences of biases in AI is that they lead to behavioral shifts to users relying on AI for decision-making, which reproduce or amplify biases existing of humans (Kordzadeh & Ghasemaghaei, 2022; Vicente & Matute, 2023).

Many studies on biases in algorithms focus on social biases, specifically racial and gender stereotypes (see Ghosh & Wilson, 2025; Kordzadeh & Ghasemaghaei, 2022). Social biases in AI and their ethical

implications are also implicit in tourism and hospitality research. For example, Law et al. (2025) highlight ethical challenges of AI adoption in tourism and hospitality sectors, urging inclusiveness of AI adoption with reduced "biased and discriminatory actions" (p.287). Hsu et al. (2024) suggests fine-tuning generative AI with tourism-specific data, noting that such models "could perpetuate stereotypes and result in discrimination" (p.2). Viglia et al. (2024) discuss using AI-generated data for tourism research, giving a specific example of bias against AI generating racist content.

Beyond ethical concerns, AI biases could undermine the sustainability and resilience of tourism sector. Several schloars mention popularity biases in generative AI as a potential threat, where algorithms favor popular options while underrepresenting less popular ones (Law et al., 2024; Lehto et al., 2025; Mellors, 2025). This bias is of practical concern because it can exacerbate challenges like over-tourism for popular destinations, while less popular destinations struggle to sustain their tourism sector.

Stiil, AI's social and popularity biases in tourism remain as concerns, with little empirical evidence. We can only speculate that generative AI produce such biases in tourism context. Generally, these models still exhibit stereotype biases like humans and show less diversity in outputs compared to human counterparts (see Abdollahpouri & Mansoury, 2020; Bai et al., 2025; Wu et al., 2024). But such biases can have positive consequences. For example, reduced diversity in generative AI outputs are beneficial in programming task, as correct and efficient solutions are being favored. Without empirical evidence, we cannot know in what ways and to what extent biases would be benefical or harmful for the tourism sector.

Defining and measuring biases in generative AI

First we need to define *bias* to understand generative AI's biases in tourism and their implications for the sector. Works on AI biases in tourism and hospitality are often vague about what is meant by these algorithms being *biased*. This issue is not unique to tourism scholarship; broader AI bias literature have also been criticized for lacking explicit definition of bias (Ghosh & Wilson, 2025). Such ambiguity in defining bias results in discrepancies between the concerns raised and the empirical evidence provided (Blodgett et al., 2020). Although, AI biases is often undefined or vauguly defined even outside of academia. For example, the EU's AI Act that mandates AI providers to assess and mitigate biases does not specify what constitutes bias (van Bekkum, 2025).

Definition of bias varies not only across disciplines but also by the purpose of examining bias. When the focus is on stereotyping, studies define bias as *act of* unjustified association between social groups and attributes (for example Bai et al., 2025). This definition is inhereted from social psychology literature on implicit asssociations (Greenwald et al., 1998). Studies examining ethical implications of AI biases often define bias in terms of *outcomes or treatment*: unequal allocation of resources or unfavorable representation of social groups (for example Blodgett et al., 2020; Gallegos et al., 2024; Kordzadeh & Ghasemaghaei, 2022). Such definitions are analogous to legal definition of discrimination, a behaviroal outcome of biases, as disparate impacts or treatement (Seiner, 2006). Others distinguish bias from harm, where bias is defined as having a particular inclination or deviation of model outputs from empirical data (for example Ghosh & Wilson, 2025; Wu et al., 2024). In this more netural definition of bias that are closely tied to statistical and computer science, bias is considered unavoidable for generalization and because data itself reflects biases of the real world (Chen et al., 2023; Ghosh & Wilson, 2025).

Even with a clear definition of bias, measuring biases in generative AI is challenging. Major obstacle is proprietary and closed-source nature of these models Ali (2025) proposes using tools such as IBM's AI Fairness 360 or Google's What-If toolkits to asses biasses in AI generated data for tourism reseachers. These bias assesment tools assume that one has access to the data and internals of the AI models. For commonly used generative AI models, however, neither the training data nor the model itself is accessible to end users and researchers (Gallegos et al., 2024). Even with access to the model, recent generative AI models go through value alignment and safety tuning stage to avoid explicit biases (Santurkar et al., 2023). Thus,

generative AI models tend to show implicit biases that are harder to detect (Bai et al., 2025). Existing methods for aessing biases in AI are also designed to quantify one type of bias at a time (see review of bias metrics by Gallegos et al., 2024; Kordzadeh & Ghasemaghaei, 2022). For example, a tourism recommendation algorithm may exhibit racial stereotype bias and popularity bias simultaneously. But current metrics do not allow for measuring multiple biases at once, which limits diagnoses as models become more complex and their biases become more subtle.

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

We propose a Baseline-Rescaling-Outcome Model to test algorithmic biases in tourism. Bias in this model is defined as a *systematic deviation of algorithmic outputs from empirically grounded expectations of tourism phenomena.* This definition separates bias from harm (Ghosh & Wilson, 2025), as what is considered harmful depends on who is being affected (Blodgett et al., 2020). Consider tourism as a system involving multiple stakeholders (Leiper, 1979). If an algorithm systematically favors one destination over another, such bias is benefical to the favored but harmful to the rest. Even for the destination being favored, the tourism sector in that destination could benefit while locals suffer from over-tourism.

Our model tests bias by modeling the divergence between algorithmic *outcome* and empirical *baseline* as a function of *rescaling* factors representing hypothesized bias mechanisms (Figure 1). Compared to existing methods for assessing biases in AI, our approach has three advantages. The model assume that the model and its modeling process are inaccessible, which allows testing biases in proprietary and closed-source models. This assumption is no different from how social scientists study human biases. We cannot directly observe the biases in cognitive processes, but we can infer countious or uncontious biases from behavioral outcomes such as speed of responses and error rates (Greenwald et al., 1998). Our approach is also flexible. As long as we can derive empirically grounded expectations, rescaling factors, and algorithmic outputs, we can test biases in algorithms beyond generative AI and outside of tourism context. Finally, the rescaling factors allows interpretable diagnosis of biases, ideal for translating findings into practical suggestions for mitigating biases. Although, we caution that the model is not, nor aiming to be, a theory of how biases arise in algorithms. Biases in algorithms are *caused by* complex socio-technical processes that includes biases in data generating processes, model architectures, and human feedback (Santurkar et al., 2023; Viglia et al., 2024). Because we often lack access to these processes and due to uninterpretable nature of some algorithms, we can only speculate on what causes biases (Bai et al., 2025).

Outcome: Projected tourism patterns under complete reliance on the algorithm

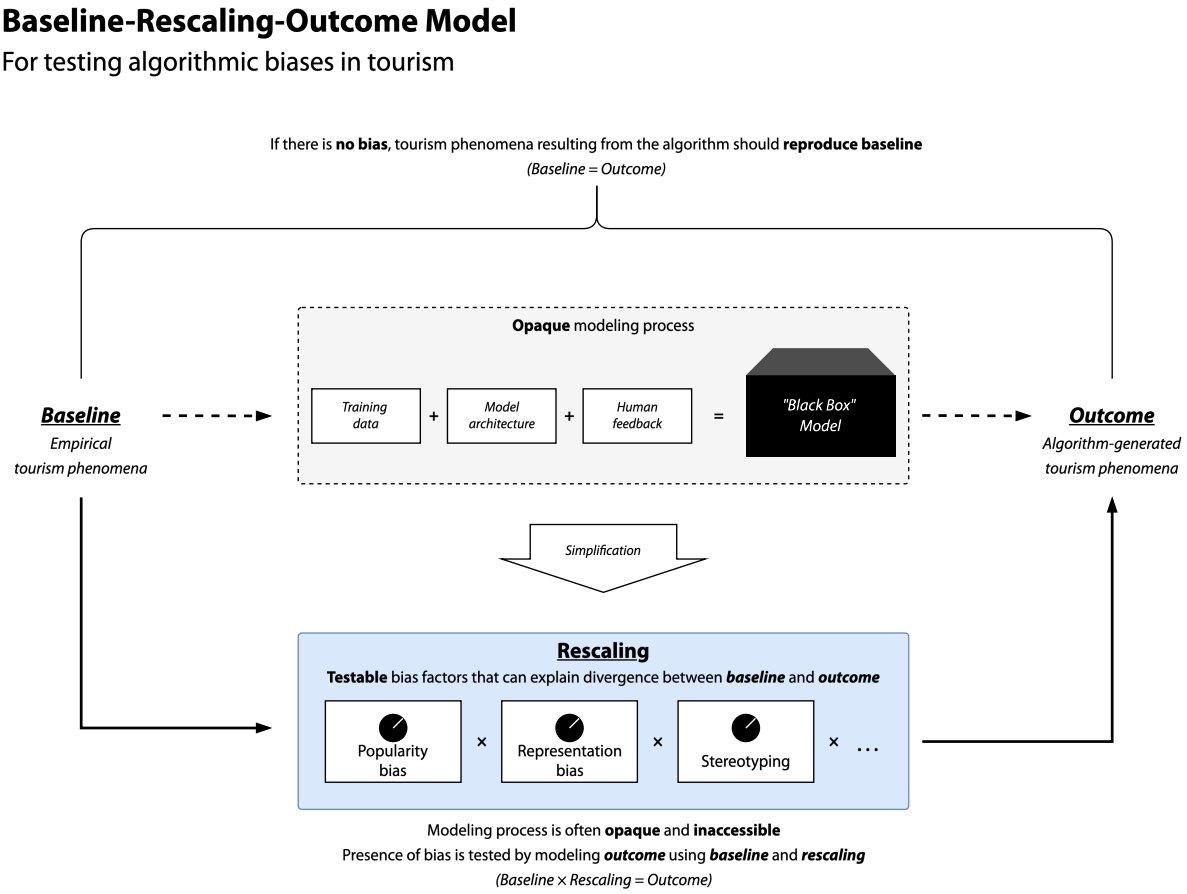
Given that we assume no access to the modeling process, output of the algorithm is the only means to assess the bias. Specifically, we use outcome of when *all* decisions are made by the algorithm under study. This follows the practice of scenario-based projection models that predict outcomes under specified conditions (Runge et al., 2024). Examples include Shared Socioeconomic Pathways for climate trajectories (IPCC, 2021) and COVID-19 diffusion models (Adam, 2020). These models contain "worst-case" scenarios that project the most extreme outcomes: climate projections without emission reductions or infection rates without government interventions. Though unrealistic, these extreme scenarios serve as benchmarks for assessing whether action is necessary. We apply the same logic to studying algorithmic biases. For example, if we want to test whether generative AI for tourist recommendation system is biased, we project a scenario where all tourists rely on the generative AI. Simiarly, biases in hiring algorithms for tourism firm can be tested by projecting a scenario where all hiring decisions are made by the algorithm.

Baseline: Empirically grounded expectations of tourism phenomena

But without a baseline, we cannot know whether the outputs are biased. Consider a situation where an algorithm has a four-in-ten chance of suggesting US domestic tourists to visit San Francisco. Could we

Figure 1

Baseline-Rescaling-Outcome Model



say that this algorithm has a systematic tendency to favor San Francisco as a tourism destination? We argue that to answer this question, one first need an empirical benchmark of what would "unbiased" suggestions look like. If four in ten Americans visit San Francisco, then the algorithm's tendency 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. Santurkar et al. (2023) implement a similar approach to assess representation biases in large language models by comparing empirical distribution of public opinions with model-generated opinions. Abdollahpouri and Mansoury (2020) similarly used real-world music and movie ratings to assess popularity biases by comparing empirical and algorithm-produced distributions of popularities.

Rescaling: Testable factors for hypothesized bias mechanisms

Rescaling is the final component of the model that tests specific mechanisms that produce biases. We do so by reproducing the algorithm outcome using the baseline and a set of rescaling factors. Under the null condition of no bias, the algorothm should reproduce the baseline hence the rescaling factors are unnecessary. If the algorithm deviates from the baseline, we can test whether such deviation is systematic by explaining the divergence using hypothesized mechanisms. This approach yields interpretable tests of specific bias in algorithms, going beyond simply measuring the degree of divergence between baseline and

algorithm output. The direction of bias is another important aspect that our model can capture. Since we adopted a neutral definition of bias, it is possible for an algorithm to exhibit bises that reduce empricially-observed asymmetries (Ghosh & Wilson, 2025). For example, algorithms could be designed to favor less popular destinations and off-peak seasons, diversifying the tourism demand across destinations and time. Our model can test for such biases as negative rescaling factors, allowing us to explain algorithm outputs as mixtures of amplification and attenuation of empirical patterns.

Study design

Empirical framework for testing popularity biases in large language model travel suggestions

This study proposes that LLMs have systematic biases in their suggestions, favoring popular tourism patterns while attenuating less popular ones. This study defines algorithmic bias as systematic deviation from expected real-world patterns. Our empirical testing framework builds on this definition. We test algorithmic bias in large language models' travel suggestions by modeling the divergence from the empirical baseline as a function of popularity factors.

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 $(Flow_{(i,j,m)}/\sum Flow_{(i,j,m)})$. Under the null condition that large language models produce "unbiased" travel suggestions, we expect:

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

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 total number of LLM-simulated 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 LLMs can perfectly replicate empirical tourism patterns $(Baseline_{(i,j,m)})$.

The focus of this study is on *popularity bias*, a specific instance of algorithmic bias where popular options are excessively favored and less popular ones gets underrepresented (Chen et al., 2023). We account for the following four types of popularity. The first two mesaure overall popularities of destinations and months independently:

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

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.

While these two factors account for destination and month popularities independently, bias may also exist in specific combinations of destinations and months. 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 compared to what would be expected if origin and destination popularities were independent.

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

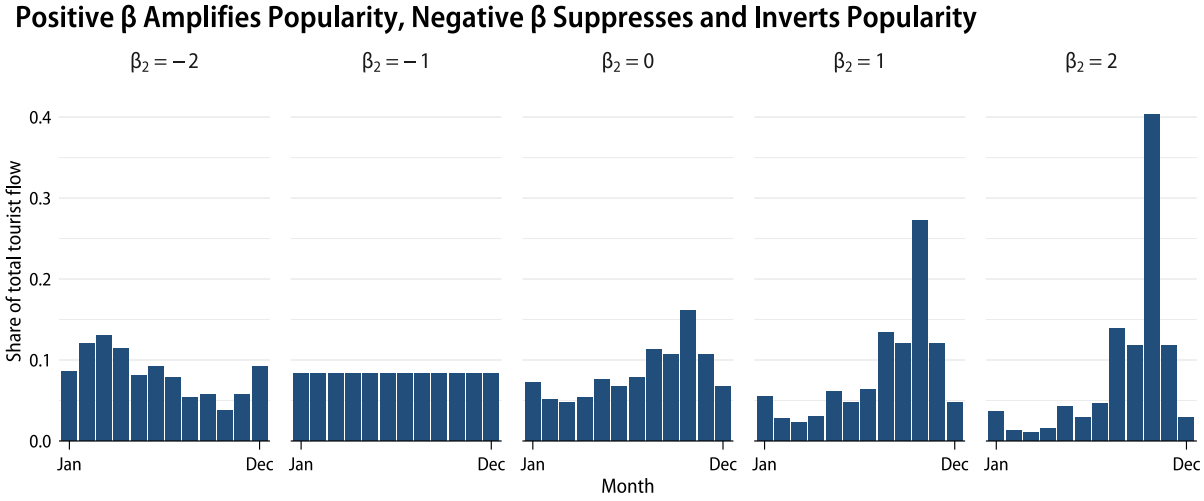
$$\frac{Flow^{LLM}_{(i,j,m)}}{Baseline_{(i,j,m)}} = (D_j)^{\beta_1} \cdot (M_m)^{\beta_2} \cdot (DM_{j,m})^{\beta_3} \cdot (OD_{i,j})^{\beta_4} \tag{5}$$

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 2 illustrate how different β_2 values (month popularity rescaling) affect the distribution of 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. The model isolate the biases toward specific combinations because $DM_{j,m}$ and $OD_{i,j}$ measure popularities beyond independent popularities of a single factor. Meaning, if we generate tourist flows with only the destination-month popularity bias ($\beta_3 \neq 0$), the resulting data would show no changes in overall destination or month popularity ($\beta_1 = 0, \beta_2 = 0$).

Figure 2

Example of how different β_2 values affect distribution of monthly tourist share



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

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

Data collection and analysis

Figure 3 summarizes our data collection and analysis process. The simulation process involves generating tourist flow under the scenario where all tourists rely on generative AI for travel suggestions. Using

empirically grounded demographic profiles, we created simulated tourists and provided their information to two large language models. Separately from the 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 regression model in Equation 6.

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 (UNWTO, n.d.), offering geographically and socioeconomically diverse destinations. 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 (Qin et al., 2025), making the US context advantageous for these models.

We used 2019-2023 American Community Survey 5-Year Public Use Microdata Sample data to derive stratum weights for the population. Table 1 summarizes sex, age, and household income proportions 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 generative AI models for generating synthetic data. Examples include using large language models to simulate social interactions (Liu et al., 2023) and human decision-making under game theory (Akata et al., 2025). Tourism scholars have also began noting the benefits and challenges of using AI-generated synthetic data (see Ali, 2025; Viglia et al., 2024).

We propose an approach that differs from prior studies in four ways. We assume that large language models do not accurately replicate real-world tourism and contain biases. Hence, these synthetic datare used to measure such biases, rather than using AI-generated data to *mimic* tourist behavior (for example Viglia et al., 2024; Xiong et al., 2024). Next, we base our simulations on empirically derived demographic profiles. This modification is to ensure that any divergence from empirical tourism patterns is attributable to algorithmic biases, instead of using arbitrary or non-representative profiles that would confound the results (for example, Andreev et al., 2025). Third, unlike studies that relied on a few handpicked responses or a single simulation run (for example Andreev et al., 2025; Mellors, 2025; Xiong et al., 2024), we ran multiple iterations of simulations to sufficiently capture the uncertainty in large language model outputs. Finally, our approach recognizes the network and temporal nature of tourism. Beyond looking at the propensity of large language models to suggest specific destinations, our simulation captures the entire tourist flow network across origins, destinations, and months.

The simulation starts with a system prompt containing the simulation context, response structure, and examples (available in Appendix A). To simplify the simulation, we assume that each person chooses one domestic destination within the US (50 US states and the District of Columbia). The large language models were instructed to act as travel agents recommending one domestic travel destination based on the provided demographic profile. The system prompt was followed by a user prompt that included demographic characteristics of each simulated individual. Twenty people were processed at a time to improve the efficiency of the simulation. This process is similar to agent-based modeling using large language models (Gao et al., 2024). But the key difference is that we prompt large language models to act as travel agents, rather than as

Figure 3

Overview of the simulation and analysis process

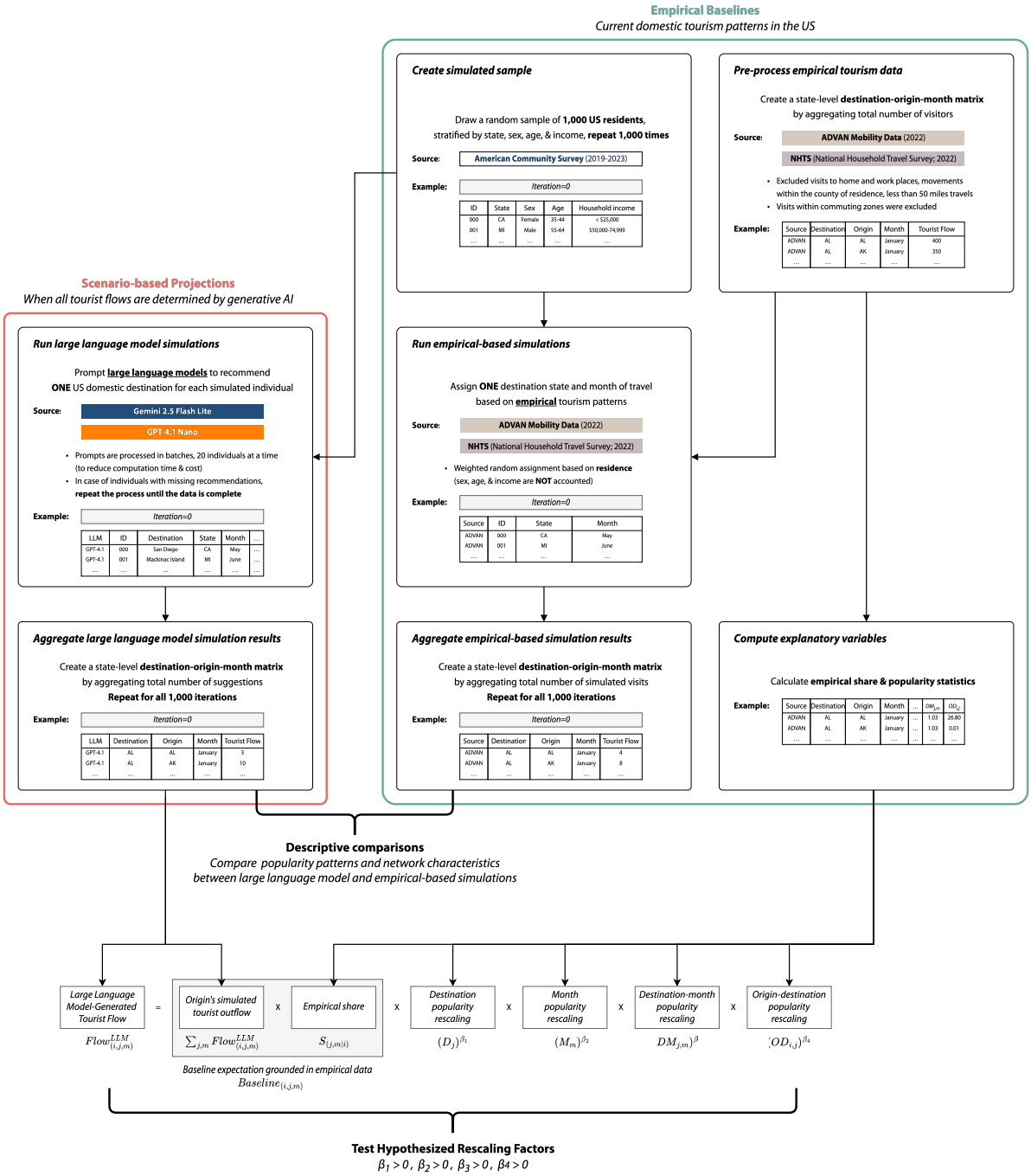


Table 1

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.

tourists. The rationale for this choice is that our goal is to project how generative AI would influence tourist flows if widely adopted, rather than using it to substitute for human travelers as research subjects.

Because outputs of large language models are probabilistic, we took an iterative approach to capture the uncertainty in their recommendations. The large language models generated recommendations for 1,000 simulated individuals across 1,000 samples. This process produces a distribution of simulated tourist visits given the demographic profile, helpful in assessing whether the differences in network characteristics between scenarios are meaningful. Studies have used similar approaches for assessing structural properties of social networks (e.g., Bearman et al., 2004). The difference is that large language models generate tourist flow networks, rather than defining a model with explicit rules about the formulation of the network.

To establish the consistency of our findings across different large language models, we used two different models for simulations: 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 large language models currently available in terms of their capabilities and market share. Because our simulations require a large number of responses, we chose their smallest variants optimized for speed and cost. Although the two large language models are closed-source, they are comparable in pricing structures ($0.10 per million input tokens; $0.30 and $0.40 per million output tokens, respectively).

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 to avoid potential bias. Data collection continued until we achieved a complete dataset. The collected data were then aggregated 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 serves two purposes in this study. It provides the baseline for real-world tourism patterns against which we compare the characteristics of AI-simulated tourist flows. 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 (see Huang et al., 2021; Z. Li et al., 2024), we rely on two different data sources to ensure the robustness of our findings. Our primary data source is the Advan Research (2022) Mobility Data, which estimates movements across US census block groups based on mobile device panels. This dataset is our primary data source due to its high spatial and temporal resolution, and availability of longitudinal data. The supplementary data source is the Department of Transportation's 2022 National Household Travel Survey (NHTS). National Household Travel Survey is the only official national travel survey in the US, which collects travel behavior data such as trip purpose, modes, and distances (Federal Highway Administration, 2022).

For both datasets, we used monthly data from January through December 2022 (the latest available for NHTS). 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 and Pennington-Gray (2025)). Additionally, trips under 50 miles one-way and trips within NHTS designated 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 in the Equation 5.

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 to generate a distribution of tourist flows. Subsequently, the results were aggregated to create destination-origin-month matrices for each iteration and two data sources.

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 estimatior, 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, 2006, 2011). Because we ran 1,000 iterations of simulations, the model is fitted separately for each iteration and model. Then, we collect the estimated coefficients across iterations to assess their significance.

Table 2

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).

Results

Descriptive analysis

Table 2 presents descriptive statistics of simulation and empirical data. The origin total outflow are closely matched across all simulations, as our simulated sample is stratified to reflect the population distribution across origins states. The median tourist flow per destination-origin-month cell is 0 for all simulations, indicating that more than half of the cells have no tourist flow. This sparsity is expected, given that each iteration assigns 1,000 tourists across 31,212 possible combinations (51 origins × 51 destinations × 12 months). Thus, even if we randomly assign each of simulated tourists, only about 3.2% of destination-origin-

Table 3

Agreement in precence 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 (23%)	15,916 (51%)	10,714 (34%)	12,283 (39%)
Any flow	701 (2.2%)	7,514 (24%)	1,885 (6.0%)	6,330 (20%)
Simulation: GPT 4.1 Nano				
No flow	7,119 (23%)	14,711 (47%)	10,648 (34%)	11,182 (36%)
Any flow	663 (2.1%)	8,719 (28%)	1,951 (6.3%)	7,431 (24%)

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

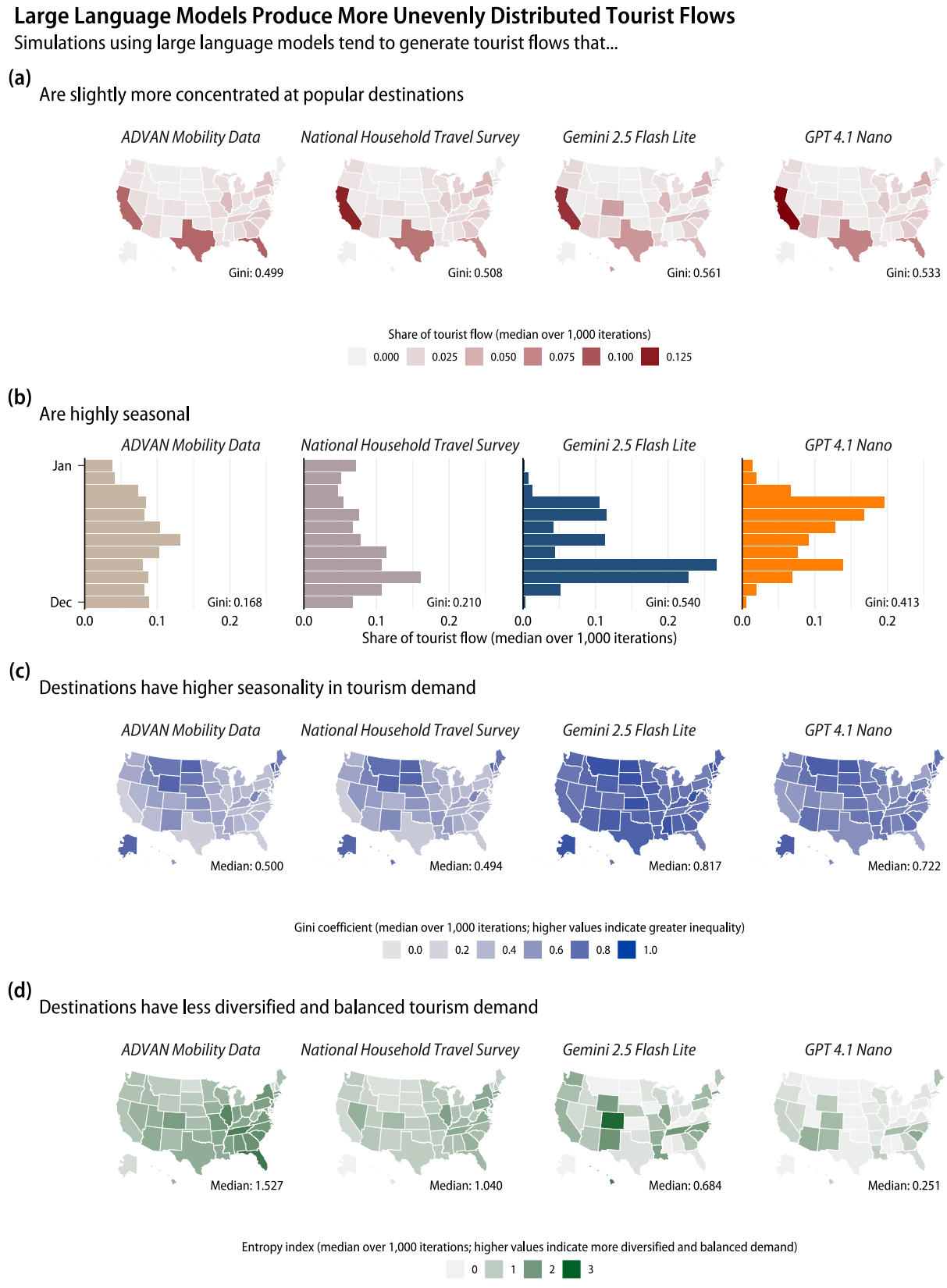
month cells would have at least one tourist flow. But the head of the distribution differs across simulations. Large language model simulations show higher concentration of tourist flows in the most popular destination-origin-month combination. The maximum tourist flow across all cells and iterations is 72 for Gemini 2.5 Flash Lite and 46 for GPT 4.1 Nano, which are higher than the two empirical-based simulations (17 for ADVAN Mobility Data and 30 for National Household Travel Survey).

Similar patterns are observed when considering presence of *any* tourist flow, without accounting for differences in number of tourists. Table 3 presents the agreement between large language model and empirical-based simulations in the precense of any tourist flow across destination-origin-month combinations. We aggregated the simulated visits across all 1,000 iterations, to reduce the sparsity arising from having only 1,000 tourists per iteration. Hence, the results indicate whether any of one million simulated tourists were assigned to a given destination-origin-month combination. In all four comparisons, large language models exclude specific destination-origin-month combinations even though they are present in empirical-based simulations. For example, 51.0% of 31,212 combinations are observable in simulation based on empirical ADVAN Mobility Data but not in flows generated by Gemini 2.5 Flash Lite. Another consistent pattern is that large language models rarely suggest destination-origin-month combinations that are absent in empirical-based simulations. Meaning, large language models typically *prune* a large portion of destination-origin-month combinations, but rarely generate new combinations that are absent in empirical data.

We further examine differences in distribution of tourist flow across simulations by looking at the four popularity factors. Since we ran 1,000 iterations of simulations, we summarize the findings by taking the median across all iterations. Mathematical definitions of the metrics are available in Appendix B. First, we examine distribution of tourist share by destination states (D_j) and months (M_m). We use the Gini index to quantify the inequality in these distributions, where a higher Gini index indicates greater concentration of tourist share among fewer states or months. In all four simulations, tourist visits concentrate in a few popular states, such as California, Florida, and Texas (Figure 4a). However, large language model simulations show higher Gini indices than empirical-based simulations, indicating greater inequality in tourist arrivals

Figure 4

Characteristics of destination, month, destination-month, and origin-destination popularities across simulations



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).

Seasonal patterns are also more pronounced in the large language model outputs than in the empirical-based scenarios (Figure 4b). The two large language model-generated networks also have substantially high Gini indices across months (Median = 0.540 and 0.413 for Gemini 2.5 Flash Lite and GPT-4.1 Nano). This level of seasonality exceeds that of the two empirical-based scenarios (Median = 0.168 and 0.210 for ADVAN Mobility Data and National Household Travel Survey). Although, the two large language models are different in which months are most popular. Gemini 2.5 Flash Lite shows a clear peak during September and October, with more than half of all simulated tourist arrivals concentrated in these two months. GPT-4.1 Nano shows peaks in April, May, and September. Further, large language models worsen the seasonality of tourism demand at the destination level. In Figure 4c, we calculated Gini index of monthly tourist share for each destination state and took the median over 1,000 iterations. Few states show relatively high seasonality in the empirical-based simulations, such as Alaska and Hawaii. However, most states exhibit high seasonality in the large language model simulations. This tendency leads to overall higher level Gini indices accross all destinations (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).

Examining patterns in origin-destination pairs require a different approach. Although Gini index can be used, it does not effectively capture whether destination states receive tourists from a diverse set of origin states or rely on a few. Such demand characteristics are of importance for tourism sector, as less diversity of tourist origins and higher reliance on a few origin markets undermines resilience to shocks (see Lee & Pennington-Gray, 2025). We use entropy index for measuring how diversified and balanced the demand for destinations is. A higher entropy value indicates that a destination receives tourists from wide range of origins (diversified demand) and that each origin have similar contribution to the total tourist arrivals (balanced demand). We calculate the entropy index for each destination and summarize the results over 1,000 iterations by taking the median.

Figure 4d shows that large language models have tendency to generate travel suggestions that destinations rely on a few origins. The empirical simulations show that states such as Florida, Illinois, and North Carolina tend to have diversified and balanced demand (higher entropy). For the two AI-simulated tourist flows, the entropy indices are overall lower than those of empirical-based simulations (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). We also observe that fewer states exhibit relatively high entropy values in the large language model simulations (for example, Colorado and Hawaii for Gemini 2.5 Flash Lite).

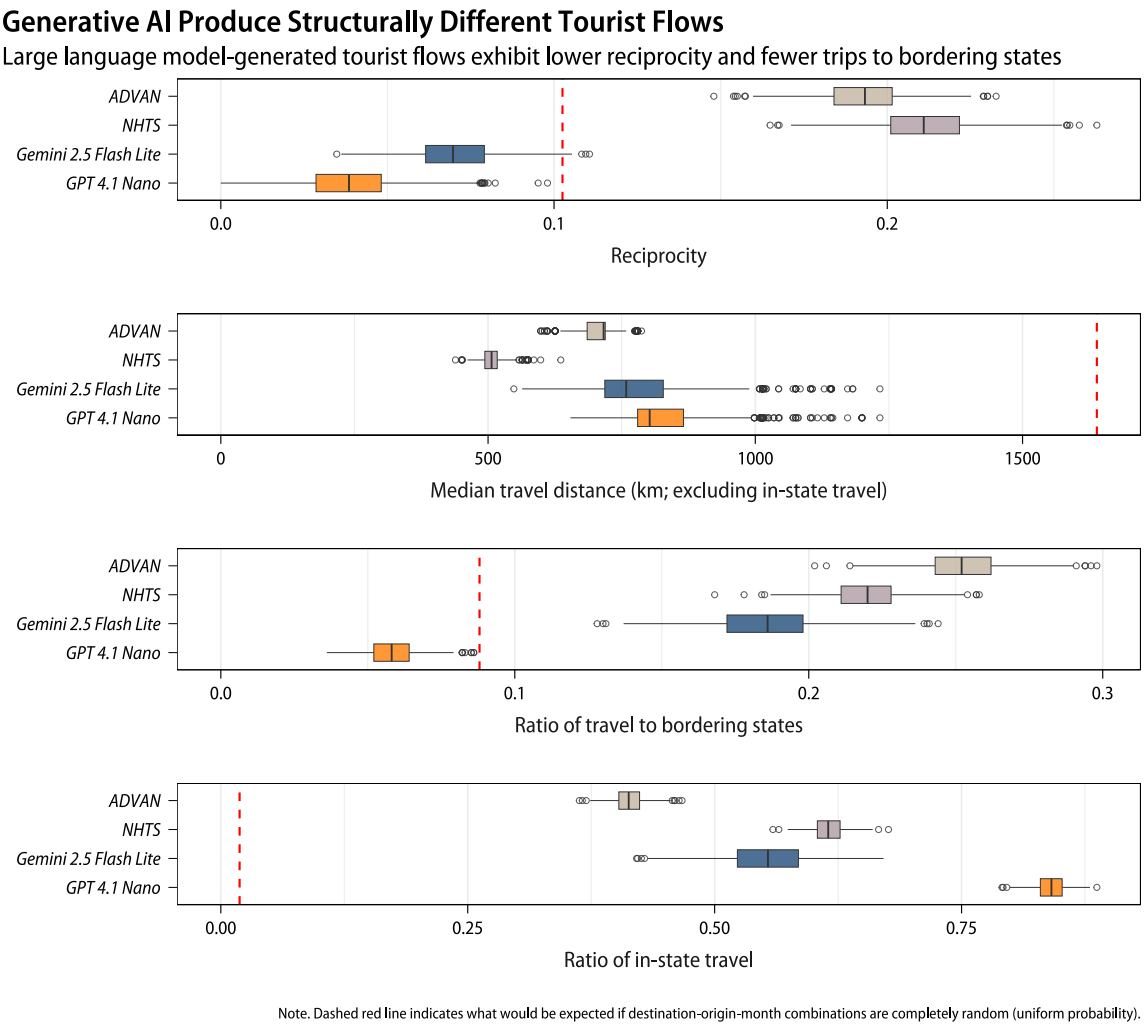
In addition to analyzing individual popularity factors, we analyze the overall structure of tourist flows. We constructed an origin-destination matrix by aggregating tourist numbers at the year level. Then we computed four statistics capturing how the global structure of tourist flows differ across simulations. Boxplots in Figure 5 present the distribution of the four network-level statistics across 1,000 iterations, colored by simulation scenario. The dotted red line indicates what can be expected by chance alone, if there is no structure in tourist flow. This expectation under complete randomness is calculated by assuming that each individual have equal probability of choosing a destination (a probability of $1/51$).

Reciprocity refers to the tendency to form mutual relationships—in our case, two states exchanging a similar number of tourists. Empirical-based simulations show higher levels of reciprocity compared to the random expectation (Median = 0.193 and 0.211 for ADVAN Mobility Data and National Household Travel Survey). However, large language model simulations show lower levels of reciprocity than expected by chance (Median = 0.070 and 0.038 for Gemini 2.5 Flash Lite and GPT 4.1 Nano). Simply put, the large language models produce tourist flows with a clear separation between states that send tourists and those that receive them.

Gemini 2.5 Flash Lite and GPT 4.1 Nano tend to suggest further destinations compared to empirical

Figure 5

Characteristics of tourist flow structure across simulations



data, when looking at median travel distance excluding in-state trips (Median of medians = 758 and 802). Similarly, Gemini 2.5 Flash Lite has lower ratio of tourist flows between bordering states compared to the empirical models (Median = 0.186). GPT-4.1 Nano shows even lower tendency to suggest bordering states, lower that what would be expected under randomness (Median = 0.058). The two models show different propensity to suggest in-state tourism. Gemini 2.5 Flash Lite shows ratio of in-state trips comparable to the empirical models (Median = 0.554). In contrast, GPT 4.1 Nano shows much stronger preference for recommending tourist to travel within their own state (Median = 0.841).

Model estimation results

Table 4 and Figure 6 summarizes how the four popularity factors can explain deviation between large language model-simulated tourist flows and the empirical expectations (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, lower bounds of 95% credible intervals for Gemini

Table 4

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

	Empirical: ADVAN			*Empirical: NHTS*		
	Median	2.5%	97.5%	Median	2.5%	97.5%
Simulation: Gemini 2.5 Flash Lite						
β_1: Destination	-0.175	-0.242	-0.115	-0.245	-0.387	-0.112
β_2: Month	0.089	-0.192	0.423	0.887	0.472	1.294
β_3: Destination-Month	0.882	-0.097	1.480	2.669	0.369	4.993
β_4: Origin-Destination	0.066	-0.027	0.141	-0.168	-0.254	-0.086
Simulation: GPT 4.1 Nano						
β_1: Destination	-0.006	-0.040	0.027	0.180	0.139	0.229
β_2: Month	-0.789	-0.913	-0.644	-0.849	-1.032	-0.675
β_3: Destination-Month	1.771	1.042	2.180	1.611	0.950	2.249
β_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.

2.5 Flash Lite crosses zero with ADVAN Mobility Data baseline and is very wide with National Household Travel Survey baseline, indicating unncertainty around the estimate.

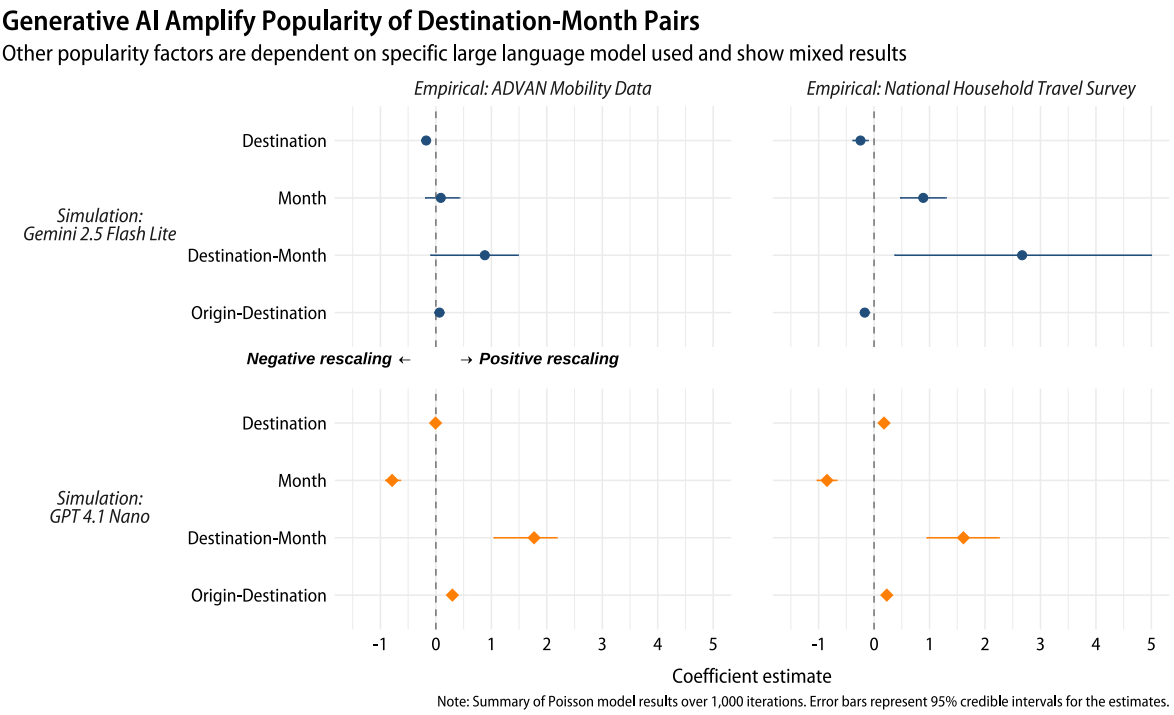
One way to interpret the coefficient for destination-month popularity is to consider it as an elasticity. For example, GPT 4.1 Nano with ADVAN Mobility Data baseline had a median coefficient of 1.771 for destination-month popularity. If real-world data shows that a particular destination-month combination 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 factors. 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 β_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 β_1 = 0.180) but not with ADVAN Mobility Data baseline (Median β_1 = -0.006). GPT 4.1 Nano with shows positive rescaling for origin-destination popularity (Median β_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 β_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 β_2 = -0.789 and -0.849 with ADVAN Mobility Data and National Household Travel Survey baselines). This is contradictory, given prior descriptive analysis indicated GPT 4.1 Nano having greater seasonality. One possible explanation is that peak months in GPT 4.1 Nano simulations do not 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.

Figure 6

Summary of popularity effect estimates across 1,000 iterations



Robustness checks

We conducted following robustness checks to test the sensitivity of our findings to different simulation choices (reported in Appendix C). 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 A). 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 C1). 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.

Changing the temperature parameter also does not alter the results. The temperature parameter controls randomness in large language model outputs. Higher temperature setting produce more diverse outputs, while lower temperature setting generate more deterministic outputs. The default temperature setting used in our main analysis is 1.0. We collected additional data using temepratures of 0.5 and 1.5 (Figure C2). GPT 4.1 Nano with temperature of 1.5 could not generate valid outputs and hence was excluded. Similiar 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 amplified destination-month popularity. 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 strongest positive effects (Figure C3). Compared to the models used in our main analysis, larger models show stronger destination-month popularity bias, not weaker.

Finally, we examined robustness of our findings to alternative approach for fitting Poisson regression model to our iterative simulation data. Our results are robust to using aggregated data 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. Coefficient estimates using aggregated data are nearly identical to median estimates from our main approach (see Table C1).

Discussions

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

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., & Schulz, E. (2025). Playing repeated games with large language models. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-025-02172-y

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

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

Bearman, P. S., Moody, J., & Stovel, K. (2004). Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*, *110*, 44–91. https://doi.org/10.1086/386272

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/

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

Dawes, J. (2024). Anthropic's new AI feature mimics human travel agents. *Skift*. https://skift.com/2024/10/23/anthropics-new-ai-feature-mimics-human-travel-agents/

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

Epstein, Z., Hertzmann, A., Investigators of Human Creativity, the, Akten, M., Farid, H., Fjeld, J., Frank, M. R., Groh, M., Herman, L., Leach, N., Mahari, R., Pentland, A. "Sandy"., Russakovsky, O., Schroeder, H., & Smith, A. (2023). Art and the science of generative AI. *Science*, *380*, 1110–1111. https://doi.org/10.1126/science.adh4451

Expedia. (2023, April 4). *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

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

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 Linguis-*

tics, *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

Grundner, L., & Neuhofer, B. (2021). The bright and dark sides of artificial intelligence: A futures perspective on tourist destination experiences. *Journal of Destination Marketing & Management*, *19*, 100511. https://doi.org/10.1016/j.jdmm.2020.100511

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

Huang, X., Li, Z., Jiang, Y., Ye, X., Deng, C., Zhang, J., & Li, X. (2021). The characteristics of multi-source mobility datasets and how they reveal the luxury nature of social distancing in the U.S. During the COVID-19 pandemic. *International Journal of Digital Earth*, *14*(4), 424–442. https://doi.org/10.1080/17538947.2021.1886358

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

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. (2025). 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

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

Leiper, N. (1979). The framework of tourism: Towards a definition of tourism, tourist, and the tourist industry. *Annals of Tourism Research*, *6*, 390–407. https://doi.org/10.1016/0160-7383(79)90003-3

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

Liu, R., Yang, R., Jia, C., Zhang, G., Zhou, D., Dai, A. M., Yang, D., & Vosoughi, S. (2023). *Training socially aligned language models on simulated social interactions*. https://doi.org/10.48550/arXiv.2305.16960

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/

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

Pan, B., Lin, M. S., Liang, Y., Akyildiz, A., & Park, S. Y. (2021). Social, ethical, and moral issues in smart tourism development in destinations. *Journal of Smart Tourism*, *1*, 9–17. https://doi.org/10.52255/smarttourism.2021.1.1.3

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

Runge, M. C., Shea, K., Howerton, E., Yan, K., Hochheiser, H., Rosenstrom, E., Probert, W. J. M., Borchering, R., Marathe, M. V., Lewis, B., Venkatramanan, S., Truelove, S., Lessler, J., & Viboud, C. (2024). Scenario design for infectious disease projections: Integrating concepts from decision analysis and experimental design. *Epidemics*, *47*, 100775. https://doi.org/10.1016/j.epidem.2024.100775

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

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. (2006). The Log of Gravity. *The Review of Economics and Statistics*, *88*(4), 641–658. https://doi.org/10.1162/rest.88.4.641

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

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. *Computer 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

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

Appendix A
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 B
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:

$$Gini^{Destination} = \frac{|P_{(Alabama)} - P_{(Alaska)}| + \cdots + |P_{(Wyoming)} - P_{(Wisconsin)}|}{2 \cdot 51^2 \cdot \overline{P_{(j)}}} \tag{B1}$$

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 B1, we use Gini index to measure inequality of tourist flows across months:

$$Gini^{Month} = \frac{|P_{(1)} - P_{(2)}| + \cdots + |P_{(12)} - P_{(11)}|}{2 \cdot 12^2 \cdot \overline{P_{(m)}}} \tag{B2}$$

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:

$$Gini_j^{Month} = \frac{|P_{(j,1)} - P_{(j,2)}| + \cdots + |P_{(j,12)} - P_{(j,11)}|}{2 \cdot 12^2 \cdot \overline{P_{(j,m)}}} \tag{B3}$$

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 (2025), we measure the diversity of tourist orgins 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:

$$Entropy_j = -\sum_i S_{(i,j)} \cdot \ln(S_{(i,j)}) \tag{B4}$$

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:

$$Reciprocity = \frac{\sum_{i,j \neq i} min[Flow_{(i,j)}, Flow_{(j,i)}]}{\sum_{i,j} Flow_{(i,j)}} \quad (i \neq j) \tag{B5}$$

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:

$$R_{Border} = \frac{\sum_{i,j} Flow^*_{(i,j)}}{\sum_{i,j} Flow_{(i,j)}} \tag{B6}$$

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:

$$R_{InState} = \frac{\sum_{i=j} Flow_{(i,j)}}{\sum_{i,j} Flow_{(i,j)}} \tag{B7}$$

Appendix C
Robustness checks

Figure C1

Summary of popularity effect estimates with alternative prompts

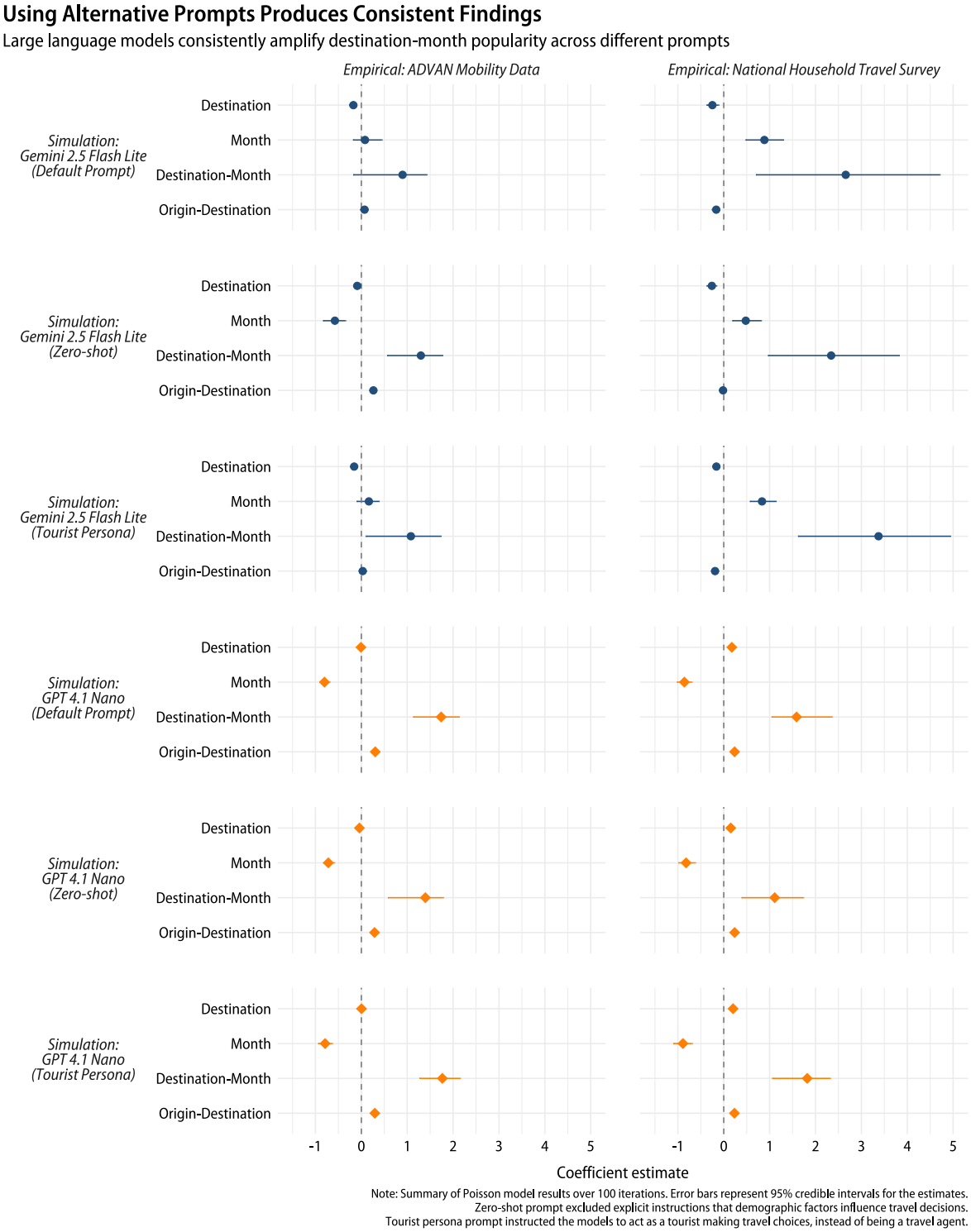


Figure C2

Summary of popularity effect estimates with different temperature settings

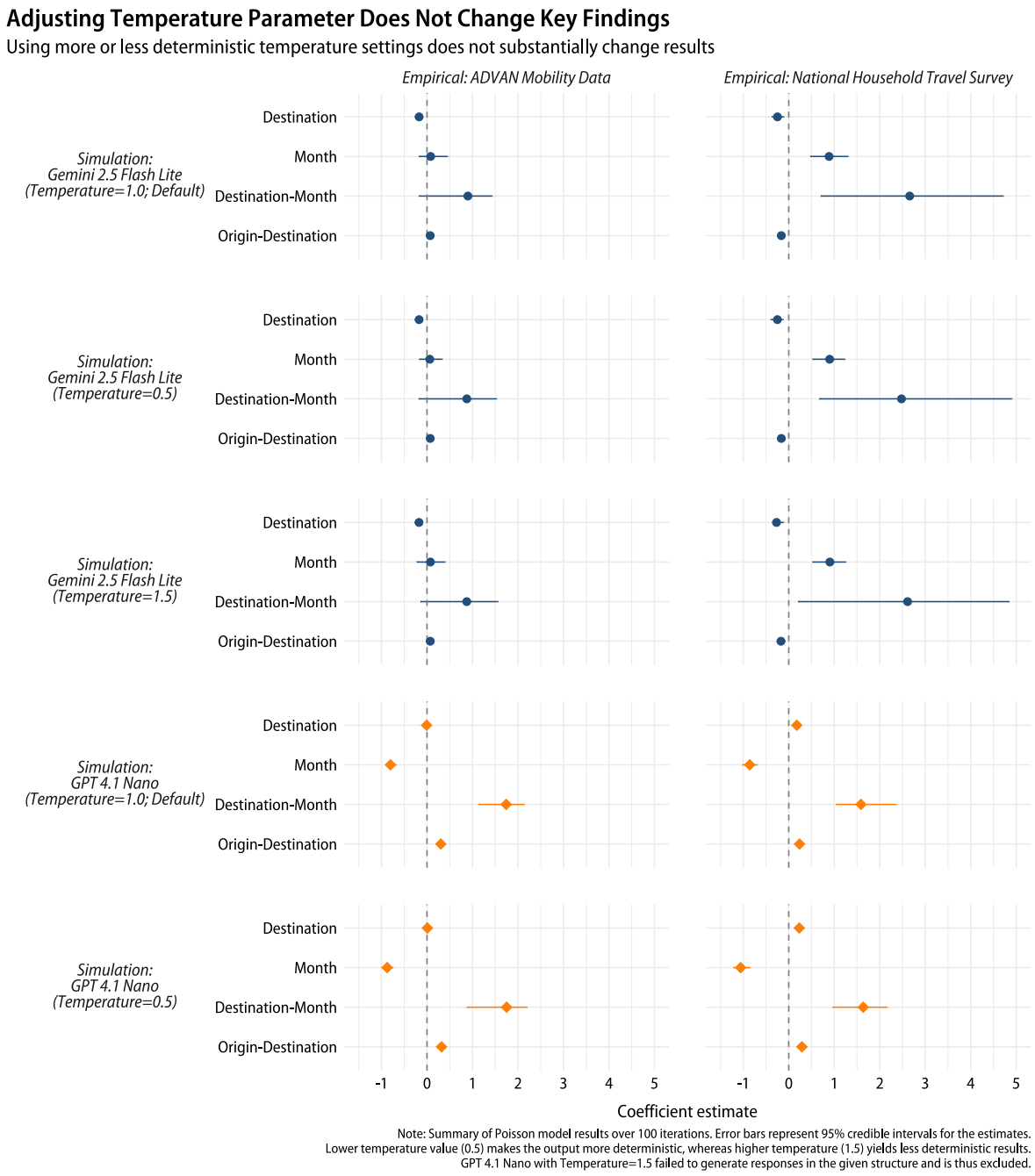


Figure C3

Summary of popularity effect estimates with additional large language models

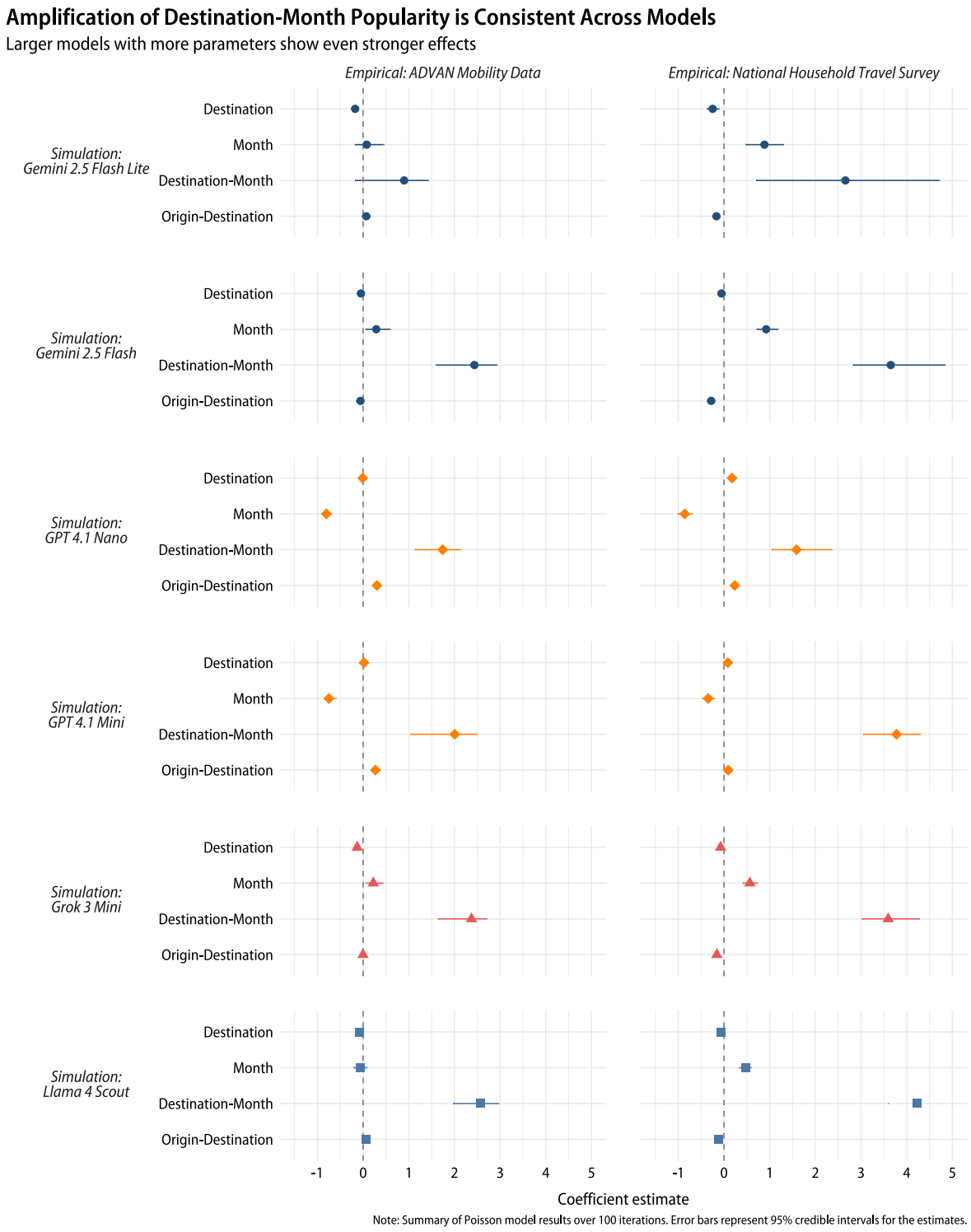


Table C1

Regression results with aggregated data

	Empirical: ADVAN			*Empirical: NHTS*		
	Estimate	Robust SE	*p*	Estimate	Robust SE	*p*
Simulation: Gemini 2.5 Flash Lite						
β_1: Destination	-0.178	(0.069)	0.009	-0.250	(0.064)	<0.001
β_2: Month	0.093	(0.168)	0.578	0.887	(0.118)	<0.001
β_3: Destination-Month	0.821	(0.306)	0.007	2.692	(0.848)	0.001
β_4: Origin-Destination	0.063	(0.023)	0.006	-0.170	(0.018)	<0.001
Simulation: GPT 4.1 Nano						
β_1: Destination	-0.007	(0.053)	0.901	0.180	(0.059)	0.002
β_2: Month	-0.788	(0.122)	<0.001	-0.850	(0.133)	<0.001
β_3: Destination-Month	1.726	(0.323)	<0.001	1.603	(0.572)	0.005
β_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.