AI

MDPI

Article

Destination (Un)Known: Auditing Bias and Fairness in LLM-Based Travel Recommendations

Hristo Andreev [1,*], Petros Kosmas [1,*], Antonios D. Livieratos [2], Antonis Theocharous [1] and Anastasios Zopiatis [1]

[1] Department of Hospitality and Tourism Management, Cyprus University of Technology, Ampelokipon 9, 8027 Paphos, Cyprus; antonis.theocharous@cut.ac.cy (A.T.); anastasios.zopiatis@cut.ac.cy (A.Z.)
[2] Department of Business Administration, National and Kapodistrian University of Athens, 14 Evripidou, 10559 Athens, Greece; alivieratos@ba.uoa.gr
* Correspondence: hd.andreev@edu.cut.ac.cy (H.A.); petros.kosmas@cut.ac.cy (P.K.)

Abstract

Large language-model chatbots such as ChatGPT and DeepSeek are quickly gaining traction as an easy, first-stop tool for trip planning because they offer instant, conversational advice that once required sifting through multiple websites or guidebooks. Yet little is known about the biases that shape the destination suggestions these systems provide. This study conducts a controlled, persona-based audit of the two models, generating 6480 recommendations for 216 traveller profiles that vary by origin country, age, gender identity and trip theme. Six observable bias families (popularity, geographic, cultural, stereotype, demographic and reinforcement) are quantified using tourism rankings, Hofstede scores, a 150-term cliché lexicon and information-theoretic distance measures. Findings reveal measurable bias in every bias category. DeepSeek is more likely than ChatGPT to suggest off-list cities and recommends domestic travel more often, while both models still favour mainstream destinations. DeepSeek also points users toward culturally more distant destinations on all six Hofstede dimensions and employs a denser, superlative-heavy cliché register; ChatGPT shows wider lexical variety but remains strongly promotional. Demographic analysis uncovers moderate gender gaps and extreme divergence for non-binary personas, tempered by a "protective" tendency to guide non-binary travellers toward countries with higher LGBTQI acceptance. Reinforcement bias is minimal, with over 90 percent of follow-up suggestions being novel in both systems. These results confirm that unconstrained LLMs are not neutral filters but active amplifiers of structural imbalances. The paper proposes a public-interest re-ranking layer, hosted by a body such as UN Tourism, that balances exposure fairness, seasonality smoothing, low-carbon routing, cultural congruence, safety safeguards and stereotype penalties, transforming conversational AI from an opaque gatekeeper into a sustainability-oriented travel recommendation tool.

Keywords: large language models; travel recommendation systems; AI bias; sustainable tourism recommendations

Academic Editor: Gianni D'Angelo

Received: 31 July 2025
Revised: 12 September 2025
Accepted: 15 September 2025
Published: 19 September 2025

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

1. Introduction

Large language models (LLMs) are rapidly becoming the first stop for trip planning. They offer travellers instant, conversational advice that once required hours of web searches or guidebook reading. Early audits indicate that LLMs can reproduce structural imbalances long seen in tourism media, such as favouring mainstream hotspots, overlooking smaller

communities, and tailoring suggestions unevenly across cultures and demographics. Despite growing concern, empirical evidence remains limited, particularly for the newest generation of LLMs and across the full spectrum of bias types identified in the literature.

Bias and fairness in AI-based travel decision making concern the ethical and practical challenges that arise when artificial intelligence recommends destinations and experiences. Selection, data, and algorithmic-design biases can misrepresent diverse traveller preferences and produce unfair outcomes [1,2]. These effects shape individual choices and the tourism landscape; algorithms may skew suggestions based on demographic assumptions, altering perceptions of affordability and accessibility [3,4]. Economic consequences can be uneven if larger firms gain disproportionate visibility, with adverse effects on local operators and cultural representation [5,6]. Mitigation efforts include data diversification, algorithmic audits, stakeholder engagement, and continuous evaluation to achieve more representative and equitable recommendations [7,8].

AI systems generate personalised recommendations by analysing past trips, social media activity, and real-time trends, producing itineraries aligned with user preferences [9,10]. Research on fairness in travel recommendation remains nascent, limited by the scarcity of public data for development and validation; many studies rely on synthetic or proprietary data that impede independent replication [1,11]. Representation of diverse interactions is therefore insufficient, which hinders algorithms that serve all demographics equitably. Addressing popularity bias and highly active user effects is essential, and incorporating stakeholder feedback from diverse groups improves performance and satisfaction [1,7]. As the industry embraces AI, it must balance innovation with safeguards so systems promote inclusion rather than reproduce disparities [8,9,12].

Audits should distinguish prompt-based bias from model-based bias. Prompt-based bias arises when user questions include stereotypes or constraints that steer answers, whereas model-based bias originates in training data, labelling, or design choices and can persist under neutral prompts [13]. Absolute neutrality in training data is unattainable, so this article defines neutrality operationally as neutral prompts and symmetric execution procedures rather than a claim that the model or corpus is neutral [14]. We kept prompts consistent and neutral across personas and used session controls to limit personalisation and carry-over effects. The patterned differences observed are best interpreted as model-based bias rather than artefacts of user input.

Most prior studies examine a single bias, user group, or model. Four gaps follow. Cross-bias comparability requires consistent metrics; small per-user differences can accumulate to system-level effects; contemporary model contrasts under controlled sessions help separate prompt-driven from model-driven patterns; and governance linkage is needed to translate audit findings into a practical re-ranking layer that serves sustainability and safety. This study targets these gaps by identifying model-induced biases that persist under neutralised input while recognising that all knowledge sources are culturally situated. Recent audits and research converge on three issues: recommendations concentrate on already popular options, creating popularity and geographic bias [15,16]; cultural representations are uneven, with WEIRD-centred perspectives and promotional clichés inherited from web training data [14,17,18]; and demographic disparities appear when advice changes with cues such as gender [19,20]. Proposed remedies emphasise data diversity, regular auditing, and transparency, yet shared public benchmarks and standardised metrics for consistent comparison remain limited [1,21].

Prior audits show that LLMs can exhibit popularity and geographic concentration, uneven cultural representation, and demographic differences, but evidence is fragmented across models, origins, metrics, and contexts. What is missing is an integrated, persona-aware evaluation that assesses multiple bias families under a single, transparent framework,

contrasts contemporary models under controlled, neutralised prompts, and links audit results to actionable governance. This paper fills that gap by analysing more than 6000 recommendations for 216 traveller personas across two state-of-the-art LLMs (ChatGPT-4o and DeepSeek-V3), one U.S.-developed and one China-developed model, using simple, standardisable metrics and proposing a public-interest re-ranking layer to improve equity without degrading utility.

1.1. Biases in AI Systems

AI systems used in travel decision making can be influenced by various types of bias, which can lead to unfair outcomes and reinforce existing inequalities. Understanding these biases is crucial to mitigating their impact on destination selection and travel experiences. Table 1 categorises and summarises the most common biases in AI systems.

Table 1. Biases in AI systems.

Bias Category	Biases	Conceptual Roots	Bias Stage	Rationale
Data & design-time (technical) biases	• Selection • Data • Algorithmic design	"Technical bias" Friedman & Nissenbaum, 1996 [22] and the representation/measurement/aggregation points in the Suresh & Guttag, 2019 [23] pipeline	Upstream—data curation, feature choice, model architecture	All three stem directly from how data are collected, labelled, or mathematically encoded before the model is ever used.
Feedback & distribution (emergent/interaction) biases	• Reinforcement • Popularity • Geographic	"Emergent bias" Friedman & Nissenbaum 1996 [22] and "interaction bias" bias on the web recommendation system and search engine algorithms (Baeza-Yates, 2018) [24]	Deployment—user loops, ranking, continual learning	These arise only after the system's outputs start guiding new behaviour or data collection, creating self-reinforcing loops that push popular or well-documented places to the top and further marginalise the rest.
Representational & societal (pre-existing) biases	• Cultural • Stereotype • Demographic	"Pre-existing bias" Friedman & Nissenbaum, 1996 [22] and representational harms vs. allocative harms due to undetected biases in big data (Barocas & Selbst, 2016) [25]	Any user-visible output—content generation, explanations	These reflect historical power imbalances and cultural assumptions already embedded in text and images on the Web; the model simply reproduces them in its recommendations.

1.1.1. Data & Design-Time (Technical) Biases

Selection bias occurs when the training data does not represent the target population. Under-represented groups receive poorer model performance and less relevant recommendations [26–28]. Data bias emerges during data collection and preparation, producing skewed signals even before modelling. It includes sampling bias, where some groups are more likely to be included than others [27,29,30]; labelling bias, where annotator judgements alter ground truth [27] and statistical bias, where historical patterns embed pre-existing inequalities that the model reproduces [27,31]. For instance, if historic travel logs reflect one group's preferences, the system will over-prioritise those preferences. Algorithmic design bias is introduced through problem framing, feature selection, and factor weighting. These choices can encode developer assumptions and produce systematic favouritism if not audited. Insufficient oversight allows such effects to persist in production [30,32,33].

1.1.2. Feedback & Distribution (Emergent/Interaction) Biases

Reinforcement bias occurs when outputs influence future inputs, which then validate the model's initial behaviour. If certain users are shown fewer options, subsequent interaction data will confirm that reduced exposure, entrenching disparities over time [28,34].

Popularity bias is the tendency to recommend items that already dominate the corpus, such as widely covered attractions [15]. As a result, itineraries converge on the same landmarks while the long tail of lesser-known sites remains under-exposed. Users receive predictable lists that may not match individual interests, and traffic concentrates in established hotspots [35]. Geographic bias appears when recommendations over-emphasise well-documented cities or regions and under-represent rural or disadvantaged areas [16,36]. This narrows itineraries to popular destinations, intensifies crowding, and reduces visibility for overlooked communities.

1.1.3. Representational & Societal (Pre-Existing) Biases

Cultural bias is the alignment of outputs with dominant cultural norms because training texts disproportionately represent those norms [14]. In tourism, models privilege experiences favoured by Western audiences and give less attention to local practices. Many models align with contexts abbreviated as WEIRD, meaning Western, Educated, Industrialised, Rich, and Democratic [17]. This reduces relevance for travellers from other cultural backgrounds. Stereotype bias occurs when the model reproduces clichés and simplified tropes about places, which flattens complexity and misinforms expectations [18]. Demographic bias arises when advice changes with cues such as gender or ethnicity, reflecting stereotypes rather than genuine personalisation [19,20,37]. For example, solo female travellers may receive over-cautious, shopping-centred itineraries, while solo male travellers are steered toward nightlife or adventure, regardless of stated preferences.

1.2. Bias and Its Consequences in AI Systems for Tourism Recommendations

Addressing bias in AI systems is essential to ensure fair outcomes in travel decision making [2]. The sector should adopt comprehensive strategies that combine diverse perspectives, explicit ethical frameworks, and robust evaluation methods to identify and mitigate bias [38]. With such measures, the industry can work toward inclusive and representative systems that benefit all travellers [19]. The implications are individual and systemic: biased outputs affect traveller decisions, industry practices, and local economies, and growing integration of AI raises concerns that recommendations may perpetuate stereotypes or inaccuracies [2,38].

Recent studies report that AI-generated recommendations can reflect demographic assumptions [10]. For instance, one study suggested that UK residents in cities such as Glasgow and Manchester prefer domestic travel, which misrepresented actual aspirations and behaviours [38]. Such patterns shape perceptions of affordability and accessibility and narrow perceived options for global users. Travellers from different socioeconomic backgrounds may receive suggestions that do not match their preferences, framing destinations as unrealistic due to geography rather than demonstrated habits [39,40]. Firms with resources to deploy advanced AI may gain disproportionate visibility, sidelining smaller local operators [1,6]. Well-known destinations benefit from popularity effects that overshadow less developed places, redirecting revenue toward a few locations, increasing overtourism in hotspots, and neglecting others. Biases also affect cultural representation and authenticity [41]. Biases also affect cultural representation and authenticity, as algorithms that emphasise heavily reviewed venues can marginalise smaller businesses, homogenise itineraries, reduce cultural diversity, misinform travellers about local practices, and strain community relations [14,42,43].

1.3. Approaches to Bias Mitigation

Mitigating bias in AI-driven travel decision making requires a lifecycle strategy centred on data quality, transparency, stakeholder engagement, and ongoing evaluation [38]. Diverse and representative training data reduce bias [30]. Organisations should avoid convenience samples and historically skewed records, intentionally cover a wide range of demographics and travel patterns, and use over-sampling or data augmentation with documented criteria and checks [3,4,7,44]. To support repeatability, teams should publish data sources, sampling frames, inclusion and exclusion rules, and version identifiers.

Algorithmic auditing evaluates models against ethical standards and documented metrics. Although the field is developing, standardised procedures for detecting and mitigating algorithmic bias are vital [21,32]. Organisations can run external audits, deploy continuous monitoring across demographic groups, and track performance shifts over time [7,8,37]. Transparency initiatives, including explainable methods, can surface influential features and correlations, which improves understanding and trust [7]. Stakeholder engagement during design and implementation improves fairness and relevance by aligning inputs, outputs, and error review with real-world needs [1,2,26]. Bias mitigation is not a one-off task; teams should build feedback and learning mechanisms into data processing, feature engineering, deployment, and model management. Scheduled re-audits, version control for models and datasets, and public change logs when training data, features, or thresholds change support equitable algorithms and better travel experiences [27,28,30,31].

The literature provides clear bias typologies and mitigation principles, yet cross-bias comparability, persona-level fairness checks, and public benchmarks remain limited. Building on this work, this study operationalises six bias families with transparent definitions and external references, tests two leading LLMs under identical prompt chains, and connects findings to sustainability-relevant re-ranking. The study advances the field from descriptive concerns to a compact, reproducible audit that can be replicated and governed in practice.

1.4. Aim and Research Questions

Proposed remedies for bias, from data diversification to audits and stakeholder engagement, are often general and rarely tested across multiple bias families on current large language models. This gap motivates a holistic audit that quantifies six user-visible bias types within one repeatable framework and links them to sustainability-relevant objectives. This study examines six commonly cited bias families: popularity, geographic, cultural, stereotype, demographic, and reinforcement. Three additional biases, selection, data, and algorithmic design, arise upstream in training data, labelling, and model architecture. Addressing those requires privileged access to datasets, annotations, or internal code that external auditors do not have. They are therefore outside the scope of this study. The aim is to address the research gap on LLM bias in travel recommendation systems by conducting a comprehensive, persona-based audit across multiple bias categories and contrasting contemporary models under controlled conditions. The study is guided by six research questions:

RQ1. Do LLMs exhibit popularity bias in destination suggestions, and how strong are model and theme differences?

RQ2. How do recommendation portfolios vary geographically across origin markets, including the balance between domestic and international options?

RQ3. To what extent do models align with or diverge from travellers' cultural contexts, and do models differ in that respect?

RQ4. How prevalent is stereotyped promotional language in generated rationales, and how do variety and intensity compare across models?

RQ5. Do recommendation patterns differ systematically by gender identity and age, and are consistent advantages or disadvantages observable for particular groups?

RQ6. When asked to refine prior advice, do models introduce genuinely new destinations or repeat earlier items?

The contribution of this work is threefold. First, it shifts the focus from isolated examples of bias to a systematic and comparative measurement framework that can be replicated by other researchers. Second, it connects technical findings on bias to key sustainability goals, including the need to spread tourism demand across time and space, reduce transport-related emissions, and improve visibility for underrepresented regions and traveller groups. Third, it outlines a practical governance approach by introducing a public-interest re-ranking layer managed by a neutral organisation such as UN Tourism, aimed at turning audit results into actionable design principles. Taken together, these contributions show how rigorous, context-aware evaluation can transform LLM-based travel recommenders from opaque gatekeepers into transparent tools that more fairly serve both travellers and destinations.

2. Methodology

2.1. Research Design

The study used a controlled, persona-based experiment to evaluate six bias classes in large language-model (LLM) travel recommendations: popularity, geographic, cultural, stereotype, demographic, and reinforcement. Personas varied origin, age, gender identity, and interests while wording and session context remained constant. Both models were run on the same personas with a fixed three-prompt chain, yielding a within-subject design that isolates model-induced differences and supports replication. Sessions were executed as fresh contexts with fixed prompts and no history to increase internal validity while mirroring typical user behaviour.

Two widely used LLMs—ChatGPT-4o (build 20250326) and DeepSeek-V3 (build 0324) were selected to maximise practical relevance and contrast systems likely trained on different corpora and alignment choices. We make no claims about proprietary training data; all inferences derive from observed outputs under identical controls.

Because there is no ground truth for generative recommendations, outputs were parsed into comparable city- and country-level items and transformed into exposure and distribution-based metrics to operationalise each bias construct. For every persona, the same prompt sequence was executed on each model; texts were parsed into destinations, rationales, and lexical features, then converted into bias-specific metrics for cross-model and cross-persona comparisons.

2.2. Persona Construction

Personas varied along four dimensions, namely country of origin, age, gender identity, and interest constraint. The countries of origin were China, the United States (USA), Germany, the United Kingdom, France, Japan, Saudi Arabia, and India. The eight origin countries were chosen by two criteria: (i) high outbound travel volume (USA, China, Germany, United Kingdom, France) and (ii) added cultural coverage across regions while retaining substantial outbound volume (Japan, India, Saudi Arabia). Ages (25, 45, 65) represent early-career adults, mid-life adults, and seniors, which are common planning cohorts in tourism research. Gender identity includes female, male, and non-binary to enable fairness probes across a protected attribute without assuming a binary structure. Interest constraints (Sun & Sea, Cultural Heritage, Wildlife) were selected to span mass-market and niche themes. All prompts were issued in English to avoid confounding by automatic translation; “country of origin” therefore encodes cultural context rather than

prompt language. The full factorial design produced 8 × 3 × 3 × 3, that is 216 unique personas see Table 2, for a summary of persona characteristics. For every persona the same prompt chain was executed on each model. Each prompt requested five destinations, which yields a theoretical maximum of 216 personas × 2 models × 3 prompts × 5 items, or 6480 destination recommendations before cleaning and deduplication.

Table 2. Persona characteristics.

Dimension	Persona Characteristics
Country of origin	China,·United States, Germany, United Kingdom, France, India, Saudi Arabia, Japan
Age bracket	25 yrs (young), 45 yrs (mid-life), 65 yrs (senior)
Gender identity	Female, Male, Non-binary
Prompt style	Generic, Single-constraint (Sun & Sea, Cultural Heritage, Wildlife), Reinforcement
LLM	ChatGPT 4o- (build 20250326), DeepSeek-v3- (Build 0324)

2.3. Prompting Protocol

For each persona, a three-step prompt chain was executed in the order Generic → Single-constraint → Reinforcement using the templates in Table 3 (Prompt Templates → variables filled from the persona profile). We used strict session controls (fresh browser context and model session; identical vendor defaults; no tools/retrieval; limited regeneration). The reinforcement prompt explicitly excluded the first recommendation from the second prompt; novelty was evaluated only relative to the second-prompt set. Full session controls and execution settings are documented in Appendix A Session controls and execution settings.

Table 3. Prompt templates.

Type	Template (Variables in <...>)
1. Generic	I am a <AGE>-year-old <GENDER> from <COUNTRY>. Please recommend five travel destinations and give reasons for each.
2. Single-constraint	I'm particularly interested in <CONSTRAINT>. Please recommend five destinations and explain why.
3. Reinforcement follow-up	Apart from the first recommendation, could you suggest five other places that fit my profile?

2.4. Experimental Controls and Data Processing

We standardised VPN endpoints to persona origin, rotated IPs, used private windows, and cleared site data between runs, applying identical controls to both models with randomised persona/model order. For each response we stored raw text, timestamps, model ID, and persona metadata. The pipeline comprised destination extraction and normalisation, deduplication, popularity tagging, cultural-distance scoring, stereotype-lexicon features, and joins to external acceptance/safety indices. Cities were mapped to countries ISO 3166-1 alpha-3 standard [45]); ambiguous cities were geocoded; regional labels were mapped to constituent countries for frequency analyses and flagged as regional. Parsing rules, edge-case handling, and robustness checks appear in Appendix A.

2.5. Data Analysis Procedure

The analysis proceeded in six streams that matched the six bias constructs.

Popularity bias is treated as a systematic preference for already popular items that reduces exposure to the long tail of relevant options, as commonly examined with coverage, diversity, and long-tail indicators [15,35,46]. This study reports the share of recommended items that appear in the Euromonitor Top-100 City Destinations Index and the WEF TTDI Top-30 countries, alongside the complementary off-list share [47,48]. Higher on-list rates indicate stronger popularity skew. Group differences are assessed with proportion tests and, where appropriate, mixed-effects logistic regression with prompt-type and persona controls. Country recommendation frequency is also related to TTDI scores using Pearson and Spearman correlations to capture linear and rank-order associations.

Geographic bias is framed as regional concentration or systematic disparities across contexts (for example, domestic vs. foreign options), following prior comparative work across origin groups [16,36,49]. For each origin, the distribution of recommended countries is compared using Jensen–Shannon distance (JSD) on normalised shares, which ranges from 0 to 1, and by the domestic share (the percentage of recommendations in the traveller's home country). Larger JSD values indicate more distinct geographic portfolios. Differences in domestic share are examined with binomial proportion tests and regression models that include random effects for persona.

Cultural bias is defined as skewed cultural representation that privileges dominant frames or steers travellers toward culturally distant or overly proximate contexts, consistent with calls for explicit cultural-coverage measures [14,17,50,51]. Cultural distance is summarised using Hofstede's six dimensions (PDI, IDV, MAS, UAI, LTO, IVR): for each origin, the analysis computes frequency-weighted gaps by dimension and an overall score, and then compares models via the Euclidean distance between their six-dimension profiles. Larger values denote culturally more distant recommendation portfolios. Exact formulas and handling of missing Hofstede values are documented in Appendix B.

Stereotype bias is operationalised as the presence of promotional clichés that flatten place-specific detail, in line with lexicon-based audits for generative text [18,31,46,52]. Three indicators are reported: cliché density (average cliché tokens), coverage (percentage of recommendations with at least one cliché), and diversity (distinct clichés divided by total cliché tokens). A verbosity-controlled rate (per 1000 characters) is also provided. Differences are evaluated with regression that includes persona and prompt controls; implementation details for tokenisation and matching are in Appendix B Metric Definitions and Formulas.

Demographic bias is examined as systematic differences across protected attributes (gender identity and age) when other factors are held constant, building on fairness diagnostics that emphasise multi-metric reporting and distributional similarity [18,46,52–54]. Within matched {origin, constraint} strata, the analysis summarises separation between groups using symmetric KL divergence on country-share distributions; larger values indicate greater between-group differences. As a safety-relevant probe, country shares are correlated with the Global Acceptance Index (GAI) for non-binary personas and with a general safety index (GSI) by gender [55,56]. Data alignment, smoothing, and exclusions are detailed in Appendix B Metric Definitions and Formulas.

Reinforcement bias is approached as reduced novelty caused by repeating earlier outputs, consistent with sequential and longitudinal audit recommendations [28,31,34,57]. For each persona–model chain, novelty at the third prompt is computed relative to the second (one minus the overlap of unique recommended countries). Higher novelty indicates less reinforcement. The study reports mean novelty and the share of chains with zero overlap, with set-size normalisation and deduplication rules specified in Appendix B. Table 4 summarises variability, metrics, and external data sources for each bias family. Full technical detail is available in Appendix B Metric Definitions and Formulas.

Table 4. Bias assessment table.

Bias	Variability	Metrics	External/Secondary Data
Popularity	None (intrinsic)	Probability of recommending destinations outside of Euromonitor's Top100 destinations Probability of recommending countries outside the top 30 WEF TTDI	• By city: Euromonitor Top-100 City Destinations Index [47] • By Country: Travel and Tourism Development Index (TTDI) World Economic Forum Top 30 [48]
Geographic	China, USA, Germany, United Kingdom, France, India, Japan, Saudi Arabia (Persona & location spoofing via VPN)	Pairwise Jensen–Shannon distance between country-frequency across geographic variables. Difference between models' JSD and domestic-share %	• None
Cultural	Country of origin as a proxy for traveller culture (Persona & location spoofing via VPN)	Cultural-distance score = Frequency-weighted mean absolute difference to each recommended country.	• Hofstede values (PDI, IDV, MAS, UAI, LTO, IVR) for the 8 countries: CN, US, DE, UK, FR, IN, SA, JP [51]
Stereotype	None (intrinsic)	Cliché-rate = Frequency and percentage of cliché use in each response based on 150 term tourism-stereotype lexicon	• 150-tourism cliché term lexicon (e.g., "hidden gem", "paradise", "Iconic")
Demographic (gender & age)	Persona string in prompt (male, female, non-binary) and age (25, 45, 65)	Symmetric KL divergence between country-frequency distributions of persona pairs (gender & age) Correlations of country frequency with LGBTI GAI and GSI	• LGBTI (lesbian, gay, bisexual, transgender, and/or intersex) Global Acceptance Index (GAI) [55] • Global Safety Index (GSI) by Country 2024 [56]
Reinforcement	Reinforcement of the second prompt's results with a third: "Apart from the first recommendation, could you suggest five other places that fit my profile?"	Percentage of novel recommendations by the 3rd prompt in comparison to the 2nd prompt's responses.	• None

3. Results

In brief, both models recommend many off-list destinations, with DeepSeek more exploratory at the city and country levels (+7.9 and +4.3 percentage points off-list) while still surfacing well-known hubs. Geographically, DeepSeek separates origin markets more often and more strongly (82 percent of pairs, mean Jensen–Shannon increase about 0.06) and recommends domestic travel more frequently overall (34.6 percent versus 22.8 percent), whereas ChatGPT yields tighter cross-origin convergence. Culturally, both route travellers toward destinations that are distant from home profiles, with DeepSeek generally farther; inter-model cultural distance averages 7.37 and is inversely related to domestic routing. Language is promotional for both at roughly one cliché per item, with DeepSeek denser and more concentrated and ChatGPT more varied. Demographically, separations are modest for female versus male but high to extreme for non-binary personas, especially in DeepSeek, while both models steer non-binary users toward more LGBTI-accepting

countries. Reinforcement is limited, since follow-up lists remain more than 91 percent novel in both models.

3.1. Popularity Bias

Both models surface many off-list destinations, and DeepSeek is more exploratory at both the city level (57.9 percent off Euromonitor Top-100 versus 50.0 percent) and the country level (34.8 percent outside WEF TTDI Top-30 versus 30.5 percent), although it concentrates that exploration within a smaller subset of countries. At the city level, ChatGPT produced 1020 valid recommendations and 50.0% were outside Euromonitor's Top-100. DeepSeek produced 1050 city recommendations with 57.9% off-list, a +7.9 pp difference. At the country level, 30.5% of ChatGPT's recommendations were outside the WEF TTDI Top-30 compared with 34.8% for DeepSeek. Both models still surface well-known hubs such as Tokyo, Kyoto, Lisbon, Barcelona, alongside less mainstream options such as Tuscany, Cape Town, Reykjavik. Country rankings are led by Japan for both, with consistently strong Portugal. The USA appears relatively infrequently despite a high TTDI score. Correlations between recommendation frequency and TTDI score are weak and not statistically significant for either model (ChatGPT r = 0.262, $p = 0.239$; DeepSeek r = 0.167, $p = 0.523$). ChatGPT names a larger set of off-list countries (45 vs. 24), whereas DeepSeek concentrates more heavily on a smaller subset including Iceland, South Africa, Thailand. Overall levels of popularity bias are similar, with DeepSeek modestly more inclined toward off-the-beaten-path destinations. Table 5 presents the top 20 country recommendations.

Table 5. Top 20 Country Recommendations ChatGPT vs. DeepSeek.

TOP 20 Country Recommendations ChatGPT				TOP 20 Country Recommendations DeepSeek			
Ranking	**Country**	**Frequency**	**%**	**Ranking**	**Country**	**Frequency**	**%**
1	Japan	192	88.89%	1	Japan	166	76.85%
2	Portugal	96	44.44%	2	Portugal	80	37.04%
3	Canada	85	39.35%	3	Indonesia	75	34.72%
4	New Zealand	64	29.63%	4	Iceland	65	30.09%
5	Italy	64	29.63%	5	New Zealand	64	29.63%
6	India	50	23.15%	6	South Africa	64	29.63%
7	Spain	48	22.22%	7	USA	30	13.89%
8	Iceland	41	18.98%	8	Spain	52	24.07%
9	Morocco	40	18.52%	9	Switzerland	52	24.07%
10	Switzerland	33	15.28%	10	Canada	50	23.15%
11	Germany	32	14.81%	11	Germany	47	21.76%
12	South Africa	32	14.81%	12	Italy	42	19.44%
13	Turkey	28	12.96%	13	Thailand	41	18.98%
14	USA	28	12.96%	14	Argentina	32	14.81%
15	Netherlands	25	11.57%	15	Taiwan	32	14.81%
16	Greece	19	8.80%	16	India	30	13.89%
17	Vietnam	19	8.80%	17	Greece	29	13.43%
18	Indonesia	18	8.33%	18	Netherlands	22	10.19%
19	Thailand	18	8.33%	19	Turkey	22	10.19%
20	Mexico	17	7.87%	20	Mexico	19	8.80%

For Sun & Sea, ChatGPT centres on the Mediterranean, led by Greece (63.89%), with Indonesia, Portugal, and Mexico also prominent. DeepSeek spreads attention across Asia and Africa, headed by Japan (52.78%), Kenya (44.44%), and the Maldives (43.06%),

with Costa Rica (41.67%) also frequent. For Cultural Heritage, ChatGPT focuses on the Mediterranean and North Africa, with Greece (55.56%), Morocco (52.78%), and India (51.39%) most frequent. DeepSeek elevates the Americas, led by Canada (63.89%) and Ecuador (50.0%), followed by Greece and India. For Wildlife, ChatGPT favours Latin American biodiversity hotspots, especially Costa Rica (58.33%) and Ecuador (54.17%). DeepSeek prioritises Indonesia (62.50%) and then a broader mix that includes Turkey, Mexico, and India. Overall, ChatGPT clusters around classic Mediterranean heritage and Latin American wildlife sites, whereas DeepSeek distributes recommendations more widely across Asia, Africa, and the Americas, consistent with Table 6.

Table 6. Top 10 Country Recommendations for special interest ChatGPT vs. DeepSeek.

Sun & Sea Top 10 ChatGPT				**Sun & Sea Top 10 DeepSeek**			
	Country	**Frequency**	**%**		**Country**	**Frequency**	**%**
1	Greece	46	63.89%	1	Japan	38	52.78%
2	Indonesia	33	45.83%	2	Kenya	32	44.44%
3	Portugal	28	38.89%	3	Maldives	31	43.06%
4	Mexico	28	38.89%	4	Costa Rica	30	41.67%
5	Spain	21	29.17%	5	Indonesia	29	40.28%
6	Italy	18	25.00%	6	Ecuador	26	36.11%
7	Maldives	17	23.61%	7	Spain	25	34.72%
8	Philippines	13	18.06%	8	Turkey	23	31.94%
9	Seychelles	13	18.06%	9	India	21	29.17%
10	Thailand	11	15.28%	10	Italy	20	27.78%
Cultural Heritage Top 10 ChatGPT				**Cultural Heritage Top 10 DeepSeek**			
	Country	**Frequency**	**%**		**Country**	**Frequency**	**%**
1	Greece	40	55.56%	1	Canada	46	63.89%
2	Morocco	38	52.78%	2	Ecuador	36	50.00%
3	India	37	51.39%	3	Greece	32	44.44%
4	Turkey	30	41.67%	4	India	29	40.28%
5	Japan	28	38.89%	5	Japan	24	33.33%
6	Egypt	26	36.11%	6	Seychelles	21	29.17%
7	Peru	25	34.72%	7	Italy	21	29.17%
8	Italy	23	31.94%	8	Turkey	19	26.39%
9	Mexico	21	29.17%	9	Thailand	18	25.00%
10	Uzbekistan	11	15.28%	10	South Africa	17	23.61%
Wildlife Top 10 ChatGPT				**Wildlife Top 10 DeepSeek**			
	Country	**Frequency**	**%**		**Country**	**Frequency**	**%**
1	Costa Rica	42	58.33%	1	Indonesia	45	62.50%
2	Ecuador	39	54.17%	2	Turkey	21	29.17%
3	Malaysia	38	52.78%	3	Mexico	19	26.39%
4	Botswana	26	36.11%	4	India	18	25.00%
5	Kenya	25	34.72%	5	Greece	17	23.61%
6	Australia	25	34.72%	6	Peru	16	22.22%
7	India	20	27.78%	7	Ecuador	13	18.06%
8	Indonesia	19	26.39%	8	Thailand	10	13.89%
9	South Africa	15	20.83%	9	Italy	8	11.11%
10	Argentina	15	20.83%	10	Spain	7	9.72%

3.2. Geographic Bias

Recommendation geographies cluster by origin in both models, with DeepSeek producing stronger segmentation in 82 percent of origin pairs (mean Jensen–Shannon increase about 0.06) and a higher overall domestic share (34.6 percent versus 22.8 percent). Distances within Europe are relatively low, while distances between Asian and Western origins are higher. DeepSeek yields larger distances than ChatGPT in 82% of origin pairs (mean increase 0.06; SD 0.05). For example, China's average distance to other origins is 0.50 in DeepSeek and 0.41 in ChatGPT, and the largest model gap occurs for China vs. USA at 0.15. Both models treat Japan as closest to China, suggesting regional affinity. For interpretation, values $\leq$0.30 indicate low separation, 0.31–0.50 indicate moderate separation, and $\geq$0.51 indicate high separation. As seen in Table 7, differences between ChatGPT and DeepSeek are shown as positive values when DeepSeek separates more, and negative values when ChatGPT separates more. A pre-specified heuristic effect threshold of $|\Delta| \geq 0.10$ is considered notable. For example, China–USA shows $\Delta = +0.15$ (DeepSeek higher), while UK–USA shows $\Delta = -0.04$ (small ChatGPT lead).

Table 7. Pairwise Jensen–Shannon Analysis ChatGPT vs. DeepSeek and their difference.

DeepSeek	**France**	**India**	**Japan**	**Saudi Arabia**	**United Kingdom**	**China**	**Germany**	**USA**
France	0.00	0.57	0.50	0.50	0.26	0.53	0.27	0.37
India	0.57	0.00	0.51	0.59	0.59	0.55	0.59	0.61
Japan	0.50	0.51	0.00	0.55	0.45	0.41	0.51	0.41
Saudi Arabia	0.50	0.59	0.55	0.00	0.51	0.46	0.46	0.51
United Kingdom	0.26	0.59	0.45	0.51	0.00	0.51	0.29	0.26
China	0.53	0.55	0.41	0.46	0.51	0.00	0.52	0.54
Germany	0.27	0.59	0.51	0.46	0.29	0.52	0.00	0.39
USA	0.37	0.61	0.41	0.51	0.26	0.54	0.39	0.00
ChatGPT	**France**	**India**	**Japan**	**Saudi Arabia**	**United Kingdom**	**China**	**Germany**	**USA**
France	0.00	0.50	0.36	0.49	0.26	0.38	0.25	0.27
India	0.50	0.00	0.51	0.52	0.47	0.45	0.50	0.50
Japan	0.36	0.51	0.00	0.49	0.35	0.36	0.37	0.39
Saudi Arabia	0.49	0.52	0.49	0.00	0.48	0.47	0.49	0.53
United Kingdom	0.26	0.47	0.35	0.48	0.00	0.40	0.26	0.30
China	0.38	0.45	0.36	0.47	0.40	0.00	0.41	0.39
Germany	0.25	0.50	0.37	0.49	0.26	0.41	0.00	0.32
USA	0.27	0.50	0.39	0.53	0.30	0.39	0.32	0.00
Difference Distance Between the Two	**France**	**India**	**Japan**	**Saudi Arabia**	**United Kingdom**	**China**	**Germany**	**USA**
France	0.00	0.07	0.14	0.01	0.01	0.14	0.01	0.10
India	0.07	0.00	0.00	0.07	0.13	0.10	0.09	0.11
Japan	0.14	0.00	0.00	0.06	0.10	0.06	0.14	0.02
Saudi Arabia	0.01	0.07	0.06	0.00	0.03	−0.02	−0.03	−0.01
United Kingdom	0.01	0.13	0.10	0.03	0.00	0.11	0.03	−0.04
China	0.14	0.10	0.06	−0.02	0.11	0.00	0.12	0.15
Germany	0.01	0.09	0.14	−0.03	0.03	0.12	0.00	0.06
USA	0.10	0.11	0.02	−0.01	−0.04	0.15	0.06	0.00

Note: Color coding: $\leq$0.30 = low separation (green), 0.31–0.50 = moderate (yellow), $\geq$0.51 = high (red).

Domestic recommendation rates reinforce these contrasts. DeepSeek recommends domestic travel more often overall (34.6% vs. 22.8% for ChatGPT). Examples include USA 59% vs. 4%, Japan 74% vs. 37%. India records the highest domestic emphasis in both models (85% DeepSeek, 93% ChatGPT). France is 0% in DeepSeek and 4% in ChatGPT;

Saudi Arabia is 0% in both. The pattern suggests that DeepSeek adopts a more segmented geographic strategy with stronger domestic routing for some origins, while ChatGPT shows tighter regional clusters and lower domestic shares. Figure 1 visualises domestic percentages.

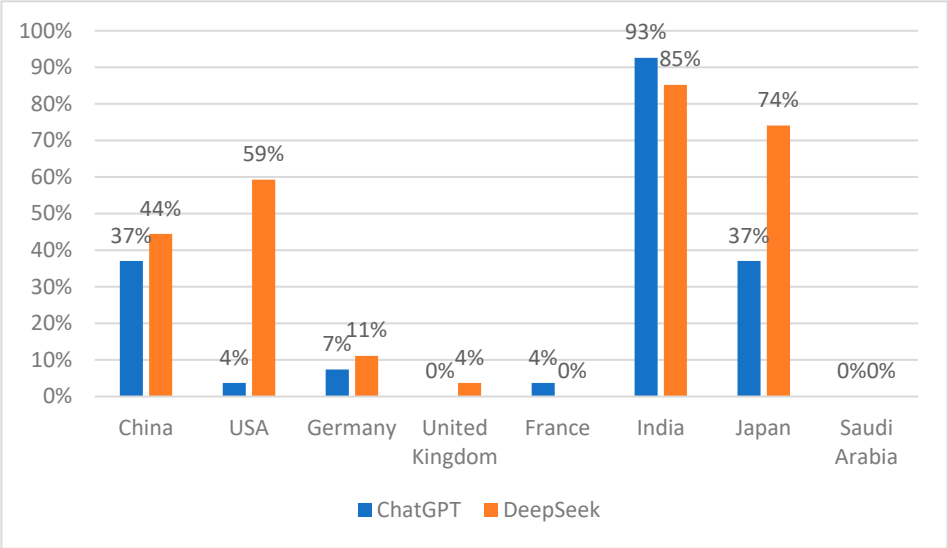

Figure 1. Domestic Travel Recommendation (%) ChatGPT vs. DeepSeek.

3.3. Cultural Bias

Across Hofstede dimensions, both models route travellers to culturally distant destinations, with DeepSeek generally farther and with inter-model cultural distance averaging 7.37 across origins. DeepSeek's mean gaps exceed ChatGPT's for most origins and dimensions, with a few clear exceptions (for example, France on MAS and Saudi Arabia on LTO). Dimension by dimension, both systems place Japan closest on PDI and MAS, and India closest on IDV, consistent with recommendations toward higher power-distance or more collectivist contexts. The largest per-origin distances appear for Saudi Arabia and China on PDI and UAI, where suggestions skew toward more egalitarian and lower uncertainty-avoidant destinations, and for France on MAS, where outputs span more masculine or feminine cultures than France's baseline. USA and United Kingdom show marked IDV distances, reflecting frequent proposals to more collectivist destinations. Table 8 reports frequency-weighted mean absolute Hofstede gaps on a 0–100 scale. Thresholds of $\leq$10 indicate very close distance, 11–25 modest distance, 26–40 marked distance, and $\geq$41 large distance. These bands are approximate heuristics for visual interpretation. Rows correspond to origins and columns to dimensions; higher values indicate greater cultural distance on that dimension.

The inter-model six-dimension (6-D) Euclidean distance averages 7.37 across origins (largest: Saudi Arabia $\approx$ 10.61; smallest: Japan $\approx$ 2.66). This difference correlates negatively with the domestic recommendation rate (ChatGPT r ≈ -0.76, DeepSeek r ≈ -0.70). When the models' cultural profiles are closer for an origin (for example, Japan, India), they recommend domestic or culturally proximate destinations more often. When the profiles diverge (for example, USA, France, Saudi Arabia), they tilt toward foreign and culturally distant options. In Figure 2, *X*-axis is inter-model cultural distance (higher = models disagree more); *Y*-axis is domestic recommendation rate (higher = more "stay/domestic"). Downward trend means greater disagreement and lower domestic routing.

Table 8. Hofstede 6-D per country of origin ChatGPT vs. DeepSeek.

Model	Origin	Power Distance (PDI)	Individualism (IDV)	Masculinity (MAS)	Uncertainty Avoidance (UAI)	Long Term Orientation (LTO)	Indulgence (IVR)
ChatGPT	China	21.4	23.5	18.36	43.74	2.21	13.76
	USA	35.21	36.17	23.8	25.03	19.51	7.14
	Germany	16.93	19.1	24.93	25.2	6.97	3.87
	United Kingdom	17.45	21.38	30.27	17.92	11.21	9.36
	France	15.94	21.59	46.75	7.69	23.59	6.06
	India	13.4	2.35	18.02	24.92	18.83	8.15
	Japan	8.21	10.58	17.37	9.47	5.53	1.58
	Saudi Arabia	38.5	20.13	20.43	44.7	25.3	12.17
DeepSeek	China	22.64	26.15	24.05	44.3	4.22	16.29
	USA	34.58	38.01	24.26	21.83	28.02	10.79
	Germany	17.37	25.48	28.59	25.94	8.1	4.59
	United Kingdom	18.12	27.49	32.12	18.45	15.03	12.98
	France	17.12	26.74	40.37	8.32	25.02	7.67
	India	14.01	3.03	21.13	25.66	20.79	9.51
	Japan	8.99	11.15	18.75	11.06	6.47	2.5
	Saudi Arabia	41.06	25.47	21.9	44.12	17.38	15.68
	Origin	**Power Distance (PDI)**	**Individualism (IDV)**	**Masculinity (MAS)**	**Uncertainty Avoidance (UAI)**	**Long Term Orientation (LTO)**	**Indulgence (IVR)**
Differences Between ChatGPT and Deepseek	China	1.24	2.65	5.69	0.56	2.01	2.53
	USA	−0.63	1.84	0.46	−3.2	8.51	3.65
	Germany	0.44	6.38	3.66	0.74	1.13	0.72
	United Kingdom	0.67	6.11	1.85	0.53	3.82	3.62
	France	1.18	5.15	−6.38	0.63	1.43	1.61
	India	0.61	0.68	3.11	0.74	1.96	1.36
	Japan	0.78	0.57	1.38	1.59	0.94	0.92
	Saudi Arabia	2.56	5.34	1.47	−0.58	−7.92	3.51

Note: Hofstede gaps are shown on a 0–100 scale. Color coding: ≤10 = very close (blue), 11–25 = modest (white), 26–40 = marked (pink), ≥41 = large distance (red). Rows represent origins and columns represent dimensions; higher values indicate greater cultural distance on that dimension.

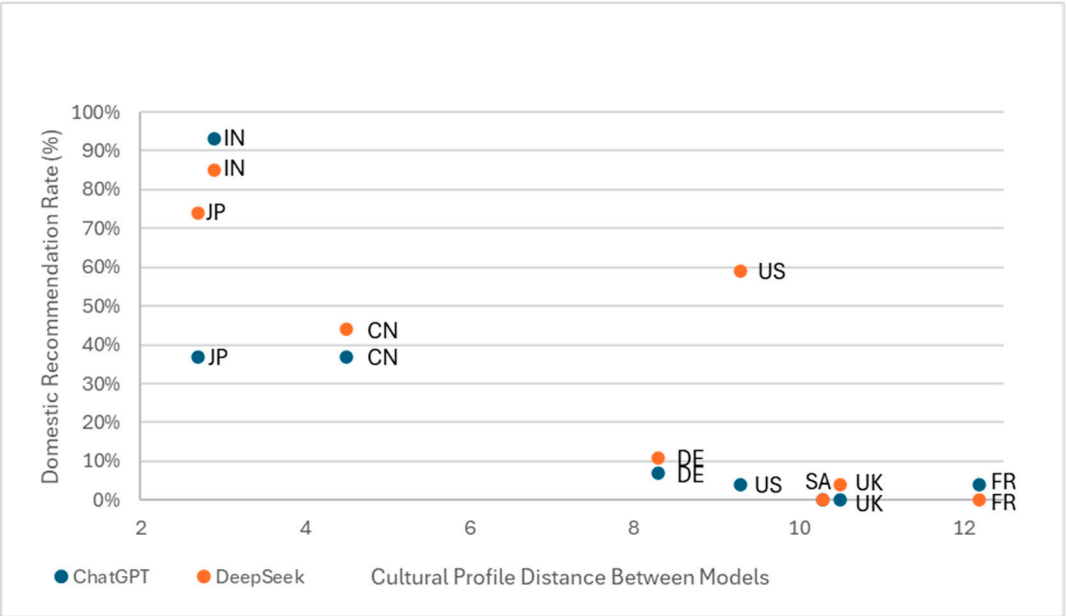

Figure 2. Model Cultural Differences vs. Domestic Rates per country and cultural profile distance ChatGPT vs. DeepSeek.

3.4. Stereotype Bias

Both systems operate in a high cliché regime at nearly one cliché per recommendation, with DeepSeek denser and more concentrated and ChatGPT more linguistically varied. Across the 2160 recommendations ChatGPT produces 1864 cliché instances (0.863 per recommendation) and DeepSeek 1931 (0.894 per recommendation), which means almost every item contains at least one trope and DeepSeek is 3.6% more cliché-dense. These levels suggest a default promotional register rather than precise, place-specific description. Figure 3 reports total cliché counts (higher means greater density) and the number of unique clichés used (higher means a broader palette). Both models sit close to one cliché per item.

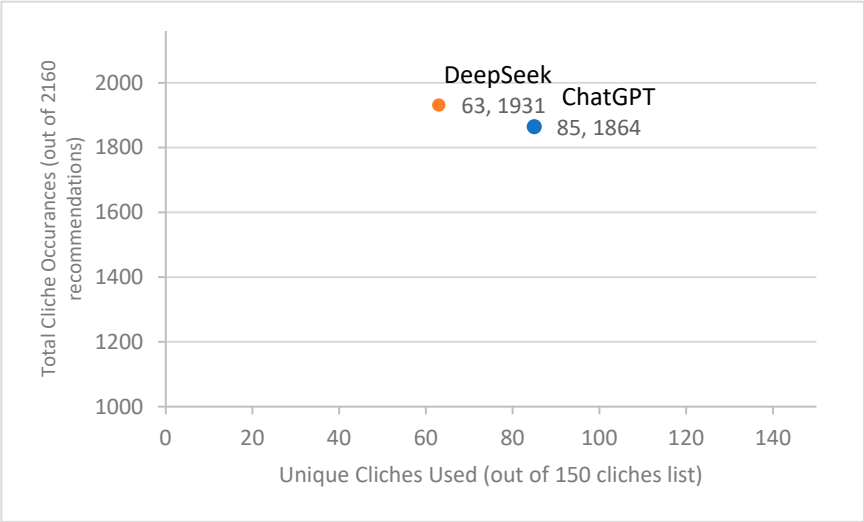

Figure 3. Cliche Usage Patterns (Unique and Total Cliche usage) ChatGPT vs. DeepSeek.

The models differ in lexical diversity versus density. ChatGPT draws on 85 distinct clichés (56.7% of the 150-term lexicon) and reaches a diversity ratio of 0.046 (distinct terms divided by total uses). DeepSeek uses 63 distinct clichés (42.0% coverage) with a diversity ratio of 0.033. In short, DeepSeek is denser, while ChatGPT is more varied. Figure 4 shows the stereotype-bias metric (higher means stronger cliché density) alongside the diversity ratio (higher means more varied language and less reuse).

Concentration and leading terms underline the contrast. DeepSeek's top-10 phrases account for 63.5% of all cliché tokens (1227 of 1931), compared with 52.2% in ChatGPT (973 of 1864). The single most frequent term in DeepSeek is "Breathtaking" with 268 uses (13.9% of all clichés). ChatGPT's top term is "Paradise" with 151 uses (8.1%). ChatGPT spreads usage over a wider set that mixes idyllic and hospitable frames ("Paradise," "Charming," "Iconic," "Friendly locals," "Classic," "Serenity"), while DeepSeek relies more on emphatic superlatives and spectacle ("Breathtaking," "World-class," "Iconic," "Vibrant nightlife," "Ultimate"). Figure 5 presents the top-10 lists and their shares. Overall, both models operate in a high stereotype-bias regime; DeepSeek is stronger through higher density and tighter concentration, and ChatGPT is more diverse yet still reproduces familiar marketing language.

Taken together, the evidence indicates a high stereotype-bias regime for both models: almost one cliché per destination on average, heavy reuse of a narrow set of promotional tropes, and limited grounding in locally specific descriptors. DeepSeek exhibits the stronger stereotype bias through higher density and higher concentration of a few superlatives. ChatGPT shows greater cliché diversity, which softens repetition but still reproduces familiar marketing language. From a user perspective, this bias risks flattening places

into interchangeable archetypes and may obscure safety, accessibility, or culturally salient details that matter for different traveler profiles.

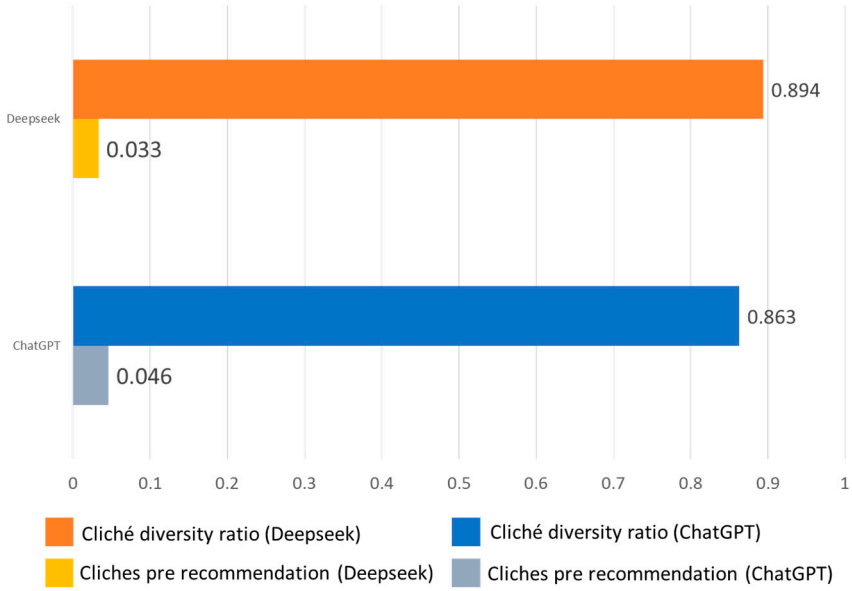

Figure 4. Stereotype Bias Metric and Cliche Diversity ChatGPT vs. DeepSeek.

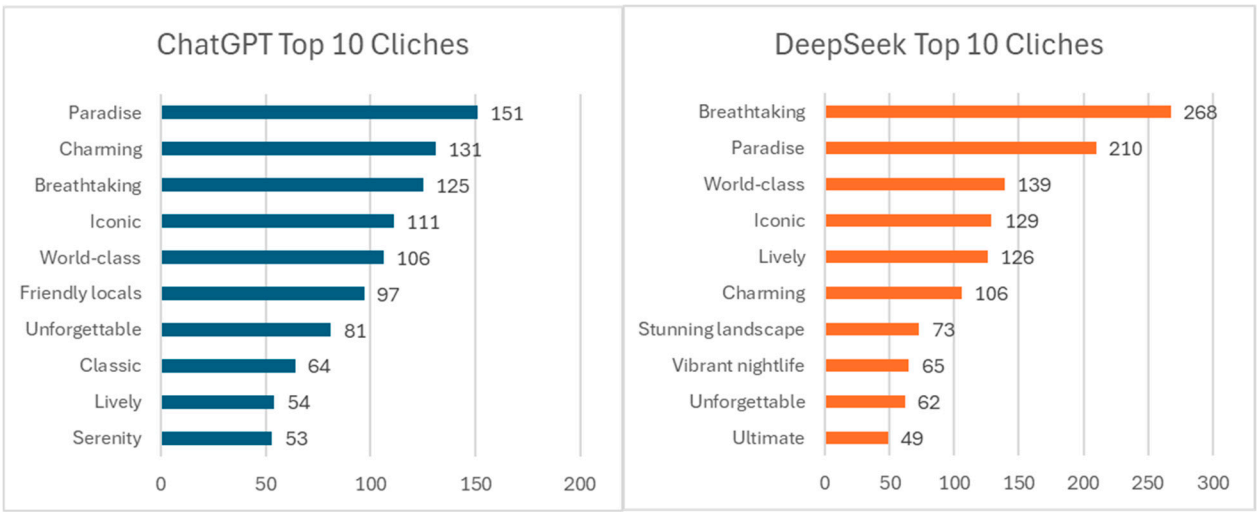

Figure 5. Top 10 Cliches ChatGPT vs. DeepSeek.

3.5. Demographic Bias

Recommendations vary systematically by gender identity and age, with the largest separations affecting non-binary personas, especially in DeepSeek, and with both models preferentially routing non-binary users toward LGBTI-accepting countries. Symmetric KL divergences between female and male personas are moderate in both models (ChatGPT 1.26; DeepSeek 1.47). Gaps increase sharply for non-binary users. For ChatGPT, female vs. non-binary = 4.87 and male vs. non-binary = 3.96; for DeepSeek, 8.77 and 5.90 respectively. DeepSeek therefore separates non-binary personas from binary personas almost twice as much as ChatGPT. According to Table 9: Female–Non-binary and Male–Non-binary show high to extreme separation (ChatGPT 4.87/3.96; DeepSeek 8.77/5.90), while Female–Male is moderate (1.26; 1.47). DeepSeek > ChatGPT on all gender pairs. Symmetric KL $\geq$ 0. 0 means identical country mix; higher = greater demographic separation.

Table 9. Symmetric KL divergences between genders ChatGPT vs. DeepSeek.

ChatGPT Gender KL				DeepSeek Gender KL			
Gender	**Female**	**Male**	**Non-Binary**	**Gender**	**Female**	**Male**	**Non-Binary**
Female	0.000	1.260	4.867	Female	0.000	1.468	8.771
Male	1.260	0.000	3.964	Male	1.468	0.000	5.897
Non-binary	4.867	3.964	0.000	Non-binary	8.771	5.897	0.000

Note: Color coding: ≤1.5 = low (blue), 1.6–3.0 = moderate (white), 3.1–5.0 = high (pink), ≥5.1 = extreme (red).

Age effects are smaller but still distinct. For ChatGPT the largest gap is 25 vs. 65 = 3.40 (high), with 25 vs. 45 = 1.58 and 45 vs. 65 = 2.31 (moderate). DeepSeek shows a similar pattern with slightly lower separation at the extremes (25 vs. 65 = 2.97; 25 vs. 45 = 1.59; 45 vs. 65 = 1.20). This indicates stronger differentiation between younger and older travellers than between adjacent age brackets. According to Table 10: The 25 vs. 65 gap is largest (ChatGPT 3.40 high; DeepSeek 2.97 high). Mid-age contrasts are smaller (25–45 ≈ 1.58–1.59 moderate; 45–65 = 1.20–2.31 low to moderate).

Table 10. Symmetric KL divergences between age groups ChatGPT vs. DeepSeek.

ChatGPT Gender KL				DeepSeek Age KL			
Age	**25**	**45**	**65**	**Age**	**25**	**45**	**65**
25	0.000	1.578	3.398	25	0.000	1.586	2.970
45	1.578	0.000	2.314	45	1.586	0.000	1.203
65	3.398	2.314	0.000	65	2.970	1.203	0.000

Note: Color coding: ≤1.5 = low (blue), 1.6–3.0 = moderate (white), 3.1–5.0 = high (pink), ≥5.1 = extreme (red).

Despite these disparities, both models show a protective signal for non-binary personas. Recommendation frequency for non-binary users correlates positively with the LGBTI Global Acceptance Index and is statistically significant in both linear and rank-order forms (ChatGPT r = 0.367, $p = 0.023$; ρ = 0.455, $p = 0.004$. DeepSeek r = 0.419, $p = 0.015$; ρ = 0.389, $p = 0.025$). Non-binary top lists cluster in high-acceptance countries such as Japan, Germany, Netherlands, Iceland, Spain and, by model, Canada/New Zealand for ChatGPT and Portugal/USA for DeepSeek. In contrast, correlations with a general safety index are weak and not significant for all genders and both models (ChatGPT r = 0.12–0.17; DeepSeek r = 0.05–0.19; all $p > 0.23$). According to Figure 6: Non-binary top lists cluster in high-acceptance countries (e.g., Japan, Germany, Netherlands, Iceland, Spain; plus Canada/NZ for ChatGPT; Portugal/USA for DeepSeek), matching the significant positive correlations reported. Higher acceptance index (point) means more LGBTI-accepting destination; higher bar means more frequent recommendation.

Recommendation frequency shows weak and non-significant correlations with a general safety index for all genders and both models (e.g., ChatGPT Pearson r = 0.12–0.17; DeepSeek r = 0.05–0.19; all $p > 0.23$). Non-binary personas have the slightly highest correlations, males the lowest, and ChatGPT is marginally more responsive than DeepSeek. In contrast to the significant LGBTI-acceptance effects, this suggests the models prioritize social acceptance over generic safety when routing different gender personas.

The models systematically vary recommendations by gender and age, with the largest and most concerning separations affecting non-binary users, particularly in DeepSeek. At the same time, both systems preferentially route non-binary personas toward more LGBTI-accepting countries, suggesting a form of safety-oriented bias alongside allocative disparities in where different groups are encouraged to travel. The combination highlights a dual mitigation need: curb excessive demographic separation while retaining explicit attention to traveller safety and acceptance.

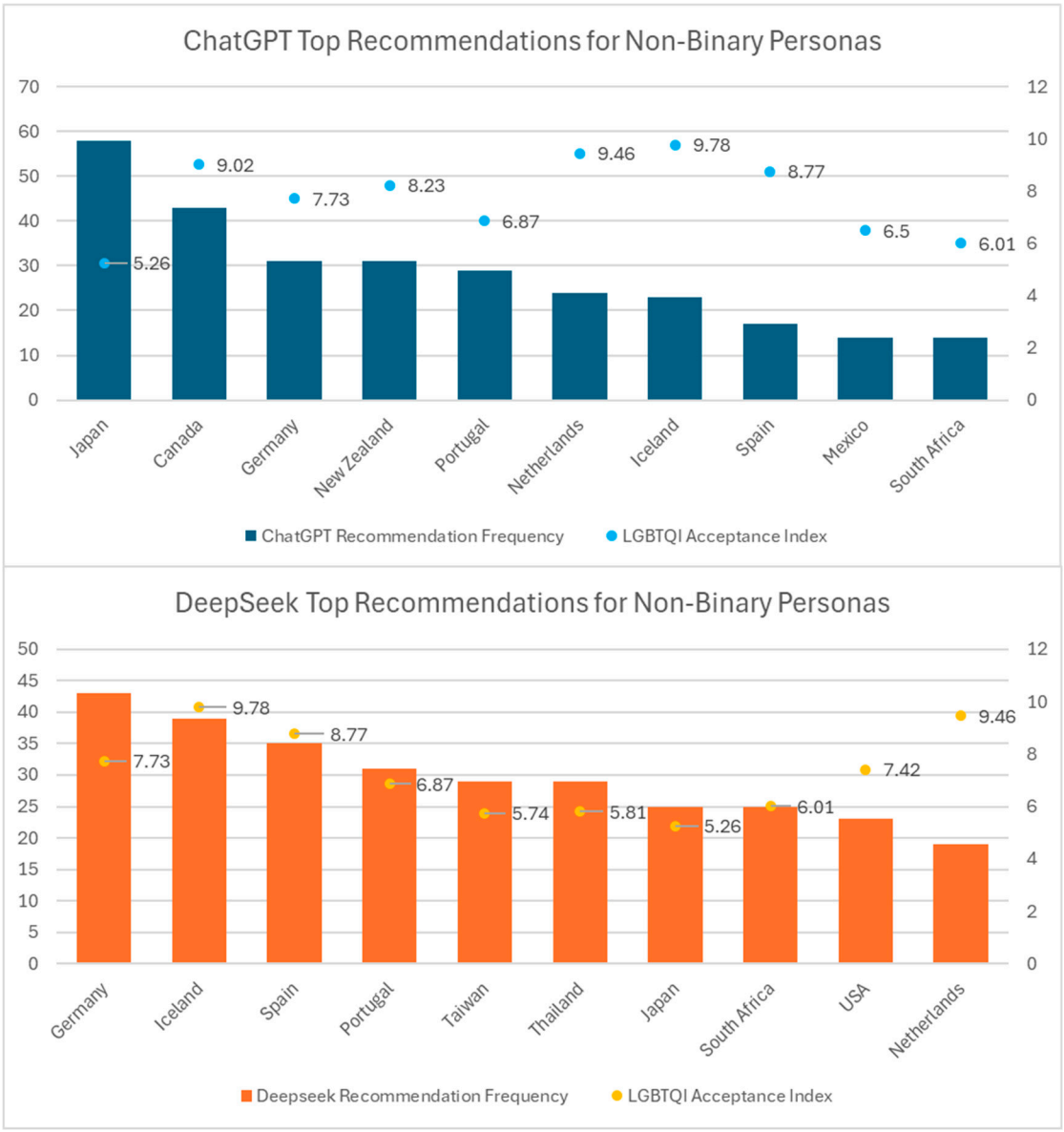

Figure 6. Top Recommendations for Non-Binary Personas with LGBTI Acceptance Index Scores ChatGPT vs. DeepSeek.

3.6. Reinforcement Bias

Reinforcement is limited for both models because follow-up lists remain highly novel (about 92 to 93 percent), with DeepSeek slightly more repetitive than ChatGPT. ChatGPT's mean overlap is 6.74% (93.26% novelty). DeepSeek's mean overlap is 8.25% (91.75% novelty). Zero-overlap cases dominate (73.1% ChatGPT; 69.9% DeepSeek). When overlap occurs it is usually small (1–20% in 17% of ChatGPT cases and 20% of DeepSeek cases). Mid-range overlap is rare (21–40% in about 7% and 6% respectively), and high overlap above 40% appears only sporadically. The pattern indicates limited reinforcement bias overall, with DeepSeek slightly more repetitive than ChatGPT while still maintaining high novelty. According to Figure 7: ChatGPT median ≈ 0% (mean 6.74% → 93.26% novelty); DeepSeek mean 8.25% (→ 91.75% novelty). Zero-overlap cases dominate (73.1% ChatGPT; 69.9% DeepSeek). 1–20% overlap is minor (17%/20%), 21–40% is rare (7%/6%), and >40% appears only sporadically. Reinforcement in a prompt chain is low for both models, with DeepSeek slightly more repetitive than ChatGPT, but high novelty is maintained for the vast majority of follow-ups.

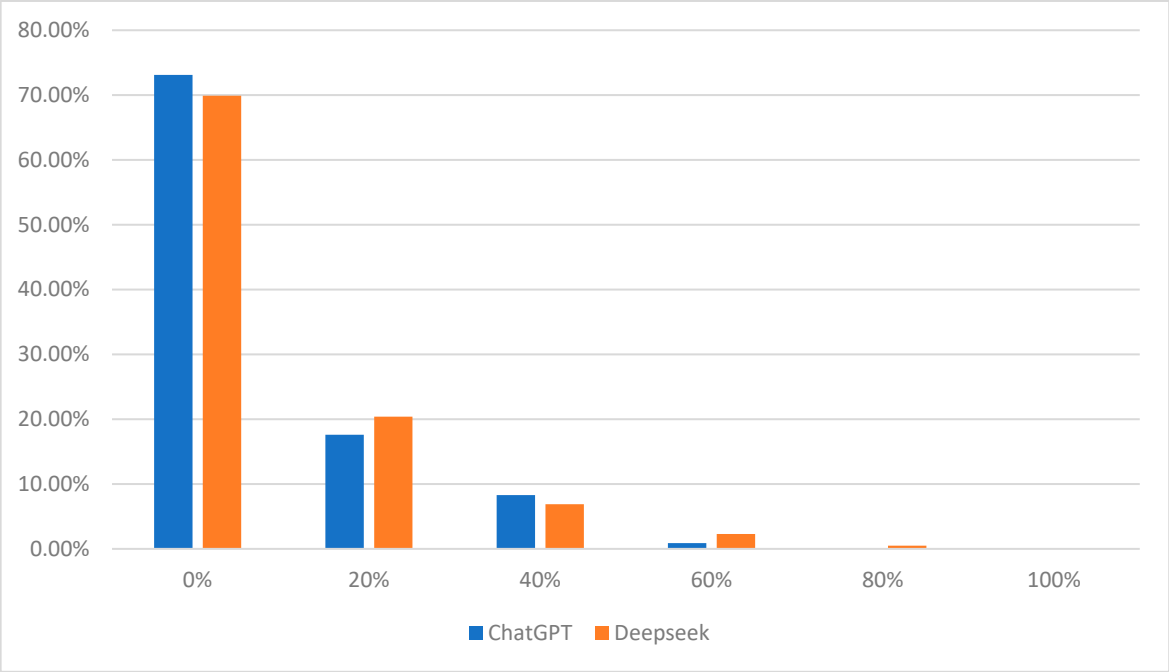

Figure 7. Distribution of reinforcement percentage ChatGPT vs. DeepSeek.

3.7. Summary of the Results

Overall, the audit finds measurable bias across all six families, with distinct signatures by model. For popularity, both systems recommend many off-list places, and DeepSeek is more exploratory at the city and country levels (+7.9 and +4.3 percentage points off-list), while weak TTDI correlations suggest that institutional rankings only loosely shape exposure. Geographic results show clustered recommendation corridors with stronger origin segmentation in DeepSeek, which has higher pairwise JSD in 82% of origin pairs (mean increase ≈0.06) and a higher domestic share overall (34.6% vs. 22.8%), whereas ChatGPT produces tighter cross-origin convergence and lower domestic routing. Cultural analysis indicates both models route travellers to culturally distant destinations, with DeepSeek generally farther; the average inter-model 6-D distance is 7.37 and it correlates negatively with domestic rates, implying that greater cultural divergence between models goes with less domestic routing. Stereotype metrics point to a promotional register in both systems at roughly one cliché per recommendation, with DeepSeek denser and more concentrated and ChatGPT more varied but still reliant on stock phrases. Demographic tests reveal the largest separations for non-binary personas (very high to extreme symmetric KL, especially in DeepSeek) and smaller age effects, yet both models steer non-binary users toward more LGBTI-accepting countries, while links to a general safety index are weak; this suggests a need to reduce allocative disparities while preserving acceptance-aware routing. Reinforcement is limited, since follow-up lists are highly novel in both models, with DeepSeek slightly more repetitive but still maintaining high novelty. Taken together, the results support targeted governance rather than one-size-fits-all fixes: dampen popularity and stereotype bias, moderate geographic and demographic separation where it is excessive, and apply cultural-distance and acceptance safeguards to keep recommendations relevant, inclusive, and practically useful.

Taken together, the results show consistent bias signatures with clear model differences. DeepSeek is more exploratory yet more segmented by origin and more cliché-dense; ChatGPT is more convergent across origins and more linguistically varied. Practically, this points to targeted governance rather than one remedy for all: moderate geographic and demographic separation where it is excessive, dampen promotional cliché density,

preserve follow-up novelty, and use cultural-distance and acceptance guardrails to keep recommendations inclusive without sacrificing utility.

4. Discussion

This study analysed six prominent bias categories in LLM travel recommendations: popularity, geographic, cultural, stereotype, demographic, and reinforcement bias. Evaluating two contemporary models under one controlled protocol provides a holistic view of how these biases co-occur and trade off. In contrast to prior audits that typically examine a single bias, or a single model, this design prioritises breadth and comparability across bias families to reveal system-level patterns rather than a deep technical treatment of any one bias in isolation [14–16,18,19,21,39,46,49].

Findings align with previous audits on regional clustering and representational skew, and they partially diverge on popularity. Prior work reports popularity concentration and geographic bias in recommendation settings, including recent evidence for LLMs [15,16,36]. It also documents WEIRD-centred cultural alignment and tourism-stereotype reproduction in AI outputs [14,17,18], and gender-conditioned differences in travel advice [19]. Here, both models surface substantial off-list content, suggesting more exploration than some earlier studies reported. A plausible reason is the broader persona coverage and controlled prompt chains used here, alongside recent alignment updates that favour variety; the task context also differs from studies focused on itinerary agents or single-persona prompts [20,43]. The weak association we observe between recommendation frequency and institutional rankings indicates that such benchmarks exert limited pull relative to corpus salience and learned associations, which is consistent with exposure-driven accounts in the literature [24].

The observed differences are modest per interaction but meaningful in aggregate. DeepSeek's higher city-level off-list rate of 57.9 percent versus 50.0 percent translates to roughly 0.4 additional off-list cities in a five-item list, or about one extra every two to three lists. At the country level, the 34.8 percent versus 30.5 percent gap is about one extra off-list country every four to five lists. Domestic routing differs more: an 11.8 percentage point gap corresponds to roughly 0.6 additional domestic items per five-item list, or about three additional domestic items over five lists. Stereotype density differs by 0.031 clichés per recommendation, which is about three extra clichés per 100 recommendations. Individually, these deltas are small; accumulated across prompts, users, and time, they shape exposure, tone, and perceived choice.

Patterns by bias family point to distinct mechanisms. Popularity and geographic results fit an emergent and interaction-bias account in which learned salience and exposure histories outweigh formal benchmarks [24,45]. Cultural, stereotype, and demographic effects are consistent with representational bias rooted in public web narratives and long-standing accounts of bias in sociotechnical systems [22,23], including WEIRD-centric frames and promotional diction identified in recent audits of LLMs and tourism content [14,17,18]. The inverse relationship between inter-model cultural distance and domestic shares suggests that internal cultural modelling, not only measured cultural gaps, conditions routing: when models align more closely on a cultural profile, they more often suggest domestic or culturally proximate trips; when they diverge, suggestions tilt outward.

Model contrasts matter in practice. DeepSeek is more exploratory overall yet more segmented by origin and more cliché-dense. ChatGPT is more convergent across origins and more linguistically varied. For travellers, two similar personas may therefore receive portfolios that differ in domestic share, regional spread, and tone. For destinations and platforms, these differences influence discoverability and distribution: stronger segmen-

tation can narrow cross-market diversity, while denser promotional diction can obscure safety, accessibility, or culturally specific details that matter to different users.

These results imply concrete governance levers that match a holistic view. Diversity-aware re-ranking and popularity-calibration can dampen concentration while reserving space for relevant long-tail items [15,35]. Geographic spread constraints and novelty floors can limit excessive origin segmentation and preserve newness across follow-ups [21,54]. Lexical quality checks that down-weight dense promotional clichés can elevate locally grounded descriptors and improve decision utility [54]. Finally, acceptance-aware routing for non-binary users should be retained, while monitoring demographic separation so that protective effects do not produce exclusionary portfolios; operationalising this requires reliable acceptance indicators such as the Global Acceptance Index [55], complemented by broader fairness and lifecycle guidance for bias management in deployed systems [7,44,53,54].

In sum, the study converges with earlier audits on core exposure and representation issues, extends them by testing six biases together under a single controlled design, and translates percentage differences into user-level counts that clarify practical significance. The effect sizes per list are small but accumulate across sessions and users, shaping what travellers see and where attention flows. Targeted re-ranking, diversity and novelty safeguards, toned-down promotional diction, and acceptance-aware guardrails offer a balanced path to improve inclusivity without sacrificing utility.

4.1. Limitations

This audit evaluates two LLMs: ChatGPT-4o and DeepSeek-V3, one build of each, under a single prompt template, default decoding, English-only interactions, and chat-only (no tools/retrieval) settings. Accordingly, the estimates speak to these specific builds, configurations, and language. They should not be generalized without caution to other model families, newer/older builds, alternative prompts (e.g., de-biased or stereotype-laden wording), non-English or right-to-left languages, tool-enabled modes, or fine-tuned deployments, where bias profiles may differ.

Several alternative explanations merit consideration. First, strict list cutoffs may overstate off-list behavior for destinations near the threshold. Second, Hofstede scores are static and applied at the national level, which limits sensitivity to regional variation and change over time. Third, parsing of regions or multi-place phrases may influence novelty estimates, although robustness checks confirm the main findings. Fourth, while session controls reduce personalization and server-side heuristics, they cannot fully eliminate these effects. Fifth, the English-only cliché lexicon may undercount culturally specific stereotypes. Sixth, the study did not manipulate prompt framing; determining how biased or stereotype-laden wording interacts with baseline model tendencies remains an open question for future research.

The training data, pre-processing pipelines, and alignment steps are proprietary and not fully documented; thus, corpus and policy influences are inferred from outputs. Safety/preference tuning and default decoding may nudge language toward generic promotion, and behavior can differ under other decoding or tool-augmented settings. Outputs also depend on model architecture and default decoding choices such as sampling temperature, nucleus probability, and length constraints; we used default chat-only generation without tools or retrieval, so behavior may differ under other settings or architectures. The scope of evaluation is bounded: two models and one build each, a single collection window, English prompts only, eight origins, three ages, three gender identities, three interest themes, and a fixed top-five recommendation task. These constraints limit causal claims and the generalizability of results across languages, time periods, deployment modes, and

model families. To limit over-generalization, the study reports build identifiers and dates, applies strict and standardized sessions, audits models in parallel on the same personas, and conducts robustness checks on regional parsing and list thresholds; replication across languages and time points, alternative cultural frameworks, and tool-enabled settings is recommended before drawing broader conclusions. This audit is cross-sectional and runs each persona and prompt in a fresh session. That choice improves internal validity by removing history effects, but it also removes any influence from earlier interactions. As a result, we do not observe path dependence, carry over, or preference shaping that may arise in longer multi-turn conversations, repeat visits, or stateful personalisation. Our reinforcement check measures novelty within a single session by comparing the second and third prompts, and it should not be taken as evidence about longer-run reinforcement or filter-bubble effects. Bias that builds up over time may therefore be underestimated. Results should be generalized with care. This study relies on Hofstede scores and on tourism lists that reflect Western travel infrastructure. Moreover, the study exclusively use English prompts with country of origin and VPN spoofing as a proxy for culture. Different cultural frameworks, non-English prompts, tool-enabled modes, or later model builds may yield different bias profiles.

Formal sensitivity analysis was outside the scope of this study. This study did not vary popularity-list cutoffs, distance metrics, cultural frameworks, decoding parameters, or language, so the estimates should be read as baseline values for this configuration. Where possible, the analysis reports both mean and rank-order associations to reduce reliance on a single statistic, but a full robustness study will require reruns that change these choices one at a time. Because the study was limited to two models, English prompts, and fixed instructions, the observed patterns may not replicate for other models (e.g., different families or guardrails), other languages (including culturally specific cliché registers), or alternative prompt designs. Formal multilingual and prompt-factorial replications are needed to establish robustness beyond this configuration. The objective here was breadth and internal control rather than parameter sweeping, and future work should extend along these sensitivity axes.

4.2. Implications

These findings describe a time-stamped snapshot of two deployed model builds under chat-only settings. They indicate what users are likely to see today, not an immutable property of all LLMs or deployment modes. The findings suggest that fairness cannot be expected from unrestricted, general-purpose LLMs. Instead, travel recommendation systems should be purpose-built with clearly defined objectives and constraints that can be adjusted under the oversight of a public-interest steward. One promising approach is a two-layer architecture for an AI system that combines LLM with parametrization mechanism: the LLM generates a broad set of candidate destinations, and a transparent re-ranking parametrization mechanism selects the final list based on multiple objectives. In practice, this is a simple two-step flow: the LLM proposes a larger shortlist (e.g., 20 places), and a transparent scoring sheet: set by a neutral steward and picks the final five based on a few pre-set goals/conditions such as: fairness, seasonality, carbon, safety, and novelty. Each recommendation is explained with plain-language reasons shown to users.

Public-interest governance could be provided by a neutral organization, such as UN Tourism, responsible for managing and configuring the re-ranking layer while openly publishing its objectives and weighting criteria. To ensure integrity, governance frameworks should explicitly prohibit pay-to-play arrangements involving financial contributions from destinations or companies. Advertising-style incentives must not affect the probability of being recommended. A public registry of model versions, objectives, and audit results

would further enhance transparency and accountability. This is practical to deploy because it mirrors how travel sites already sort results.

The re-ranker should optimise recommendations for a balanced set of goals:

- Exposure fairness: Quotas or minimum exposure budgets for underrepresented regions and for off list destinations while preserving relevance. Budgets can be set per origin market and per theme.
- Seasonality smoothing: Time aware objectives that lift destinations in low and shoulder periods and gently suppress those at peak load. Inputs include historical arrivals, accommodation occupancy and event calendars. Access to such data enables dynamic adjustment of recommendation priorities based on real-time or predicted demand patterns. In doing so, the system can be parametrized to favor more sustainable travel behaviors by encouraging visits to less congested areas and periods, mitigating over-tourism, and enhancing traveler experience and destination resilience.
- Low carbon routing: A transport aware objective that prioritises itineraries with lower estimated emissions. Defaults should favour domestic or short haul options when comparable in utility and increase the score for train, coach and ferry access while discouraging unnecessary long haul flights. Emission estimates can be computed with standard factors and shown to users. When emissions are made visible during the decision-making process, users are more likely to shift their preferences toward more eco-friendly travel options, reinforcing sustainable behaviour through informed choice.
- Cultural congruence bounds: Limits should be applied to cultural distance to avoid systematically directing users from high power distance or high uncertainty avoidance societies to culturally mismatched destinations, unless the user explicitly requests novelty.
- Safety and acceptance safeguards: Positive weights should be assigned to LGBTI acceptance and general safety indices for minority groups, with adjustable thresholds. The system should notify users when recommended destinations fall below predefined safety or acceptance levels.
- Stereotype penalties: Cliché density in destination descriptions should be minimized, while concrete and place-specific details such as names of protected areas, museum collections, trail difficulty levels, or community initiatives should be rewarded. The goal should be to associate authenticity and real characteristics to destinations instead of cliché marketing phrases.
- Diversity and novelty: Similarity controls should be applied to prevent near-duplicate recommendations within and across sessions, supported by clearly defined novelty targets.
- Bias dashboards and audits: Platforms should provide live dashboards displaying key metrics such as off-list rates, domestic destination shares, geographic Jensen–Shannon distances by origin, demographic symmetric KL divergence gaps, cliché density and lexical diversity, as well as indicators for seasonality and carbon emissions. Predefined thresholds should trigger alerts and prompt automatic adjustments to weighting parameters. Independent audits based on persona matrices should be conducted regularly, with full audit reports made publicly available.
- User experience: Interfaces should clearly communicate public interest objectives in accessible language and allow users to adjust settings within safe limits, such as opting for more environmentally friendly or off-peak travel options. Explanations should include the reasons for selecting a destination, the estimated CO2 emissions of the itinerary, and how the recommendation aligns with fairness or seasonality goals.

- Because many biases are not immediately visible at the point of use, interfaces should incorporate lightweight transparency features and guardrails that function effectively at scale. Examples include "why this was recommended" explanations, simple diversity or novelty indicators, and optional user controls that permit minor adjustments within safe limits. Such measures do not require users to conduct audits but make systemic safeguards more transparent. They also help align the objectives of the re-ranking layer with user understanding and trust.

Underrepresented regions can enhance their visibility by publishing open, machine-readable datasets that include information on seasonality, infrastructure capacity, transport connectivity, trail networks, cultural events, accessibility features, and sustainability certifications. Supplying reliable emission factors for common travel routes and making LGBTI acceptance and safety information available in multiple languages can support re-ranking systems in meeting key safeguards. Destination managers may also work in partnership with the steward organization to define exposure budgets aligned with local carrying capacity and community objectives.

4.3. Suggestions for Future Research

Future research should assess multi-objective re-ranking in real-world settings, measuring outcomes such as the distribution of visitor flows, reductions in CO_2 emissions, alleviation of overcrowding, and traveler satisfaction. This should include small live pilots or simple A/B tests with tourism boards or travel sites to see how recommendations work in practice. Multilingual audits are essential to determine whether patterns of stereotype and cultural bias persist beyond English-language outputs; these should compare native-language prompts (for example Arabic, Hindi, Chinese, Spanish), English, and straightforward machine-translated versions, and should also cover right-to-left scripts. Longitudinal studies can examine model drift and the effectiveness of seasonality-focused objectives over time. More detailed intersectional analyses that combine gender identity, age, country of origin, and interest theme will help identify where disparities have the greatest impact; results should also make clear where improving one bias may worsen another, so readers can see the trade-offs. Human evaluations of explanation clarity and cultural sensitivity will offer valuable insights alongside automated metrics. Finally, partnerships with public-interest stewards can provide a testing ground for governance frameworks that preserve recommendation neutrality while promoting measurable sustainability outcomes.

The prompt design in this study elicited short rationales for each recommended destination ("...give reasons for each"; "...explain why"). In the present analysis, these texts were used only to extract destinations and were not analysed qualitatively. Future work should apply qualitative content analysis (for example, thematic coding) and simple sentiment analysis to these rationales to examine tone, persuasiveness, hedging, and cliché usage alongside the stereotype lexicon. These checks should be repeated in multiple languages to see if the framing changes by language. Such analyses would show whether the explanations themselves embed promotional or demographic or cultural skews, how this framing interacts with destination exposure, and whether patterns differ across languages.

Another direction for future research is to understand how prompt wording interacts with model-based bias. A factorial design could test variations in prompt framing (neutral, stereotype-laden, or de-biased), constraint specificity (general, thematic, or budget or safety related), and language (English vs. other major languages). Bias outcomes could be measured using the same metrics applied here. Mixed-effects models could then estimate both main effects and interactions, clarifying how much bias stems from user input versus the model, and whether prompts amplify, reduce, or reverse underlying biases.

4.4. Final Remarks

The persona-based audit demonstrates that conversational LLMs already display measurable biases in travel recommendations, including popularity, geographic, cultural, stereotype, demographic, and reinforcement biases. The extent and nature of these biases vary by model. DeepSeek-V3 is more inclined to suggest off-list destinations, applies stronger segmentation by user origin, recommends domestic travel more frequently, and relies on a concentrated set of superlative descriptors. In contrast, ChatGPT-4o produces more convergent recommendations across different origins and uses a broader vocabulary, though it still leans heavily on clichéd language. Both models show significant demographic variation in recommendations, particularly for non-binary users, while also demonstrating a tendency to favor destinations with higher levels of LGBTI acceptance. Reinforcement bias remains minimal, indicating that prompt-level controls are effective in maintaining novelty.

Because models, decoding defaults, and guardrails evolve, periodic re-audits are needed before applying these conclusions to future versions or other languages. These results suggest that fairness in AI travel recommendation systems cannot be assumed based solely on personalization or accuracy. Achieving equitable outcomes requires clearly defined goals around diversity, cultural alignment, and user safety, supported by ongoing audits and transparent reporting. With appropriate safeguards in place, LLM-based travel planning has the potential to promote more inclusive and balanced discovery, ensuring broader representation of destinations and fairer distribution of attention across travelers.

Author Contributions: Conceptualization, H.A., P.K., A.D.L., A.T. and A.Z.; methodology, H.A., P.K., A.D.L., A.T. and A.Z.; software, H.A.; validation, H.A., A.T., and A.Z.; formal analysis, H.A.; investigation, H.A., P.K. and A.D.L.; resources, A.T., and A.Z.; data curation, H.A.; writing—original draft: H.A., P.K. and A.D.L.; preparation, H.A.; writing—review and editing, H.A., P.K., A.D.L., A.T. and A.Z.; visualization, H.A., supervision, A.T. and A.Z.; project administration, H.A. and P.K.; funding acquisition, P.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets are available upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Session Controls and Execution Settings

A fresh browser context and a new model session were created before each persona–model chain. No plugins, tools, or retrieval features were enabled, and no manual edits were made to model outputs. The team issued each prompt once. If the model returned fewer than five destinations, one regeneration was allowed within the same session. If the response still contained fewer than five, the chain was rerun in a new instance after clearing cookies and changing the IP via the VPN service. If a refusal or safety notice occurred, a fresh session was started once. All settings used vendor defaults; temperature, top p, system prompts, safety filters, tools, and retrieval were not changed. This improves internal consistency but limits the examination of alternative settings. The reinforcement prompt explicitly excluded the first recommendation from the second prompt ("Apart from the first recommendation. . ."), and novelty was evaluated only relative to the set returned by the second prompt. The 1st prompt (generic) establishes a baseline of recommendations by using only geographic and demographic parameters, and it requests reasons for each suggestion so that rationales can be analysed for stereotyping. The single-constraint prompt then introduces a specific interest or tourism focus to obtain more targeted results and to

test how preference information reshapes the distribution of destinations. The reinforcement follow-up prompt asks for five other places apart from the first recommendation in the second prompt to test whether the model produces genuinely new destinations rather than repeating earlier choices, which probes novelty and reinforcement effects. The fixed sequence from baseline to constrained to follow-up mirrors user behaviour and supports within-persona comparisons under identical wording, which strengthens repeatability and connectivity across prompts. For every response, the dataset stored the raw text, timestamps, model identifier, and persona metadata. The cleaning pipeline comprised destination extraction and normalisation, deduplication, popularity tagging, cultural distance scoring, stereotype-lexicon features, and joins to an acceptance index for demographic fairness. Destination extraction parsed each recommendation into a city, when present, and a country. Country names were mapped to ISO 3166-1 alpha-3 [45]: codes, and ambiguous city names were geocoded to the corresponding country. Items referring to regions, such as the Balkans, were mapped to all constituent countries for frequency analyses and flagged as regional. Deduplication collapsed repeated items within a response.

Appendix B. Metric Definitions and Formulas

Popularity bias, operationally defined as a systematic preference for already popular items that reduces exposure to the long tail of infrequent yet relevant options. In the literature it is commonly measured with coverage and diversity indices and long-tail exposure analyses [15,35,45]. This study measures popularity as the percentage of recommendations that appear in widely used popularity indices (on-list rate) and its complement, the off-list share. City-level items are compared with Euromonitor's Top 100 City Destinations Index [46] and country-level items with the World Economic Forum Travel and Tourism Development Index (TTDI) Top 30 [47]. Higher on-list rates indicate stronger popularity skew; higher off-list shares indicate broader exposure. The team produced ranked frequency tables of countries and cities, including overall Top 20 lists and theme-specific lists for Sun and Sea, Cultural Heritage, and Wildlife. City and country frequencies were cross-referenced with the two indices to estimate the probability that an item fell outside each list. Model differences were tested with proportion tests and mixed-effects logistic regression including prompt type and persona covariates. Country recommendation frequency was also correlated with TTDI scores using Pearson and Spearman coefficients to assess linear and rank-order associations.

Geographic bias, operationally defined as regional concentration or systematic disparities in recommendations across geographic contexts, such as domestic versus foreign options or over- and under-representation by region. Prior work contrasts distributions across origin groups [16,36,48]. Here, measurement uses (i) the Jensen–Shannon distance (bounded 0 to 1) between normalised country-share distributions by origin and (ii) the domestic share, defined as the percentage of recommendations that fall in the traveller's home country. For each origin–model pair, recommendations were mapped to countries (cities mapped to their country; regional labels handled per the data-cleaning rules) and pooled across the 27 personas for that origin. Origin-level distributions were formed by dividing each country's count by the origin total. Pairwise Jensen–Shannon distances among origins were computed within each model and summarised via matrix averages and agglomerative clustering; larger values indicate greater dissimilarity in geographic emphasis. Differences in domestic share were tested with binomial proportion tests and with regression including random effects for persona.

Cultural bias, operationally defined as skewed cultural representation that privileges dominant frames or routes users toward culturally distant or overly proximate contexts. Audits call for explicit cultural-coverage measures [14,17,49]. This study quan-

tifies cultural distance using Hofstede's six dimensions (PDI, IDV, MAS, UAI, LTO, IVR). Let o denote the traveller's origin country and c, a recommended country; let $k \in \{PDI, IDV, MAS, UAI, LTO, IVR\}$ index Hofstede dimensions. Let $H_k(x) \in [0, 100]$ be the Hostede score of country x on dimension k For each origin–country pair (o,c), the per-dimension gap is:

$$dk(o, c) = \mid Hk(o) - Hk(c) \mid$$

For a set of recommendations with frequencies w_c (renormalized to $\sum w_c = 1$ after any exclusions), the mean gap on dimension k is

$$\overline{d}_k(o) = \sum_c w_c d_k(o, c)$$

The aggregate cultural distance for origin o is the simple average across dimensions:

$$\overline{d}(o) = \frac{1}{6} \sum_k \overline{d}_k(o)$$

Cities are mapped to countries. Items missing a Hofstede value on a given dimension are excluded for that dimension only and the remaining w_c are renormalised. Multi-country items are split equally across listed countries (sensitivity checks vary this split). For each origin, the inter-model cultural-profile distance between ChatGPT and DeepSeek is the Euclidean distance between their six-element vectors of mean gaps:

$$D(o) = \left\| d^{(ChatGPT)}(o) - d^{(DeepSeek)}(o) \right\|_2,$$
$$where\ d^{(model)}(o) = \left(\overline{d}_{PDI}(o), \overline{d}_{IDV}(o), \overline{d}_{MAS}(o), \overline{d}_{UAI}(o), \overline{d}_{LTO}(o), \overline{d}_{IVR}(o) \right)$$

All distances are in the units of the Hofstede scale (0–100 per dimension); thus $D(o) \in \left[0, \sqrt{6} \times 100\right]$, with 0 indicating identical cultural profiles.

Stereotype bias, operationally defined as the propagation or amplification of learned societal stereotypes, often realised as promotional clichés that flatten place-specific detail. Detection commonly uses association tests and lexicon-based audits for generative text [18,31,45,51]. This study measures stereotype bias as (i) cliché density (mean cliché tokens per recommendation), (ii) cliché coverage (percentage of recommendations containing at least one cliché), and (iii) cliché diversity (distinct clichés divided by total cliché tokens). Counts derive from a 150-term tourism cliché lexicon. Text was lower-cased, normalised for hyphen, space, and Unicode variants, and matched with token-boundary-aware patterns using longest-match precedence to avoid double counting. Metrics were computed per recommendation and then aggregated; a verbosity-controlled rate (clichés per 1000 characters) was also reported. Top-phrase lists were produced for each model. Model differences in density and coverage were tested with regression including persona and prompt controls; robust standard errors were used. A relative bias index and severity bands are reported as descriptive, study-specific summaries.

Demographic bias, operationally defined as systematic disparities in recommendation portfolios across protected attributes, here gender identity and age, when other factors are held constant. Foundational fairness metrics include demographic parity and equalised odds; recent work recommends demographic-similarity diagnostics and multi-metric reporting [18,45,51–53]. Following this rationale, demographic bias is measured as the symmetric KL divergence between country-share distributions across gender and across age groups, with small additive smoothing to avoid undefined values. Larger values indicate greater between-group separation. To hold origin and constraint constant, distributions were computed within matched {origin, constraint} strata, symmetric KL was calculated

within each stratum, and results were macro-averaged across strata for each model. As a safety-relevant probe, the analysis also computed Pearson and Spearman correlations between country-level recommendation shares and the Global Acceptance Index for non-binary personas [54], and between recommendation shares and a general safety index for each gender group [55]. Analyses were conducted separately by model using the set of countries present in both sources after standardising names and removing missing values. Symmetric KL divergences were calculated for every pair of groups to form square matrices for gender and age within each model, then summarised into bias-severity bands and contrasted across models.

Reinforcement bias, operationally defined as amplification in which systems repeat or intensify prior outputs, reducing novelty and entrenching exposure patterns. Audits recommend sequential and longitudinal checks [28,31,34,57]. This study uses a within-session sequential measure: novelty percentage, defined as the percentage of destinations in the third prompt that were not present in the second prompt for the same persona–model chain. Its complement is the overlap rate. Higher novelty indicates minimal reinforcement; higher overlap indicates stronger reinforcement. Researchers computed novelty for each chain, summarised averages across personas, and counted the share of chains with zero overlap. Let $S_{2,i}$ and $S_{3,i}$ be the sets of unique recommended countries returned at the second and third prompts for chain i (cities are mapped to countries; duplicates within a prompt are deduplicated). Let $|A|$ denote set cardinality. The overlap rate and its complement, novelty, are:

$$overlap_i = \frac{|S_{2,i} \cap S_{3,i}|}{|S_{3,i}|},\ novelty_i = 1 - overlap_i\ (\times 100\%\ for\ percentage)$$

Typically $|S_{3,i}| = 5$ (five items requested); if fewer are parsed, we normalise by $|S_{3,i}|$ and report robustness checks that exclude ambiguous cases. We summarise mean novelty across personas and the share of chains with zero overlap.

Table 4 presents the bias assessment framework, detailing variability, metrics, and secondary data for each bias.

References

1. Banerjee, A.; Banik, P.; Wörndl, W. A review on individual and multistakeholder fairness in tourism recommender systems. *Front. Big Data* **2023**, *6*, 1168692. [CrossRef]
2. Seyfi, S.; Kim, M.J.; Nazifi, A.; Murdy, S.; Vo-Thanh, T. Understanding tourist barriers and personality influences in embracing generative AI for travel planning and decision making. *Int. J. Hosp. Manag.* **2025**, *126*, 104105. [CrossRef]
3. Bulchand Gidumal, J.; Secin, E.W.; O'Connor, P.; Buhalis, D. Artificial intelligence's impact on hospitality and tourism marketing: Exploring key themes and addressing challenges. *Curr. Issues Tour.* **2023**, *27*, 2345–2366. [CrossRef]
4. Chu, C.H.; Donato-Woodger, S.; Khan, S.S.; Nyrup, R.; Leslie, K.; Lyn, A.; Shi, T.; Bianchi, A.; Rahimi, S.A.; Grenier, A. Age related bias and artificial intelligence: A scoping review. *Humanit. Soc. Sci. Commun.* **2023**, *10*, 1–17. [CrossRef]
5. Leonard, J.; Mohanapriya, J.D.C. Research on artificial intelligence in tourist management. *Int. J. Multidiscip. Res. Sci. Eng. Technol.* **2025**, 8. [CrossRef]
6. Sousa, A.E.; Cardoso, P.; Dias, F. The use of artificial intelligence systems in tourism and hospitality: The tourists' perspective. *Adm. Sci.* **2024**, *14*, 165. [CrossRef]
7. O'Flaherty, M. Bias in Algorithms—Artificial Intelligence and Discrimination (FRA Report No. 8). European Union Agency for Fundamental Rights. 2022. Available online: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf (accessed on 13 June 2025).
8. Suanpang, P.; Pothipassa, P. Integrating generative AI and IoT for sustainable smart tourism destinations. *Sustainability* **2024**, *16*, 7435. [CrossRef]
9. Rasheed, H.M.W.; Chen, Y.; Khizar, H.M.U.; Safeer, A.A. Understanding the factors affecting AI services adoption in hospitality: The role of behavioural reasons and emotional intelligence. *Heliyon* **2023**, *9*, e16968. [CrossRef]
10. Arıbaş, E.; Dağlarlı, E. Transforming personalized travel recommendations: Integrating generative AI with personality models. *Electronics* **2024**, *13*, 4751. [CrossRef]

11. Bhardwaj, S.; Sharma, I.; Kaur, G.; Sharma, S. Personalization in tourism marketing based on leveraging user generated content with AI recommender systems. In *Redefining Tourism with AI and the Metaverse*; IGI Global: Hershey, PA, USA, 2025; pp. 317–346. [CrossRef]
12. Kong, H.; Wang, K.; Qiu, X.; Cheung, C.; Bu, N. Thirty years of artificial intelligence research relating to the hospitality and tourism industry. *Int. J. Contemp. Hosp. Manag.* **2023**, *35*, 2157–2177. [CrossRef]
13. He, K.; Long, Y.; Roy, K. Prompt-Based Bias Calibration for Better Zero/Few-Shot Learning of Language Models. *arXiv* **2024**, arXiv:2402.10353. [CrossRef]
14. Tao, Y.; Viberg, O.; Baker, R.S.; Kizilcec, R.F. Cultural bias and cultural alignment of large language models. *PNAS Nexus* **2024**, *3*, 346. [CrossRef] [PubMed]
15. Klimashevskaia, A.; Jannach, D.; Elahi, M.; Trattner, C. A survey on popularity bias in recommender systems. *User Model. User Adapt. Interact.* **2024**, *34*, 1777–1834. [CrossRef]
16. Manvi, R.; Khanna, S.; Burke, M.; Lobell, D.; Ermon, S. Large language models are geographically biased. *arXiv* **2024**, arXiv:2402.02680. [CrossRef]
17. Mihalcea, R.; Ignat, O.; Bai, L.; Borah, A.; Chiruzzo, L.; Jin, Z.; Kwizera, C.; Nwatu, J.; Poria, S.; Solorio, T. Why AI is WEIRD and shouldn't be this way: Towards AI for everyone, with everyone, by everyone. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 28657–28670. [CrossRef]
18. Zhu, J.; Zhan, L.; Tan, J.; Cheng, M. Tourism destination stereotypes and generative artificial intelligence (GenAI) generated images. *Curr. Issues Tour.* **2024**, *28*, 2721–2725. [CrossRef]
19. Chen, Y.; Li, H.; Xue, T. Female gendering of artificial intelligence in travel: A social interaction perspective. *J. Qual. Assur. Hosp. Tour.* **2023**, *26*, 1057–1072. [CrossRef]
20. Ren, R.; Yao, X.; Cole, S.; Wang, H. Are large language models ready for travel planning? *arXiv* **2024**, arXiv:2410.17333. [CrossRef]
21. Mitchell, S.; Potash, E.; Barocas, S.; D'Amour, A.; Lum, K. Algorithmic fairness: Choices, assumptions, and definitions. *Annu. Rev. Stat. Its Appl.* **2021**, *8*, 141–163. [CrossRef]
22. Friedman, B.; Nissenbaum, H. Bias in computer systems. *ACM Trans. Inf. Syst. TOIS* **1996**, *14*, 330–347. [CrossRef]
23. Suresh, H.; Guttag, J.V. A framework for understanding unintended consequences of machine learning. *arXiv* **2019**, arXiv:1901.10002. [CrossRef]
24. Baeza-Yates, R. Bias on the web. *Commun. ACM* **2018**, *61*, 54–61. [CrossRef]
25. Barocas, S.; Selbst, A.D. Big data's disparate impact. *Calif. Law Rev.* **2018**, *104*, 671. [CrossRef]
26. Belenguer, L. AI bias: Exploring discriminatory algorithmic decision making models and the application of possible machine centric solutions adapted from the pharmaceutical industry. *AI Ethics* **2022**, *2*, 771–787. [CrossRef]
27. Chen, F.; Wang, L.; Hong, J.; Jiang, J.; Zhou, L. Unmasking bias in artificial intelligence: A systematic review of bias detection and mitigation strategies in electronic health record based models. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 1172–1188. [CrossRef]
28. Ferrara, E. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci* **2024**, *6*, 3. [CrossRef]
29. Nazer, L.H.; Zatarah, R.; Waldrip, S.; Ke, J.X.C.; Moukheiber, M.; Khanna, A.K.; Hicklen, R.S.; Moukheiber, L.; Moukheiber, D.; Ma, H.; et al. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLoS Digit. Health* **2023**, *2*, e0000278. [CrossRef]
30. Pessach, D.; Shmueli, E. Improving fairness of artificial intelligence algorithms in privileged group selection bias data settings. *Expert. Syst. Appl.* **2021**, *185*, 115667. [CrossRef]
31. Pulivarthy, P.; Whig, P. Bias and fairness: Addressing discrimination in AI systems. In *Ethical Dimensions of AI Development*; Bhattacharya, P., Hassan, A., Liu, H., Bhushan, B., Eds.; IGI Global: Hershey, PA, USA, 2025; pp. 103–126. [CrossRef]
32. Pasipamire, N.; Muroyiwa, A. Navigating algorithm bias in AI: Ensuring fairness and trust in Africa. *Front. Res. Metr. Anal.* **2024**, *9*, 1486600. [CrossRef]
33. Jain, L.R.; Menon, V. AI algorithmic bias: Understanding its causes, ethical and social implications. In Proceedings of the 2023 IEEE 35th International Conference on Tools with Artificial Intelligence, Atlanta, GA, USA, 6–8 November 2023; IEEE: New York, NY, USA, 2023; pp. 460–467. [CrossRef]
34. Ferrer, X.; Van Nuenen, T.; Such, J.M.; Coté, M.; Criado, N. Bias and discrimination in AI: A cross-disciplinary perspective. *arXiv* **2020**, arXiv:2008.07309. [CrossRef]
35. Forster, A.; Kopeinik, S.; Helic, D.; Thalmann, S.; Kowald, D. Exploring the effect of context awareness and popularity calibration on popularity bias in POI recommendations. *arXiv* **2025**, arXiv:2507.03503. [CrossRef]
36. Dudy, S.; Tholeti, T.; Ramachandranpillai, R.; Ali, M.; Li, T.J.J.; Baeza-Yates, R. Unequal opportunities: Examining the bias in geographical recommendations by large language models. In Proceedings of the 30th International Conference on Intelligent User Interfaces, Cagliari Italy, 24–27 March 2025; Association for Computing Machinery: New York, NY, USA, 2025; pp. 1499–1516. [CrossRef]

37. Gupta, O.; Marrone, S.; Gargiulo, F.; Jaiswal, R.; Marassi, L. Understanding Social Biases in Large Language Models. *AI* **2025**, *6*, 106. [CrossRef]
38. Marsh. AI's Travel Bias: Study Reveals UK Cities Most Affected by Artificial Intelligence's Travel Recommendations. *The Scotsman*, 25 January 2024. Available online: https://www.scotsman.com/travel/glasgow-ranked-as-top-uk-city-impacted-by-ais-travel-bias-according-to-study-5095134 (accessed on 15 May 2025).
39. Voutsa, M.C.; Tsapatsoulis, N.; Djouvas, C. Biased by Design? Evaluating Bias and Behavioral Diversity in LLM Annotation of Real-World and Synthetic Hotel Reviews. *AI* **2025**, *6*, 178. [CrossRef]
40. Tsai, C.Y.; Wang, J. A personalized itinerary recommender system: Considering sequential pattern mining. *Electronics* **2025**, *14*, 2077. [CrossRef]
41. Jamader, A.R.; Chowdhary, S.; Dasgupta, S.; Kumar, N. From promotion to preservation and rethinking marketing strategies to combat overtourism. In *Solutions for Managing Overtourism in Popular Destinations*; IGI Global: Hershey, PA, USA, 2025; pp. 379–398. [CrossRef]
42. Foka, A.; Griffin, G. AI, cultural heritage, and bias: Some key queries that arise from the use of GenAI. *Heritage* **2024**, *7*, 6125–6136. [CrossRef]
43. Singh, H.; Verma, N.; Wang, Y.; Bharadwaj, M.; Fashandi, H.; Ferreira, K.; Lee, C. Personal large language model agents: A case study on tailored travel planning. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, Miami, FL, USA, 12–16 November 2024; pp. 486–514. [CrossRef]
44. Werder, K.; Cao, L.; Ramesh, B.; Park, E.H. Empower diversity in AI development. *Commun. ACM* **2024**, *67*, 31–34. [CrossRef]
45. International Organization for Standardization. *ISO 3166-1:2020(en), Codes for the Representation of Names of Countries and Their Subdivisions—Part 1: Country Codes (Alpha-3 Code)*; International Organization for Standardization: Geneva, Switzerland, 2020; Available online: https://www.iso.org/standard/72482.html (accessed on 27 June 2025).
46. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. CSUR* **2021**, *54*, 1–35. [CrossRef]
47. Euromonitor International. Top 100 City Destinations Index 2024. 2024. Available online: https://www.euromonitor.com/top-100-city-destinations-index/report (accessed on 27 June 2025).
48. World Economic Forum. Travel & Tourism Development Index 2024. 2024. Available online: https://www.weforum.org/publications/travel-tourism-development-index-2024/ (accessed on 27 June 2025).
49. Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdl, W.; Vidal, M.-E.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; et al. Bias in data-driven artificial intelligence systems: An introductory survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1356. [CrossRef]
50. Elsharif, W.; Alzubaidi, M.; Agus, M. Cultural bias in text-to-image models: A systematic review of bias identification, evaluation, and mitigation strategies. *IEEE Access* **2025**, *13*, 122636–122659. [CrossRef]
51. The Culture Factor. Country Comparison Tool—Hofstede. Available online: https://www.theculturefactor.com/country-comparison-tool (accessed on 30 June 2025).
52. Dominguez-Catena, I.; Paternain, D.; Galar, M. Metrics for dataset demographic bias: A case study on facial expression recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5209–5226. [CrossRef]
53. Aninze, A. Artificial intelligence life cycle: The detection and mitigation of bias. In Proceedings of the International Conference on AI Research (ICAIR), Lisbon, Portugal, 5–6 December 2024; Volume 4. [CrossRef]
54. Schwartz, R.; Vassilev, A.; Greene, K.; Perine, L.; Burt, A.; Hall, P. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*; US Department of Commerce, National Institute of Standards and Technology: Gaithersburg, MD, USA, 2022; Volume 3. [CrossRef]
55. Williams Institute. Global Acceptance Index (LGBTI), 2017–2020. Available online: https://williamsinstitute.law.ucla.edu/projects/gai/ (accessed on 3 July 2025).
56. Numbeo. Safety and Crime Indices—Methodology. 2024. Available online: https://www.numbeo.com/crime/indices_explained.jsp (accessed on 3 July 2025).
57. Samala, A.D.; Rawas, S. Bias in artificial intelligence: Smart solutions for detection, mitigation, and ethical strategies in real-world applications. *IAES Int. J. Artif. Intell.* **2025**, *14*, 32–43. [CrossRef]