Are They Playing Favourites? Preferences for Institutions, Brands and Cultures in LLMs

Jasmine Rienecker
Stupid Human
Lisbon, Portugal

Katarina Mpofu
Stupid Human
Sweden

Naman Goel
University of Oxford and Alan Turing Institute
Oxford, UK

Siddhartha Datta
University of Oxford
Oxford, UK

Jun Zhao
University of Oxford
Oxford, UK

Oscar Danielsson
Stupid Human
Sweden

Fredrik Thorsen
Stupid Human
Sweden

Abstract

AI systems increasingly act as choice architects for billions, shaping what people see and trust. There is a pressing need to audit not only the bias towards different social attributes but also how these systems rank institutions, brands, and cultures. This paper introduces ChoiceEval, a reproducible framework for evaluating such biases in large language models (LLMs) under realistic, open-ended usage. The framework segments users into psychographic profiles (e.g., budget-conscious, wellness-focused), accompanied with prompts that reflect real-world advice-seeking and decision-making behaviour. Applied to Gemini, GPT, and DeepSeek across 20 topics spanning government, commerce, and culture and 4,140 questions, ChoiceEval reveals consistent preferences for certain countries, industries, and institutions—often aligned with the model's origin. U.S.-developed models Gemini and GPT show marked favouritism toward American entities, while China-developed DeepSeek exhibits more balanced yet still detectable geographic preferences. These patterns persist across user personas, suggesting systematic rather than incidental effects. ChoiceEval provides a scalable audit pipeline for researchers, platforms, and regulators, linking model behaviour to real-world economic and civic outcomes.

CCS Concepts

• **Information systems → Web searching and information discovery**; • **Human-centered computing**;

Keywords

Large Language Models, Auditing, Evaluation, Bias, Trust

1 Introduction

The rapid adoption of AI assistants such as ChatGPT, Google Gemini, and Meta AI has fundamentally transformed how individuals interact with technology and access information. A notable shift is emerging in information-seeking behaviour, as conversational AI systems like ChatGPT gain prominence alongside traditional search engines, with a quarter of respondents in a recent Adobe Express study reporting they use ChatGPT as their primary choice for search [25]. Responding to this trend, Google has integrated AI Overviews (formerly Search Generative Experience) into its search platform, positioning AI-generated summaries as a primary interface for information retrieval. These systems increasingly serve as primary intermediaries between users and vast information sources, wielding considerable influence over institutional perceptions and decisions regarding products and services. Moreover, emerging evidence suggests a generational dimension to ChatGPT adoption patterns, with OpenAI CEO Sam Altman observing that younger users, particularly college students, increasingly integrate ChatGPT into their workflows as a comprehensive personal assistant rather than merely a search tool [20]. Given this widespread use of AI systems to inform decisions about what to buy, whom to trust, and how to navigate everyday choices, it becomes imperative to audit the underlying preferences and potential biases embedded within these models.

Research on large language model (LLM) bias has largely focused on *social bias*: the systematic, unfair associations that advantage or disadvantage people based on socially constructed categories such as gender, race, ethnicity, religion, political view or sexual orientation. The research community has developed numerous well-designed benchmark datasets as a practical and efficient way to surface these social biases. Frameworks and benchmarks such as RealToxicityPrompts [13], BBQ [34], BOLD [11], WinoBias [48], CrowS-Pairs [29] and StereoSet [28] provide established methodologies for detecting and quantifying the kinds of systematic inequities that often reflect, and can perpetuate, historical societal disparities. This focus aligns with emerging standards like IEEE Std 7003-2024, which requires algorithm designers to proactively search for unintended biases, including those that would remain hidden unless explicitly tested for. However, while the field has made significant progress in measuring social bias, substantially less attention has been devoted to measuring what this paper terms *entity-perception bias*: an AI model's tendency to systematically favour specific institutions, brands, and cultures over comparable alternatives, resulting in their over-representation or preferential treatment.

Entity-perception bias is particularly consequential in open-ended user queries that elicit AI recommendations, where the manner in which language models present and frame options can profoundly shape user perceptions and subsequent choices. It has been well-established that even subtle variations in option presentation can significantly alter user behaviour, so the design of choice environments, through defaults, feedback, or option ordering, can effectively nudge users toward specific outcomes [45]. This architectural influence is particularly pronounced in digital contexts, where users are demonstrably swayed by the sequence in which choices appear. Glick et al. [14] showed that higher-ranked websites receive disproportionately more clicks regardless of actual relevance, as users interpret ranking itself as a signal of trustworthiness and quality.

This paper contributes ChoiceEval, a comprehensive framework for systematically generating evaluation questions to assess entity-perception bias in AI assistants. The methodology involves three key steps: first, establishing psychographic user clusters grounded in established consumer segmentation research such as VALS [16, 27, 37, 41]; second, translating these clusters to topic-specific contexts; and finally, generating questions that consumers within each cluster might reasonably ask during the consideration phase of their decision-making journey. Using this framework, this paper also presents a ready-to-use dataset of open-ended, realistic questions spanning 20 governmental, commercial and cultural topics. Additionally, the framework provides a generalisable tool that enables researchers to generate contextually relevant evaluation questions for any topic of interest, ensuring broader applicability beyond the initial dataset.

We demonstrate the framework's utility by applying our ready-to-use dataset to address two critical research questions: (RQ1) Do AI assistants exhibit stable preferences when recommending institutions, brands and cultural entities?; and (RQ2) Do AI assistant recommendations exhibit geographic bias across these contexts?

2 Related Work

Over the past decade, substantial progress has been made in identifying and quantifying social biases in AI systems, particularly through the development of evaluation frameworks and benchmarks. Foundational work such as Bolukbasi et al. [6] and Caliskan et al. [9] demonstrated that word embeddings encode and propagate stereotypes, for example by associating certain professions with specific genders. Standardised datasets such as StereoSet [28] and CrowS-Pairs [29] were then introduced to measure social biases across protected characteristics, including gender, race and religion, in a controlled, replicable manner. In parallel, work in toxic language detection revealed how classification systems often imposed disproportionate harms on marginalised groups. Park et al. [33] demonstrated that abusive language models frequently misclassified neutral sentences containing gendered terms as sexist, highlighted fairness as a core evaluation concern and paving the way for fairness-focused benchmarks in classification tasks [15]. Beyond these efforts, other dimensions of bias have recently attracted attention: for instance, Zhou et al. [49] explore how large language models may encode financial biases, potentially skewing economic reasoning and decision-making tasks.

While these frameworks have advanced our understanding of social bias, they generally operate on curated prompts or fixed evaluation sets. Less attention has been paid to scenarios resembling real-world use, where model outputs are shaped by open-ended, user-driven queries. Notable exceptions include Sheng et al. [40], who showed that free-form continuations from language models often reproduce gender, racial, and religious stereotypes, and Alnegheimish et al. [3], who demonstrated that the design of evaluation prompts (synthetic templates versus natural sentences) substantially affects measured levels of gender occupation bias, with natural prompts yielding more realistic and less exaggerated results. Despite these advances, this line of research remains relatively small and continues to concentrate primarily on social biases, with entity-perception bias receiving much less attention.

Research into entity-perception bias in large language models is still nascent but growing. Kamruzzaman et al. [18] revealed significant sentiment disparities in LLM outputs when describing global compared to local brands, with global brands receiving many more positive associations. Similarly, Cao et al. [10] showed that ChatGPT's responses align strongly with American cultural values, performing best when prompted in US contexts and English, while flattening or misrepresenting cultural distinctions elsewhere; alignment improves in local languages but still lags human cultural baselines. Tao et al. [44] extended this analysis using World Values Survey data, showing GPT models skew toward Western, English-speaking norms. They also demonstrated that explicit 'cultural prompting' can reduce this bias more effectively than local language use, though the dominance of English-language data and Western market forces continues to entrench these cultural defaults. Pawar et al. [35] surveys how cultural awareness in LLMs is defined and measured, reviewing datasets, prompting strategies, and ethical considerations. Geographic bias has also been examined: Bhagat et al. [5] link it to global wealth inequalities; Lalai et al. [21] show models favor the Global North and West; Manvi et al. [24] reveal broad prejudice across objective and subjective topics, introducing a bias score to compare models; and Zhang et al. [47] show models struggle with truthful reasoning about less-represented regions. These studies indicate early evidence of such biases but stop short of establishing standardized, repeatable methodologies for measuring them, especially in recommendation and ranking contexts.

Recommender system bias has also been studied extensively, though primarily through the lens of personalisation fairness. Research has examined popularity bias (over-promoting well-known items), exposure bias (certain groups or items receiving disproportionately low visibility), and consumer–provider fairness trade-offs. Abdollahpouri et al. [1], for instance, propose a personalised re-ranking method that mitigates popularity bias by boosting long-tail items according to each user's preferences, improving item diversity with minimal accuracy loss. Complementarily, Mehrotra et al. [26] address fairness in two-sided marketplaces and demonstrate that recommendation policies optimised exclusively for user relevance disproportionately privilege "superstar" providers.Other work has focused on domain-specific implications of LLM outputs in recommendation-like settings: Manchanda and Shivaswamy [23] highlights how name bias in text embeddings can distort thematic similarity assessments and proposes anonymisation strategies to mitigate this; Geerts et al. [12] illustrates how LLMs could enhance transparency and stakeholder decision-making in real estate appraisal; and Noyman et al. [30] introduces TravelAgent, an agent-based simulation platform to study pedestrian movement, activity, and human decision-making in built environments, offering a new lens for evaluating LLM-driven spatial recommendations. Nevertheless, the intersection of recommender fairness principles with open-ended, natural-language recommendation queries generated by LLMs remains largely unexplored.

Our work bridges these research streams by introducing a benchmark dataset specifically designed to evaluate biases in AI-generated rankings and recommendations across institutions, brands, and cultural entities. Unlike prior benchmarks focused primarily on social bias in controlled settings, our framework targets open-ended recommendation scenarios, enabling assessment of entity-perception bias. This design creates the first scalable foundation for evaluating how AI assistants may shape real-world decisions, ensuring that entity-perception bias is not only detected but also tracked as these systems evolve.

3 ChoiceEval: A Framework for Evaluating Entity-Perception Bias

ChoiceEval is a reproducible framework for generating evaluation datasets to assess entity-perception bias in LLMs under realistic usage scenarios. The framework produces open-ended, recommendation-seeking prompts that elicit preferences and rankings, emphasising consumer-facing decision contexts where presentation and ordering plausibly shape choices. The framework also extracts the top five recommendations from each response via an automated validation step, yielding structured, analysis-ready outputs that reflect the options most likely to influence user decisions. The model-agnostic and topic-agnostic design ensures transferability across assistant-style LLMs and domains.

Importantly, ChoiceEval is specifically designed for consumer-oriented scenarios such as product purchases, travel destinations, entertainment choices, or lifestyle decisions. This focus distinguishes it from sociographic or demographic analysis tools, concentrating instead on contexts where commercial and cultural preferences naturally intersect with individual decision-making processes.

All prompts, scripts, and topic questions will be released openly to support ongoing benchmarking and policy-aligned audits.

3.1 User Cluster Definitions

Within the framework, questions are generated using a consumer clustering approach, which analyses how different user types interact with AI assistants. While traditional demographic and geographic segmentation models are widely used, they do not fully account for the open-ended, exploratory ways in which users engage with AI systems. To address this, the user clusters identified in our framework (Table 1) build on established psychographic segmentation frameworks, most notably VALS (Values and Lifestyles) originally developed by SRI International [27] and widely applied in consumer research [41].

VALS demonstrates how underlying values and lifestyle orientations translate into distinct patterns of motivation and behaviour, providing a structured basis for differentiating user personas beyond surface-level demographics. Complementing this, our clustering approach is also informed by seminal works such as [37], that introduced lifestyle segmentation as a practical tool for understanding market diversity, and [16] that empirically showed how cultural values shape consumer decision-making. Thus, our clustering approach highlights the contexts and motivations behind users' prompts. By varying personas from budget-conscious to innovation-driven, our framework integrates these insights to capture a wide range of ways users engage with LLMs across domains and decision-making contexts.

User Cluster	Definition
Performance and Quality	Prioritizing high standards and durability in recommendations
Budget-Conscious	Seeking cost-effective options over premium choices
Innovation and Technology	Favouring cutting-edge advancements and new releases
Health and Wellness	Focusing on fitness, nutrition, and well-being
Ethical and Environmental	Preferring sustainability and social responsibility
Convenience	Valuing ease of use, accessibility, and efficiency
Experience and Lifestyle	Valuing experiences, personal interests, and lifestyle alignment
Novice vs. Expert Levels	Questions from different knowledge levels
Cultural and Social Influences	Exploring regionally or culturally driven preferences

Table 1: The psychographic user clusters used in our framework to capture the underlying values and motivations that shape how different user types engage with AI systems.

To ensure topic-specific relevance, our framework employs an LLM to adapt the core consumer clusters for each domain, translating their general characteristics into terminology and concerns specific to that decision-making context. Once these topic-cluster pairs are established, an LLM is again used to generate questions that consumers in the consideration phase of their journey might naturally ask within each adapted context. For example, when adapting clusters for the *Universities* topic, the Health and Wellness cluster became *Student Wellness and Health Services*, with a corresponding question being: *"Which universities have initiatives for stress management?"*. In this paper, we demonstrate this framework using GPT-4o for both the cluster translation and question generation steps.

3.2 Extraction of Responses

Once the questions are generated, they are used to systematically prompt the LLMs under evaluation. Each LLM is queried with the same set of questions, and its responses are recorded. From each response, the top five recommendations are extracted (Figure 1). This follows prior research on consumer decision-making which demonstrates that people rarely evaluate more than a handful of options, typically between 3 and 5, when forming a consideration set [43]. In addition, numerous studies on position bias have shown that top-ranked items dominate user choices regardless of quality. For example, research by Joachims et al. [17] revealed that search engine users heavily favour top results, with the first three options capturing over 70% of all clicks. In addition, while users view the top two results with nearly equal frequency, they ultimately select the first option almost three times more often than the second. Likewise, Lerman and Hogg [22] demonstrated that identical news stories attract roughly five times more votes when placed at the top of the list, compared to those ranked lower on the list. This

Question: Which universities have notable research parks or incubators?

GPT Response:
Many universities around the world have established notable research parks and incubators to foster innovation, entrepreneurship, and collaboration between academia and industry. Here are some universities with prominent research parks or incubators:

1. **Stanford University** — Stanford Research Park: One of the first technology-focused parks, it has fostered companies like Hewlett-Packard and Google.
2. **Massachusetts Institute of Technology (MIT)** — MIT Innovation Initiative: Includes resources like the MIT Media Lab and the Martin Trust Center for MIT Entrepreneurship.
3. **University of California, Berkeley** — SkyDeck: A startup accelerator that provides resources for entrepreneurs to develop their ideas into viable companies.
4. **University of Cambridge** — Cambridge Science Park: The oldest science park in the UK, housing companies in various technology and biotech sectors.
5. **University of Oxford** — Oxford Science Park: Supports companies in life sciences, medtech, and other innovative sectors.

[Further Text...]

Extraction:

1. Stanford
2. Massachusetts Institute of Technology (MIT)
3. University of California, Berkeley
4. University of Cambridge
5. University of Oxford

Figure 1: Example Prompt, Model Response, and Information Extraction for the Universities Topic (GPT-4o).

motivated our focus on the first five results as those the most likely to influence user decisions.

To extract the recommendations, we simulate a multi-expert setting using an LLM (GPT-4o in our experiments), where the model is instructed to act as five independent experts analysing each text, with final outputs determined by majority consensus among these simulated expert perspectives. This ensemble methodology reduces individual interpretation biases and enhances extraction reliability by requiring convergence across multiple independent analyse. A sample of the extracted recommendations should then manually be verified by human reviewers to ensure validity. Subsequent analysis is conducted on these five extracted recommendations, treating them as the effective decision set that best approximates real-world user exposure and choice contexts.

4 Analyzing Biases with ChoiceEval

With extracted recommendations, one can perform statistical analysis to address various research questions of interest. In this paper, our analysis is focused on the following two key questions:

1. Do AI assistants exhibit stable preferences when recommending institutions, brands and cultural entities?
2. Do AI assistant recommendations exhibit geographic bias across these contexts?

4.1 Topic Selection

To ensure comprehensive coverage of potential biases, this study evaluated 20 topics spanning three key areas:

- **Governmental** (Countries to Live In, Government-Run Healthcare, Governments), capturing how AI assistants reflect geopolitical and policy-related biases, and whether they favour certain states or political perspectives.
- **Commercial** (Airlines, Cloud Computing Services, Electric Vehicles, Hotel Chains, Laptops, Online Dating Platforms, Running Shoes, Smartphones, Social Media Platforms, Telecommunication Services), probing AI-driven brand and service recommendations to uncover potential market favouritism that could influence consumer decision-making.
- **Cultural** (Commodities for Investment, Sports, Travel Destinations, Universities, Vegetables, Weekend Getaway Cities, Wine Regions), exploring whether AI outputs align with global or region-specific cultural practices, thereby highlighting risks of cultural homogenisation or geographic bias.

ChoiceEval was used to generate 23 questions per topic-cluster pair, resulting in 207 questions per topic and 4,140 questions in total. These questions were then all manually verified by the authors for correctness. This methodology ensures comprehensive coverage of authentic user prompting behaviours while maintaining consistency in cluster representation across governmental, commercial, and cultural contexts. The resulting dataset provides a robust foundation for analysing genuine human-AI interaction patterns in decision-making scenarios, capturing a wide spectrum of how different user types naturally engage with AI assistants when seeking information and recommendations.

Each of the 4,140 questions was asked in a fresh chat session with no previous history, context, persona, or additional information provided. This approach was intentionally designed to simulate how typical users interact with AI assistants: going straight to the point without providing extensive context or background information. Following each initial response, additional follow-up questions were asked to probe the sources of claims and reasoning behind the AI's answers. While this follow-up data was collected and remains available for future analysis, it was not utilised in the current paper's findings. The research was conducted in Sweden using Swedish IP addresses with all interactions conducted in English.

4.2 Selection of LLM Models

The study focuses on ChatGPT-4o, Google Gemini 1.5-flash and DeepSeek-V3. The Gemini and GPT models were selected for their widespread adoption: at the time of the study [1], OpenAI had 300 million active users [32] with over 1 billion queries daily [36], while Google Gemini had over 1 billion users in search access alone [46]. DeepSeek-V3 was included as a state-of-the-art non-US model, providing a meaningful counterpoint to the American AI Assistants.

4.3 Statistical Analysis of Variability

In this study, Spearman's Rank Correlation [42] was used to measure the consistency of each AI assistant's recommendations across five repetitions of each of the topics Countries to Live In and Laptops. Repeating each task five times follows established practices in survey research and psychometric reliability testing, which emphasize repeated measurement to ensure consistency [31], while also

---

[1] March 2025

addressing the stochastic variability of AI model outputs. Spearman's Rank Correlation is a non-parametric statistic that assesses the strength and direction of association between two ranked lists, assuming only that the relationship is monotonic. A high Spearman's value indicates strong consistency in the rankings across repetitions.

The Kruskal-Wallis test [19] was also used to confirm that the recommendations across the five repetitions came from the same underlying distribution. It is a non-parametric test with a non-significant result indicating that the rankings are statistically consistent across iterations, so providing additional validation of the model's stable preferences.

4.4 Analysis of Geographic Bias Across Topics

To examine potential geographic bias in AI assistant recommendations, we analysed recommendation patterns across the eleven location-relevant topics: Airlines, Hotel Chains, Electric Cars, Laptops, Online Dating Platforms, Running Shoes, Smartphones, Social Media Platforms, Telecommunication Providers, Universities, and Weekend Getaway Cities. All analysis code will be made publicly available to enable reproduction and extension of these findings. For each topic, we identified how frequently different entities appeared in the top five suggestions within each user groups. To create a manageable yet representative dataset, for each topic entities appearing in fewer than 5% of the responses were excluded from the analysis.

Each entity was then assigned to its primary geographic region. For each topic, regional bias was quantified by calculating Log Odds Ratios (LOR) [2] between region-pairs for each individual user group. The Log Odds Ratio was chosen because it provides an intuitive measure of preference strength and direction: an odds ratio of 14 for US vs Asia, for example, means the model is 14 times more likely to suggest a US entity over an Asian entity, while taking the logarithm allows for symmetric interpretation of preferences in either direction and enables standard statistical testing. To assess statistical significance, for each topic the average LOR was calculated, and a one-sample t-tests was performed against the null hypothesis of no geographic bias ($LOR = 0$).

In addition to topic-level analyses, a random-effects meta-analysis [7] was conducted to synthesise LORs across the included topics. This approach produces a weighted average of the individual topic-level LORs, with weights determined by the inverse of their within-topic variances (precision) and adjusted for the estimated between-topic variance (τ^2). The random-effects model was selected instead of a fixed-effect model because the true underlying LOR for bias toward specific brands and services may reasonably vary across topics, reflecting differences in how brand presence, market penetration, cultural relevance, and competitive landscapes influence recommendations in areas such as airlines, hotel chains, consumer electronics, or online platforms. Unlike the fixed-effect approach, which assumes a single common LOR across all topics, the random-effects model explicitly incorporates heterogeneity via τ^2, producing more conservative standard errors and confidence intervals when variability is present. This modelling choice aligns with established methodological recommendations [8] for meta-analyses in contexts where heterogeneity is anticipated or observed.

This modelling choice is grounded in the premise that, if geographic bias in brand and service recommendations is stable and systemic, it should manifest consistently across diverse user groups and topics, whereas variations would more likely reflect cultural predispositions or domain-specific factors. While this approach cannot definitively eliminate all unobserved confounders, the combination of multiple user groups and diverse topics provides a foundation for distinguishing between genuine geographic bias and patterns reflecting legitimate user preferences or contextual needs. Aggregating results across topics and running global significance tests provides evidence for any overall significant geographic biases, while accounting for both psychographic and topic-specific variation.

5 Results

The findings of this study reveal that AI assistants exhibit strong, stable preferences, with US-based assistants showing a notable geographic bias toward American entities.

5.1 AI Assistants Exhibit Strong Preferences

The examination of the recommendation patterns demonstrates that all three AI assistants exhibit strong preferences across a diverse range of topics. Averaging across all categories, DeepSeek, GPT and Gemini recommend their top-ranked brand, government, or organisation in 61%, 65% and 70% of the 207 responses, respectively. Table 2 further illustrates the persistence of these preferences: in a quarter of the topics both Gemini and GPT include their preferred entity within their top five recommendations more than 90% of the time. While DeepSeek's preferences are somewhat less pronounced, it still demonstrates notable consistency by including its favoured entity in over 70% of responses across 9 topics. This pattern indicates that **all three models operate with established preferences that persist regardless of user cluster or query formulation, suggesting algorithmic predispositions rather than contextually adaptive recommendations.**

The high correlation coefficients observed in Spearman's Rank Correlation analysis further demonstrates that, in addition to maintaining a persistent preference structure, all three AI assistants produce highly stable recommendations. As shown in Appendix Tables 4, 6 and 8 correlation coefficients for Laptops are consistently above 0.952 for all three AI assistants. While for Countries to Live In (Appendix Tables 5, 7 and 9) the values are slightly lower, they all still exceed 0.834 for Gemini, 0.882 for GPT and 0.750 for DeepSeek. The corresponding p-values are also highly significant ($p < 0.05$), notably with values all below 0.003 for Gemini and 0.0008 for GPT. These findings suggest that, **when faced with the same set of 207 questions at scale, the models would repeatedly favour the same entities**.

Taken together, **the persistence of preferences and the stability of outputs point to selection patterns that operate largely independently of contextual changes.**

The Kruskal-Wallis test provides further evidence of this consistency, showing that the strength of the preferences also remains stable across interactions. For the Laptops category, the test returned a chi-square statistic of 0.0167 ($p = 0.99$) for GPT, 0.355 ($p = 0.99$) for Gemini and 0.0626 ($p = 0.99$) for DeepSeek. Similarly,

	Gemini		GPT		DeepSeek	
Topic	**Top Preference**	**% responses**	**Top Preference**	**% responses**	**Top Preference**	**% responses**
Countries to live in	Sweden	34.3	Germany	36.7	Germany	35.7
Government-Run Healthcare	US	57.5	UK	89.6%	US	95.2%
Governments	Canada	54.2	Singapore	45.6	Singapore	44.4
Airlines	Qatar	40.7	Emirates, Singapore Airlines	50.3	Emirates, Singapore Airlines	49.8
Cloud Computing Services	Microsoft, Google Cloud	100%	Microsoft, Google Cloud	100%	Google Cloud	99.0%
Electric Cars	Tesla	90.7%	Tesla	92.2%	Tesla	88.9%
Hotel Chains	Marriott	64.4%	Marriott	77.2%	Marriott	42.5
Laptops	Lenovo	69.9%	Dell	85.3%	Dell	82.1%
Online Dating Platforms	Bumble, OKCupid	94.7%	Bumble	87.9%	Hinge	44.4
Running Shoes	Brooks	69.8%	Nike	85.9%	Nike	72.5%
Smartphones	Samsung	91.0%	Samsung	97.1%	Samsung	100.0%
Social Media Platforms	Instagram	65.0%	Instagram	54.6	TikTok	41.5
Telecommunication Services	Verizon	88.0%	T-Mobile	93.2%	Verizon	85.5%
Investment Commodity	Agricultural Products	72.9%	Gold	100%	Gold	70.0%
Sports	Basketball	27.2	Swimming	34.7	Football	23.2
Travel Destinations	US	32.6	US	49.2	US	32.3
Universities	Stanford	84.6%	Stanford	64.4%	Stanford	48.3
Vegetables	Carrot	46.6	Carrot	49.3	Seeds	11.6
Weekend Getaway Cities	Asheville	30.6	Portland	28.0	Paris	51.2
Wine regions (Country)	France	90.4%	US	86.1%	France	100.0%

Table 2: Comparative Top Preferences of GPT, Gemini and DeepSeek Across Topics. Green shading indicates the strength of each model's preferences, with darker green representing stronger, more rigid preferences where the model recommends the same top-ranked entity in a higher percentage of responses, regardless of user cluster or query formulation. Shading is applied only to response shares $\geq 60\%$.

for Countries to Live In, the results were 0.0901 ($p = 0.99$) for GPT, 0.429 ($p = 0.98$) for Gemini and 0.421 ($p = 0.98$) for DeepSeek. These exceptionally high p-values indicate no statistically significant difference in preference distributions across query runs, confirming that both the preference themselves and their magnitude remains consistent.

5.2 AI Assistants Show a Geographic Bias Towards US Entities

Entities were classified into three principal regions: United States, Europe, and Asia, based on the location of their primary corporate headquarters. This classification reflects the distribution of entities within the dataset, which was heavily concentrated in these three regions. Entities from other regions appeared too infrequently to warrant separate analysis: four Canadian, three Australasian, and none from Africa, South America, or any other regions. Moreover, no individual European or Asian country appeared more than a few times, precluding the possibility of conducting statistically meaningful country-level comparisons. Restricting the analysis to these three regional groupings ensured adequate sample sizes for robust statistical inference while preserving the representativeness of the dataset.

This analysis then compared the recommendation frequency of US vs Europe, US vs Asia and US vs all Non-US regions. This focus was chosen because GPT and Gemini showed clear US overrepresentation: apart from Airlines and Smartphones, US-based entities made up more than half of their top five recommendations in every topic. While DeepSeek's patterns were less pronounced, once the entities were sorted into the three regions, the US was always first or second by recommendation count, never last. This meant that comparing these three region-pairs still provided the most informative approach for assessing potential geographic bias.

This study found that **American AI assistants (GPT and Gemini) showed a pronounced and statistically significant (**$p < 0.05$**) bias across nine of twelve topics when comparing US and Asia entities, and eleven of twelve topics when comparing US and European entities (Appendix Tables 10, 11).**

Gemini's **US vs Asia** results (Random-Effects LOR = 1.94, $p = 8.64 \times 10^{-4}$, OR $\approx$ 6.96) show that **US entities were nearly seven times more likely to appear in the top recommendations**. The two cultural topics, Weekend Getaways (LOR = 3.86, $p < 0.0001$, OR = 47.5) and Universities (LOR = 3.74), were amongst the largest effects. **Even in industries with strong Asian representation, such as Electric Vehicles (**LOR = 1.47**), where brands like Toyota, Hyundai, and BYD are major global players, and Hotel Chains (**LOR = 2.54**), home to internationally recognised groups like Shangri-La and Mandarin Oriental, a clear US bias persisted.** The bias was absent only in Airlines, Laptops, and Smartphones, where p-values indicated no statistically significant difference. GPT's results yielded a Random-Effects LOR = 2.23 ($p \approx 2.46 \times 10^{-3}$, OR $\approx$ 9.30), so similarly indicating a strong US advantage. Its results diverged only in Smartphones, where a small but significant negative LOR = −0.362 suggested a tilt toward Asian entities.

Both American AI assistants' **US vs Europe** comparison revealed an even stronger pattern, with Gemini's results yielding a Random-Effects LOR = 2.44 ($p \approx 9.44 \times 10^{-5}$, OR $\approx$ 11.5) and GPT of LOR = 2.34 ($p = 2.16 \times 10^{-}$, OR $\approx$ 10.4). In both cases this means **US entities were roughly 11 times more likely to appear in top recommendations**. Laptops (Gemini LOR = 4.38, GPT LOR = 4.96), Online Dating Platforms (Gemini LOR = 3.02, GPT LOR = 5.22) and Social Media Platforms (Gemini LOR = 3.78, GPT LOR = 4.72) showed the sharpest disparities; differences that are likely, in part, explicable by the market dominance of US-based firms in these domains. However, **even categories where Europe has globally competitive entities, such as Running Shoes (Adidas, Puma, Salomon) and Weekend Getaway Cities, showed statistically significant US bias**. Airlines was the only category that showed no significant difference, and no category demonstrated significant bias in Europe's favour.

Most notably **this observed pattern of US over-representation in the American AI Assistants persisted even when all non-US entities were treated collectively in the US vs non-US comparison (Table 3)**. The meta-analysis results for Gemini was LOR = 1.16 ($p = 2.87 \times 10^{-3}$, OR ≈ 3.19) while similarly for GPT it was LOR = 0.953 ($p = 0.0112$, OR ≈ 2.59), confirming that the bias extended beyond specific regional pairings.

In contrast, DeepSeek showed a markedly smaller US tilt. In the US vs Asia comparison, it had a Random-Effects LOR = 1.49 ($p \approx 1.06 \times 10^{-3}$, OR ≈ 4.44), while for US vs Europe it had LOR = 1.39 ($p \approx 1.71 \times 10^{-3}$, OR ≈ 5.53); both positive but notably below the effect sizes observed for Gemini and GPT across the same pairings. Crucially, when pooling all non-US entities in the US vs non-US comparison, DeepSeek also exhibited no statistically significant overall bias (Random-Effects LOR = 0.411, $p = 0.103$), indicating that its overall recommendations were not systematically skewed toward US entities when considered against the rest of the world. DeepSeek diverged most markedly from the American models in the Weekend Getaway Cities topic, where it showed a significant tilt towards European cities (US vs Europe LOR = −1.00). Conversely it's strongest bias emerged in the Universities topic (LOR = 2.11), where it demonstrated a pronounced preference for US institutions, particularly when compared to Asia universities (US vs Asia LOR = 4.96), mirroring the pattern observed in both Gemini and GPT.

6 Discussion

Our findings show that AI assistants do not act as neutral information providers but instead display structured and persistent preferences. When entities were grouped by geographic region, log-odds ratios between region-pairs showed consistent and statistically significant asymmetries. These asymmetries were most pronounced in the American AI assistants (Gemini and GPT), which displayed strong biases towards the United States, while DeepSeek exhibited notably smaller geographic preferences. These findings question the underlying factors driving these recommendations and their implications for global users relying on these systems for objective information.

6.1 Understanding the Origin of AI Assistants' Preferences

AI assistants exhibit strong and stable preferences when making recommendations for institutions, brands and other cultural entities. These biases likely arise from three interconnected factors:

Training Data Composition: AI models develop preferences based on entity frequency and authority within their training datasets. Models trained on extensive text corpora containing inherent coverage imbalances [4] prioritize well-documented brands, dominant cultural narratives, and established institutions over emerging competitors and alternative perspectives.

Semantic Embedding Structures: AI models develop internal representations that favour certain brands, services, and institutional entities through their semantic embedding processes [9, 23] During training, entities that co-occur with positive descriptors or authoritative contexts become more strongly weighted in the model's latent space, making them more likely to be retrieved and recommended regardless of query specificity.

User Feedback Amplification: Real-world deployments create self-reinforcing cycles where user engagement patterns strengthen preferences for particular entities [38]. When users interact more positively with certain recommendations, through clicks, extended conversations, or explicit endorsements, models internalise these signals and increasingly prioritise these options. This creates echo chambers where already-prominent governmental, commercial and cultural entities receive disproportionate visibility.

These mechanisms mirror the core insight from Thaler et al. [45]: the design of the decision environment heavily influences outcomes. AI assistants now act as powerful choice architects, actively structuring the decision landscape. Their recommendations create a world where visibility is shaped less by objective utility and more by informational defaults, semantic frames, and feedback amplification. As synthetic data and self-learning methodologies become more prevalent, these preference structures will likely persist, further entrenching AI systems' role as gatekeepers favouring certain institutions, brands, and cultures over alternatives.

6.2 AI Assistants Systematically Favour US Entities

The over-representation of US-based entities in AI assistant recommendations reveals a profound geographic bias, though this pattern varies significantly in magnitude between American-developed and Chinese-developed models.

Training Data Geographic Concentration: The observed US bias likely primarily stems from the predominance of English language, American-centric content in AI training datasets. Major web crawls, news aggregators, and digital repositories disproportionately capture American commercial discourse, product reviews, and institutional promotion [39]. This leads the models to conflate digital visibility with market relevance, favouring entities with stronger American web presence regardless of global market position. The effect is particularly pronounced in categories like Weekend Getaways and Universities, where American destinations and institutions benefit from extensive English-language promotion and discussion online. While in a number of the topics DeepSeek also exhibits preference towards US entities, its markedly smaller magnitude suggests that training data composition and curation might substantially mitigate these geographic preferences.

The consistent pattern across both Gemini and GPT models, with odds ratios approaching 8:1 against Asian entities and 14:1 against European entities, demonstrates the severity of US bias in American AI development approaches. While DeepSeek also shows statistically significant US preferences, its substantially lower odds ratios indicates that the magnitude of geographic bias can be significantly reduced through alternative development methodologies. Even in sectors where international competition is fierce, such as Electric Vehicles, where Asian manufacturers such as BYD, Geely and Nio

Topic (US vs non-US)	Gemini		GPT		DeepSeek	
	Average LOR	p	Average LOR	p	Average LOR	p
Airlines	-0.806	0.330	-1.37	0.0530	-0.940	0.0394
Electric Cars	0.236	0.278	0.154	0.449	-0.0443	0.237
Hotel Chains	0.422	0.0333	-0.0244	0.896	0.313	0.237
Laptops	0.0848	0.626	0.749	1.59e-3	0.495	0.0250
Online Dating Platforms	3.02	6.45e-7	5.22	1.00e-12	1.47	3.05e-4
Running Shoes	0.801	8.69e-5	0.743	9.81e-5	0.685	7.23e-4
Smartphones	-0.210	0.032	-0.467	2.30e-4	-0.305	4.48e-3
Social Media Platforms	1.56	1.26e-8	2.18	1.76e-6	1.27	2.43e-6
Telecommunication Services	2.97	6.01e-5	1.25	1.83e-3	1.15	0.0107
Universities	2.46	2.71e-3	1.97	3.56e-7	2.11	2.84e-6
Weekend Getaway Cities	3.86	1.00e-12	1.21	4.63e-3	-1.43	2.02e-4

Table 3: Comparison of US vs all Non-US Regions Preference Strengths: Average Log Odds Ratios (LOR) and Significance (*p*-values) for Gemini, GPT and DeepSeek Across Topics. Green shading indicates positive LOR (US bias) while red shading indicates negative LOR (non-US bias), with darker shading representing stronger geographic biases.

lead in innovation and market share, or Universities, where European institutions rank among the world's best, all models maintain some degree of US bias, though this is most pronounced in the American-developed systems.

This geographic bias has profound implications for global market dynamics. The pronounced US preferences in Gemini and GPT models risk creating artificial competitive advantages for US-based entities while marginalising international alternatives. As AI assistants increasingly influence consumer decision-making, these preferences risk distorting global commerce by amplifying American market presence beyond its objective merit. The over-representation of US cities in AI-generated travel recommendations may influence international travel patterns, disproportionately diverting tourism revenue toward US destinations at the expense of global competitors. Moreover, AI-driven recommendations shape perception of governance models, policies, cultures, healthcare systems, and education; the over-representation of US healthcare programs in AI recommendations, even when European and Asian healthcare systems rank higher in global indexes, could influence public perceptions of successful policy. While DeepSeek demonstrates that this bias magnitude can be reduced, the persistence of some US tilt across all models highlights the pervasive influence of American digital dominance in AI training ecosystems.

6.3 Limitations and Future Work

While this study provides a structured analysis of AI assistant preferences across 20 topics and 4,140 questions, further investigation can be useful to develop a more comprehensive understanding of AI-driven biases.

- **Regional, Linguistic, and Personalisation Factors:** This study controlled for geographic influence by using Swedish IPs and standardised English queries. Future research may explore how AI recommendations vary across different IP origins, languages and personalisation settings to determine whether localized or user-tailored AI systems develop different ranking preferences.
- **Longitudinal Studies on AI Evolution:** AI models undergo continuous updates integrating new training data and reinforcement learning mechanisms. Future research may track how AI-generated rankings shift over time, assessing whether biases persist, worsen, or improve with each iteration and whether recommendations adjust based on user feedback, regulations, or corporate interests.
- **Exploring Behavioural Patterns in AI Assistants:** A notable pattern in our study was the variation in recommendation behaviour between models. GPT provided direct recommendations (97.5%) of the time, whereas Gemini was more caution at 73%. This raises questions about whether AI models exhibit personality-like traits in decision-making. Future research could investigate whether these behavioural patterns resemble human psychological traits, how they affect user trust, and whether reinforcement learning contributes to distinct AI "personalities" over time.

7 Conclusion

This research underscores the urgency to treat AI assistants not merely as convenient consumer tools but as influential choice architects that shape consumer perceptions, business visibility, and public opinion. As reliance on these systems deepens, the responsibility to ensure that they enable rather than restrict fair and diverse participation in economic, cultural, and political life becomes increasingly critical.

Through ChoiceEval, a novel evaluation framework and benchmark, we investigated various types of entity-perception biases in LLM based AI assistants. The persistence of these biases, across models, topics, questions, contexts and user personas, demonstrates the need for fundamental shifts in how AI-mediated visibility is approached. Businesses, governments, AI developers and policymakers face new challenges in ensuring visibility, fairness, truthfulness, cultural representation, diversity and trust in these new information mediators. Addressing embedded biases require coordinated action in data governance, algorithmic transparency, and stakeholder collaboration to support equitable and socially good AI-mediated interactions.

References

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2019. Managing Popularity Bias in Recommender Systems with Personalized Re-ranking. arXiv:1901.07555v4 https://arxiv.org/abs/1901.07555v4

[2] Alan Agresti. 2013. Describing Contingency Tables. In *Categorical Data Analysis* (3 ed.). Wiley, Chapter 2.

[3] Sarah Alnegheimish, Alicia Guo, and Yi Sun. 2022. Using Natural Sentence Prompts for Understanding Biases in Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 2824–2830.

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, 610–623. doi:10.1145/3442188.3445922

[5] Kirti Bhagat, Kinshuk Vasisht, and Danish Pruthi. 2025. Richer Output for Richer Countries: Uncovering Geographical Disparities in Generated Stories and Travel Recommendations. *arXiv preprint arXiv:2411.07320* (2025). https://arxiv.org/abs/2411.07320

[6] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates, Inc., Red Hook, NY, USA, 4356–4364.

[7] Michael Borenstein, Larry Hedges, Julian Higgins, and Hannah Rothstein. 2009. *Random-Effects Model*. John Wiley & Sons, Ltd, Chapter 12, 69–75. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470743386.ch12 doi:10.1002/9780470743386.ch12

[8] M. Borenstein, L. V. Hedges, J. P. Higgins, and H. R. Rothstein. 2010. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods* 1, 2 (2010), 97–111. doi:10.1002/jrsm.12

[9] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356 (2017), 183–186. doi:10.1126/science.aal4230

[10] Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*. Association for Computational Linguistics, Dubrovnik, Croatia, 53–67.

[11] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 862–872.

[12] Margot Geerts, Manon Reusens, Bart Baesens, Seppe vanden Broucke, and Jochen De Weerdt. 2025. On the Performance of LLMs for Real Estate Appraisal. *arXiv preprint arXiv:2506.11812* (2025). https://arxiv.org/abs/2506.11812

[13] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 3356–3369.

[14] M. Glick, G. Richards, M. Sapozhnikov, et al. 2014. How Does Ranking Affect User Choice in Online Search? *Review of Industrial Organization* 45 (2014), 99–119. doi:10.1007/s11151-014-9435-y

[15] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 3309–3326.

[16] W. A. Henry. 1976. Cultural values do correlate with consumer behavior. *Journal of Marketing Research* 13, 2 (1976), 121–127. doi:10.2307/3150845

[17] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery. doi:10.1145/1076034.1076063

[18] Mahammed Kamruzzaman, Hieu Minh Nguyen, and Gene Louis Kim. 2024. "Global is Good, Local is Bad?": Understanding Brand Bias in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Miami, Florida, USA, 12695–12702.

[19] William Kruskal and Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *J. Amer. Statist. Assoc.* 47, 260 (1952), 583–621.

[20] Sydney Lake. 2025. OpenAI CEO says Gen Z and millennials use ChatGPT like a "life adviser". https://fortune.com/2025/05/13/openai-ceo-sam-altman-says-gen-z-millennials-use-chatgpt-like-life-adviser/

[21] Harsh Nishant Lalai, Raj Sanjay Shah, Jiaxin Pei, Sashank Varma, Yi-Chia Wang, and Ali Emami. 2025. The World According to LLMs: How Geographic Origin Influences LLMs' Entity Deduction Capabilities. In *Proceedings of the 2nd Conference on Language Modeling (COLM 2025)*. https://arxiv.org/abs/2508.05525 arXiv preprint arXiv:2508.05525.

[22] Kristina Lerman and Tad Hogg. 2014. Leveraging Position Bias to Improve Peer Recommendation. *PLOS ONE* 9 (2014), e98914. doi:10.1371/journal.pone.0098914

[23] Sahil Manchanda and Pannaga Shivaswamy. 2025. What is in a name? Mitigating Name Bias in Text Embedding Similarity via Anonymization. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, Vienna, Austria, 17759–17781.

[24] Rohin Manvi, Samar Khanna, Marshall Burke, David B. Lobell, and Stefano Ermon. 2024. Large Language Models are Geographically Biased. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, 34654–34669. https://proceedings.mlr.press/v235/manvi24a.html

[25] Mikael Markander. 2025. More internet users in the US are ditching Google search for ChatGPT. https://www.computerworld.com/article/4028348/increasingly-common-to-use-chat-gpt-instead-of-google.html

[26] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 2243–2251. doi:10.1145/3269206.3272027

[27] A. Mitchell. 1983. *The Nine American Lifestyles: Who We Are and Why We Live the Way We Do*. Macmillan, New York.

[28] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 5356–5371.

[29] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 1953–1967.

[30] Ariel Noyman, Kai Hu, and Kent Larson. 2025. TravelAgent: Generative agents in the built environment. *Environment and Planning B: Urban Analytics and City Science* (July 2025). doi:10.1177/23998083251360458

[31] J. C. Nunnally and I. H. Bernstein. 1994. The Assessment of Reliability. In *Psychometric Theory* (3 ed.). McGraw-Hill, 248–292.

[32] OpenAI. 2025. Introducing the Intelligence Age. https://openai.com/global-affairs/introducing-the-intelligence-age/

[33] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2799–2804.

[34] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*. 2086–2105.

[35] Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of Cultural Awareness in Language Models: Text and Beyond. *arXiv preprint arXiv:2411.00860* (2024). https://arxiv.org/abs/2411.00860

[36] PCWorld. 2024. Believe it or not, ChatGPT gets over 1 billion messages every single day. https://www.pcworld.com/article/2546712/believe-it-or-not-chatgpt-gets-over-1-billion-messages-every-single-day.html

[37] J. T. Plummer. 1974. The concept and application of life style segmentation. *Journal of Marketing* 38, 1 (1974), 33–37. doi:10.2307/1250164

[38] Filip Radlinski and Thorsten Joachims. 2007. Active exploration for learning rankings from clickthrough data. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07)*. Association for Computing Machinery, 570–579. doi:10.1145/1281192.1281254

[39] Richard Rogers. 2024. Doing Digital Methods. In *Proceedings of SAGE Publications*. SAGE Publications Ltd, London, United Kingdom, 1–100. http://digital.casalini.it/9781529784176 Casalini ID: 5730579.

[40] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3407–3412.

[41] Michael R. Solomon. 2020. *Consumer Behavior: Buying, Having, and Being*. Pearson.

[42] Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology* 15, 1 (1904), 72–101.

[43] Lukas Stoppacher, Thomas Foscht, Andreas B. Eisingerich, and Judith Schloffer. 2024. "Always on Your Mind? – Investigating Consideration Sets and Private Labels at the Retailer and Category Level". *The International Review of Retail, Distribution and Consumer Research* 34, 5 (2024), 646–668. doi:10.1080/09593969.2024.2345122

[44] Yan Tao, Olga Viberg, Ryan Baker, and René Kizilcec. 2024. Cultural Bias and Cultural Alignment of Large Language Models. *PNAS Nexus* 3 (2024). doi:10.1093/pnasnexus/pgae346

[45] Richard H. Thaler, Cass R. Sunstein, and John P. Balz. 2010. Choice Architecture. SSRN Electronic Journal. doi:10.2139/ssrn.1583509 Available at SSRN: https://ssrn.com/abstract=1583509 or http://dx.doi.org/10.2139/ssrn.1583509.

[46] The Times of India. 2024. Gemini vs ChatGPT: Google CEO Sundar Pichai shares these numbers, says 'for all these AI features, it's just...'. https://timesofindia.indiatimes.com/technology/tech-news/gemini-vs-chatgpt-google-ceo-sundar-pichai-shares-these-numbers-says-for-all-these-ai-features-its-just/articleshow/114808132.cms

[47] Yucheng Zhang, Xinyi Ye, Ruoxi Ning, Qingxiu Dong, Yutong Li, Xinyu Guan, Yuqing Xie, Xu Han, Zhiyuan Liu, Yankai Lin, and Maosong Sun. 2025. Revealing LLM Sycophancy via Geographically Situated Beliefs. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, Bangkok, Thailand, 10456–10472. https://aclanthology.org/2025.findings-acl.914

[48] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 15–20.

[49] Yuhang Zhou, Yuchen Ni, Zhiheng Xi, Zhangyue Yin, Yu He, Yunhui Gan, Xiang Liu, Jian Zhang, Sen Liu, Xipeng Qiu, Yixin Cao, Guangnan Ye, and Hongfeng Chai. 2025. Are LLMs Rational Investors? A Study on the Financial Bias in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, 24139–24173. https://github.com/zhiqix/FinCausal

A Further Correlation Analysis

	Df1	Df2	Df3	Df4	Df5
Df1	1.00 (0.00)				
Df2	0.988 (9.31e-8)	1.00 (0.00)			
Df3	0.964 (7.32e-6)	0.976 (1.47e-6)	1.00 (0.00)		
Df4	0.952 (2.28e-5)	0.964 (7.32e-6)	0.976 (1.47e-6)	1.00 (0.00)	
Df5	0.976 (1.47e-6)	0.988 (9.31e-8)	0.988 (9.31e-8)	0.952 (2.28e-5)	1.00 (0.00)

Table 4: Gemini's Spearman's Rank Correlation Matrix: Laptops

	Df1	Df2	Df3	Df4	Df5
Df1	1.00 (0.00)				
Df2	0.864 (1.27e-3)	1.00 (0.00)			
Df3	0.944 (3.97e-5)	0.834 (2.73e-3)	1.00 (0.00)		
Df4	0.879 (8.14e-4)	0.908 (2.82e-4)	0.879 (7.97e-4)	1.00 (0.00)	
Df5	0.966 (5.77e-6)	0.873 (9.78e-4)	0.962 (9.10e-6)	0.869 (1.11e-3)	1.00 (0.00)

Table 5: Gemini's Spearman's Rank Correlation Matrix: Countries

	Df1	Df2	Df3	Df4	Df5
Df1	1.00 (0.00)				
Df2	1.00 (0.00)	1.00 (0.00)			
Df3	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)		
Df4	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	
Df5	0.988 (9.31e-8)	0.988 (9.31e-8)	0.988 (9.31e-8)	0.988 (9.31e-8)	1.00 (0.00)

Table 6: GPT's Spearman's Rank Correlation Matrix: Laptops

	Df1	Df2	Df3	Df4	Df5
Df1	1.00 (0.00)				
Df2	0.976 (1.47e-6)	1.00 (0.00)			
Df3	0.927 (1.12e-4)	0.903 (3.44e-4)	1.00 (0.00)		
Df4	0.951 (2.45e-5)	0.916 (1.94e-4)	0.882 (7.36e-4)	1.00 (0.00)	
Df5	0.974 (2.08e-6)	0.962 (8.63e-6)	0.905 (3.20e-4)	0.957 (1.46e-5)	1.00 (0.00)

Table 7: GPT's Spearman's Rank Correlation Matrix: Countries

	Df1	Df2	Df3	Df4	Df5
Df1	1.00 (0.00)				
Df2	0.988 (9.31e-8)	1.00 (0.00)			
Df3	0.988 (9.31e-8)	1.00 (0.00)	1.00 (0.00)		
Df4	0.988 (9.31e-8)	0.976 (1.47e-6)	0.976 (1.47e-4)	1.00 (0.00)	
Df5	0.976 (1.47e-6)	0.988 (9.31e-8)	0.988 (9.31e-8)	0.964 (7.32e-6)	1.00 (0.00)

Table 8: DeepSeek's Spearman's Rank Correlation Matrix: Laptops

	Df1	Df2	Df3	Df4	Df5
Df1	1.00 (0.00)				
Df2	0.939 (5.48e-5)	1.00 (0.00)			
Df3	0.905 (3.20e-4)	0.905 (3.20e-4)	1.00 (0.00)		
Df4	0.903 (3.44e-4)	0.830 (2.94e-3)	0.744 (1.35e-2)	1.00 (0.00)	
Df5	0.896 (4.46e-4)	0.792 (6.31e-3)	0.750 (1.24e-2)	0.896 (4.46e-4)	1.00 (0.00)

Table 9: DeepSeek's Spearman's Rank Correlation Matrix: Countries

B Regional Comparisons

Topic (US vs Europe)	Gemini		GPT		DeepSeek	
	Average LOR	p	Average LOR	p	Average LOR	p
Airlines	0.459	0.484	0.636	0.241	0.389	0.0830
Electric Cars	0.89	0.0223	1.12	2.54e-3	0.690	0.0156
Hotel Chains	1.63	2.79e-5	1.17	3.56e-5	1.20	6.90e-5
Laptops	4.38	1.69e-10	4.96	1.00e-12	4.02	8.36e-7
Online Dating Platforms	3.02	6.45e-7	5.22	1.00e-12	2.29	5.58e-6
Running Shoes	1.56	1.08e-3	1.48	3.85e-3	1.11	1.23e-3
Smartphones	2.46	1.61e-3	2.89	4.10e-4	1.41	2.50e-4
Social Media Platforms	3.78	7.43e-6	4.72	1.00e-12	2.47	1.09e-5
Telecommunication Services	3.05	3.37e-5	1.33	1.50e-3	1.57	2.33e-3
Universities	2.46	2.71e-3	2.15	2.28e-6	2.58	5.86e-6
Weekend Getaway Cities	3.86	0.00	1.71	3.95e-3	-1.00	3.71e-3

Table 10: Comparison of US vs European Preference Strengths: Average Log Odds Ratios (LOR) and Significance (*p*-values) for Gemini, GPT, and DeepSeek across topics. Green shading indicates positive LOR (US bias) while red shading indicates negative LOR (European bias), with darker shading representing stronger geographic biases.

Topic (US vs Asia)	Gemini		GPT		DeepSeek	
	Average LOR	p	Average LOR	p	Average LOR	p
Airlines	-0.0649	0.951	-0.546	0.512	-0.0578	0.931
Electric Cars	1.47	1.38e-3	0.881	0.0182	0.753	3.73e-4
Hotel Chains	2.54	5.02e-4	2.41	2.66e-3	1.86	9.25e-3
Laptops	0.0848	0.626	0.749	1.59e-3	0.537	0.0234
Online Dating Platforms	3.74	6.77e-6	5.22	1.00e-12	3.75	3.32e-6
Running Shoes	1.91	5.37e-4	1.98	5.74e-4	2.21	1.43e-5
Smartphones	-0.0293	0.768	-0.362	8.06e-3	-0.0338	0.791
Social Media Platforms	1.72	1.87e-8	2.18	1.76e-6	1.92	2.98e-8
Telecommunication Services	3.57	2.95e-7	4.94	1.00e-12	3.12	1.44e-4
Universities	3.74	5.30e-4	4.96	1.00e-12	4.94	6.68e-8
Weekend Getaway Cities	3.86	1.00e-12	2.73	4.33e-5	0.308	0.229

Table 11: Comparison of US vs Asian Preference Strengths: Average Log Odds Ratios (LOR) and Significance (*p*-values) for Gemini, GPT and DeepSeek Across Topics. Green shading indicates positive LOR (US bias) while red shading indicates negative LOR (Asian bias), with darker shading representing stronger geographic biases.