Unequal Opportunities: Examining the Bias in Geographical Recommendations by Large Language Models

Shiran Dudy
EAI
Northeastern University
Boston, MA, USA
s.dudy@northeastern.edu

Thulasi Tholeti
EAI
Northeastern University
Boston, MA, USA
t.tholeti@northeastern.edu

Resmi Ramachandranpillai
Northeastern University
Boston, MA, USA
r.ramachandranpillai@northeastern.edu

Muhammad Ali*
EAI
Northeastern University
Boston, MA, USA
ali.muh@northeastern.edu

Toby Jia-Jun Li
University of Notre Dame
Notre Dame, IN, USA
toby.j.li@nd.edu

Ricardo Baeza-Yates
EAI
Northeastern University
Boston, MA, USA
rbaeza@acm.org

Abstract

Recent advancements in Large Language Models (LLMs) have made them a popular information-seeking tool among end users. However, the statistical training methods for LLMs have raised concerns about their representation of under-represented topics, potentially leading to biases that could influence real-world decisions and opportunities. These biases could have significant economic, social, and cultural impacts as LLMs become more prevalent, whether through direct interactions—such as when users engage with chatbots or automated assistants—or through their integration into third-party applications (as agents), where the models influence decision-making processes and functionalities behind the scenes. Our study examines the biases present in LLMs recommendations of U.S. cities and towns across three domains: relocation, tourism, and starting a business. We explore two key research questions: (i) How similar LLM responses are, and (ii) How this similarity might favor areas with certain characteristics over others, introducing biases. We focus on the consistency of LLM responses and their tendency to over-represent or under-represent specific locations. Our findings point to consistent demographic biases in these recommendations, which could perpetuate a "rich-get-richer" effect that widens existing economic disparities.

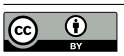
CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Social and professional topics** → *Economic impact*.

Keywords

Cultural representation, LLM biases, under-represented topics, geographical divide, LLM auditing.

*Work done while author was at Northeastern University.



IUI '25, Cagliari, Italy

ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712111

ACM Reference Format:
Shiran Dudy, Thulasi Tholeti, Resmi Ramachandranpillai, Muhammad Ali, Toby Jia-Jun Li, and Ricardo Baeza-Yates. 2025. Unequal Opportunities: Examining the Bias in Geographical Recommendations by Large Language Models. In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3708359.3712111

1 Introduction

Geographic and socio-economic factors significantly influence the assessment, representation, and dissemination of knowledge. Historically, knowledge systems such as Wikipedia and OpenStreetMap (OSM) were the main sources that millions of people relied on for accurate information, either directly or through applications that use these information sources. These sources were created by community members with the interest and resources to help educating others and shaping an online presence. However, the ability to contribute is not uniformly distributed across communities. Research by Johnson *et al.* [31] highlight the fact that there is a clear urban-rural divide in these knowledge platforms, where urban areas are more thoroughly documented than rural ones. Similarly, Lorini *et al.* [43] demonstrated that disasters in wealthier countries tend to receive more coverage compared to those in less affluent nations. This reflects broader structural inequalities and leads to the underrepresentation of less affluent regions. This underrepresentation perpetuates biases within these data sources.

With the advent of Large Language Models (LLMs), the landscape of information consumption has transformed. With over 100 million active users each month[1], some of these LLM-enabled tools are increasingly replacing traditional methods for information-seeking [63], among other purposes. However, research indicates that LLMs may amplify existing data biases, potentially exacerbating inequality through algorithmic bias [5, 23, 33].

Building on the recognition of existing biases, this study aims to examine the geographic and socio-economic *representation* within LLMs, focusing on the consistency of responses across different geographic areas and identifying which locations are over or underrepresented. This study assesses the inclusivity of LLMs as a widely utilized socio-technical tool. We investigate how equitably

[1]https://www.statista.com/statistics/1368657/chatgpt-mau-growth/

LLMs serve information representing communities of diverse demographic backgrounds, examining who gains advantages and who may be left out. Our findings reveal disparities in the representation of certain demographics, which would have real-world socioeconomic implications for the culture, economy, and politics of cities and towns of the affected regions, as well as for historically underrepresented groups due to the increasingly wide use of LLMs in information search and decision-making assistance, either directly (e.g., the Gemini panel in Google and the Copilot panel in Bing) or indirectly through other applications that are powered by LLMs.

Our research analyzes six state-of-the-art LLMs, examining their responses to twelve queries across three domains: relocation, business establishment, and tourism, within a U.S. context. We conduct a comprehensive analysis to pinpoint diversity in LLMs and representational gaps through two research questions: (RQ1) are LLMs similar in their responses and (RQ2) what kind of locations are recommended.

The main contributions of this paper, to the best of our knowledge, are the following:

(1) An analysis of LLM-recommended locations, their diversity and justifications, focusing on free-form text, unlike previous studies that mainly used ratings or numerical data.
(2) A comparison of the LLM responses with the U.S. city database sourced from the U.S. Census Bureau, to evaluate the representational gaps,
(3) An analysis showing that groups historically or socioeconomically underserved tend to be underrepresented in the recommended cities, highlighting how such LLM biases can exacerbate existing disparities among communities.

The rest of the paper is organized as follows. Section 2 presents related work while Section 3 detail our methods. In Sections 4 and 5 we show the results for our two research questions. In Sections 6 and Sections 7 we discuss our results and the limitations of our work. We end with the conclusions in Section 8.

2 Related Work

In this section, we review the existing literature on the role of LLMs as information-seeking tools, study biases in Generative AI systems, especially focusing on uneven geographic and socioeconomic representations.

2.1 LLMs as Information-Seeking Tools

LLMs have made tremendous progress in Natural Language Processing, transforming how we search, retrieve, and interact with information. Traditional information-seeking processes involved keyword-based retrievals such as search engines, which, while efficient, often fall short in understanding long queries and context. LLMs emerged as powerful systems, capable of handling long and diverse questions and offering more conversational responses.

Several researches focus on comparing LLM-powered conversational systems and traditional web search engines. One domain where LLMs have been employed for information seeking is healthcare [22, 39, 69]. A comprehensive analysis has been made in [22], within this domain comparing LLMs, traditional search engines, and retrieval-augmented generation (RAG) approaches, highlighting the strengths and weaknesses of each method. The authors found that, while LLMs provide more accurate responses to health-related questions, they are also highly sensitive to input prompts. Academic research is another area where LLMs have been shown to progress as information-seeking tools. The authors in [68] discuss how LLMs can reduce the effort involved in academic information retrieval, particularly when accessing APIs. In addition, [37] compares LLMs (such as ChatGPT) with traditional web search engines for writing SQL queries. The study reveals that while LLMs offer benefits, such as potentially higher-quality query outputs and reduced mental demand for students, the process of interaction with these models can be more demanding compared to web search.

Visual Question-Answering (VQA) is a more complex information-seeking task and the studies show the potential of LLMs coupled with external tools in enhancing performance capabilities. The AVIS framework [29], is one such tool that leverages LLMs for autonomous information-seeking in VQA. This framework combines LLMs with tree search and external APIs to answer complex questions that require external knowledge beyond the visual content. The authors used user studies to collect data on decision-making processes and employ this information to create a system that mimics human behavior in tool usage and reasoning. AVIS achieves state-of-the-art performance on knowledge-based VQA benchmarks, underlining the potential of LLMs to extend their functionality beyond text-based queries into more multimodal tasks.

2.2 Biases in LLMs Responses

A significant body of literature studied how LLMs are biased in relation to race, gender, age, and other demographic factors [2, 6, 25, 35, 58]. A study [12] found that many LLMs enforce racial and gender stereotypes, leading to skewed or harmful outputs. These biases are often sourced from the data used to train, which may over-represent/under-represent certain groups. Similarly, in [54], the authors analyze how LLM responses vary in accuracy, factuality, and refusal rates based on three user factors: English proficiency, education level, and country of origin. The authors found that users with lower English proficiency, lower education levels, and those from countries outside the U.S. experience more undesirable behaviors compared to their counterparts.

In [7], the authors found that LLMs trained on diverse datasets may still exhibit and amplify racial stereotypes toward specific ethnic groups. Additionally, a study on age biases in LLM responses has been done in [41] and found that LLMs generate language that disproportionately favors/disadvantages based on an individual's age and age-related stereotypes, neglecting the unique needs and priorities of certain age groups.

Domains such as healthcare and finance, are more prone to domain-specific biases that can influence the performance of LLMs. In healthcare, LLMs inherit biases from the medical literature or datasets, significantly affecting diagnoses or treatment recommendations as studied in [55]. In finance, LLMs may reflect biases present in credit scoring or loan approval processes, leading to discriminatory credit decisions [72].

Another research direction is about cultural representation in LLMs [1, 3, 34, 49]. In [20] the authors examine how LLMs represent cultural aspects of emotions in mixed-emotion situations raising concerns about potential biases towards Anglo-centric values due

to predominantly Western training data. The authors found that LLMs showed limited alignment with established cultural literature and that the chosen language had greater influence on responses than its textual content.

Our work is most closely related to Salinas et al. [60], but it differs in its focus, addressing city recommendations rather than hiring. Additionally, our study leverages ground truth data from U.S. city datasets and conducts a comprehensive analysis of demographic attributes.

2.3 Geographic and Socioeconomic Representations in GenAI Systems

Recently, various studies have investigated geographic bias in LLMs. A study that evaluates the geographic disparities in 16 mainstream LLMs highlights the potential negative consequences of biased representations of different regions [19]. Another benchmark, WorldBench [48], addresses geographic disparities using World Bank data to compare LLM performance across countries, finding significantly higher error rates for African nations compared to North American ones. Moreover, research into brand bias [32] has demonstrated that LLMs tend to favor global and luxury brands over local ones, which could exacerbate economic inequalities. Additionally, demographic bias in LLMs has been investigated through job recommendations, finding patterns where models suggest lower-paying jobs to Mexican workers or secretarial roles to women, highlighting intersectional biases related to gender and nationality [61]. Luo et al. [44] explored how attitudes toward immigrant cuisines in Yelp reviews reflect broader social prejudices illustrating the impact of biases. Additionally, the study finds that reviews generated by LLMs reproduce harmful framing tendencies, indicating that biases in these platforms can lead to the retention and reinforcement of harms in AI-generated content.

2.4 Evaluation of Consistency in Responses

Some studies [11, 59] have shown that the generated responses are irrelevant or inconsistent with the provided context, and LLMs often *hallucinate*. A common method for evaluating the consistency of LLM responses is using textual similarity-based metrics, which compare the generated response against a reference text. Widely used metrics in this category include BLEU [53] and ROUGE [40]. These metrics rely on word or n-gram overlaps between the generated responses and reference texts, evaluating the lexical similarity. One major drawback of lexical similarity-based metrics is that they have a weak correlation with human judgment, as they only capture the surface-level similarity, not the semantic relationship between words or phrases. On the other hand, researchers have also employed semantic similarity-based metrics as they can capture semantics in the responses even if the wordings differ. Metrics such as cosine similarity [56], BERTScore [71], and Semscore [10] can encode the semantic meaning of words or sentences.

Natural Language Interface (NLI)-based metrics, such as AlignScore [70] and Summac [38], offer a reference-free alternative to evaluate consistency. However, their performance is limited by generalization issues, often necessitating custom-trained models for specific tasks, limiting their wide acceptance. Another line of evaluating the consistency of responses is to use LLMs themselves as evaluators such as those proposed in [16, 24, 26, 42], but the accuracy of such evaluation is also often limited especially in domains that request specialized human expertise [64, 65]. The effectiveness of these metrics is often tied to the prompts used, which are tailored for particular datasets or tasks.

Finally, human evaluation is considered a gold standard for evaluating generated text. Numerous research used human-in-the-loop for evaluating consistency across various domains and applications [4, 9, 67]. However, it is labor-intensive, time-consuming, and requires subject-matter expertise, making it expensive for large-scale applications. Moreover, discussions on the biases and expertise in human judgment have also emerged, raising concerns about solely depending on human evaluators [15, 27, 62].

3 Methods

3.1 Investigating Location-based Information Seeking Queries on Reddit

Our research questions focus on analyzing LLM responses to various queries. To ground our study in real-world concerns and preferences, we sourced open-ended queries from genuine Reddit[2] community members discussing geographical locations. [3] Reddit, one of the largest online communities, hosts millions of active participants contributing to diverse discussions. Its forums reflect real user interests, trends, and challenges, while enabling extensive analysis for studying LLM recommendations.

We then identified queries of our primary areas of interest: *relocation*, *opening a business*, and *tourism* as those are the most commonly sought areas for information by Reddit users. We specifically looked for open-ended queries constrained to a particular region. Our search process involved keywords-based filters to extract queries related to geographic locations. In the *relocation* criteria, we targeted posts looking for best places to live or asking advice for moving. For *business*, we collected posts about the best cities or regions to start a business, as well as discussions on local business infrastructure to identify specific industries such as restaurant startups, coffee shops etc. In tourism, we focused on inquiring about travel destinations, sightseeing advice, and specific travel interests. To source the queries of interest, we used several relevant communities, including *r/relocating* (4.5K members), *r/traveladvice* (7.2K members), *r/AskReddit* (49M members), *r/tourism*(6.9K members), *r/StartingBusiness* (1.7K members), and *r/smallbusiness* (1.8M members).

3.2 Forming the Query Template and Experimental Setup

Following Section 3.1, for each domain, we derived two types of query conditions:

(1) *single-constraint* prompts: formed by a tuple of (<state>, <domain>, <constraint>)
(2) *generic* prompts: formed by a tuple of (<state>, <domain>)

For each domain, we crafted four prompts[4], each for *single-constraint* and *generic*, resulting in a total of 24 prompts for eliciting

[2] www.reddit.com
[3] The authentic queries can be found in the supporting material.
[4] The terms prompt and query are used interchangeably throughout this work.

Domain		Single-constraint
Opening a business	(i)	I want to open a coffee bookstore somewhere in **Oregon**. and I'm trying to find the best place to do it I'm looking for a place with *many people in their 20s and 30s*.
	(ii)	hi, I have been looking at moving to **Massachusetts** to carry out my career as a dog trainer. I'm looking for a *public-transit* friendly area to open my business.
	(iii)	If I were to open a restaurant in **Maryland**, where would you open one that is in a *walkable area*?
	(iv)	I'm looking to open a high quality bread/pastry bakery in a *safe* area in **Kansas**.
Relocation	(i)	I am making the move to **New Jersey**. I'm looking for a place with a *good bike score*.
	(ii)	We are moving to **Florida**. We are looking for communities who are at their *retirement age*.
	(iii)	We are planning to move to **Ohio**. We are looking for an *affordable area*.
	(iv)	We are looking to move to **Michigan**. We are looking for a *small town*.
Tourism	(i)	I'm visiting **Wyoming**. I am interested in visiting *wildlife habitats* and am looking to find places to stay nearby these sites.
	(ii)	I'll be touring **Arkansas**. I'm interested in visiting its *state parks* and am looking to find places to stay nearby these sites.
	(iii)	I'm visiting **Alabama**. I'd like to find *public fishing ponds* to visit and am looking to find places to stay nearby these sites.
	(iv)	We are visiting **Tennessee**. We're interested in visiting places of *historical heritage* and also looking to find places to stay nearby these sites.

Table 1: *single-constraint* prompts across the investigated domains of interest; business, relocation, and tourism. The <state> is highlighted in bold, while the specific <constraint> is italicized.

responses from LLMs. Table 1 illustrates *single-constraint* prompts across the three domains of interest. (the corresponding *generic* prompts are given in the supporting materials). Note that, in the table, <state> is highlighted in bold, while the specific <constraint> is italicized. The following state-of-the-art LLMs were evaluated — Claude-3.5 [8], Gemma [28], GPT-3.5 [51], GPT-4o [52], Llama-3.1 [45], and Mistral [47] and are shown in Table 2.

Each model was evaluated using its default settings, as we assume that everyday users are generally unfamiliar with the various options and configurations available. The selection of these LLMs was driven by their widespread use among everyday users, with some systems boasting over 100 million active users monthly. This aligns the chosen models with the context of our experiment, ensuring relevance to real-world LLM inference patterns.

3.3 Eliciting LLM Responses

In order to elicit responses we employed the Langchain[5] Python package and OpenRouter[6], where through its API key we could access the LLMs mentioned above for model inference. The input to OpenRouter consisted of a query (*single-constraint* or *generic*) together with the following instruction: "Can you recommend 5 cities or towns with multiple reasons for each recommendation". The resulting output was a JSON file with 5 different locations, and their corresponding justifications. While we request LLMs to generate justifications/reasons of locations provided, we do not suggest that LLMs possess logical reasoning capabilities. In this paper, we use the terms city, town, and location interchangeably throughout. The LLM responses and the code (for its collection and analysis presented in the following sections) can be found in `github.com/mohummedalee/cities-data-collection`.

Table 3 describes the overview of various notations used throughout the paper. For *single-constraint* as well as *generic* conditions, a total of $n_s = 40$ responses are generated per prompt, employing $n_L = 6$ LLMs across $n_p = 12$ unique prompts. This results in a total of $n_P \times n_s \times n_L = 2880$ responses per condition. Our goal was to analyze LLM responses on aggregate for each of the queries and for this purpose each LLM produced $n_s \times n_c = 200$ locations per query, and a total of $2,400$ locations per LLM to effectively investigate responses across multiple contexts.

3.4 Evaluation Measures

For our specific case of evaluating the locations within responses, we are interested in the similarity of the cities[7] generated by LLMs as well as the justifications provided for those cities. Table 4 summarizes the metrics employed to study these aspects of similarity in a quantitative fashion, and the corresponding data portions to which they were applied. We direct readers to the supporting materials for the precise notation and implementation details.

4 RQ1: Are LLMs Similar in Their Responses?

Analyzing the similarity between LLM responses can highlight both their differences and areas of alignment. A greater diversity in responses suggests a more inclusive experience that accommodates individuals from various backgrounds and starting points. In this section, we break down our first research question into the following components:

[5]https://github.com/langchain-ai/langchain
[6]https://openrouter.ai

[7]Throughout this paper, we use the terms city, town, and location interchangeably to refer to the same concept.

LLM	Model Version	Model Size (B)	Temperature	Max Tokens
Mistral	`mistralai/mistral-nemo`	7B	0.7	2048
Llama-3.1	`meta-llama/llama-3.1-405b-instruct`	405B	0.6	4096
GPT-4o	`openai/gpt-4o`	1.7T	0.7	8192
Claude-3.5	`anthropic/claude-3.5-sonnet:beta`	10B	0.7	4096
GPT-3.5	`openai/gpt-3.5-turbo`	175B	0.7	4096
Gemma	`google/gemma-7b-it`	7B	0.8	2048

Table 2: Default settings of selected Large Language Models (LLMs) with specific versions.

Notation	Description	Value
n_P	Total number of prompts per condition	12
n_s	Number of samples generated for every prompt	40
n_c	Number of locations requested for every prompt	5
n_L	Number of Large Language Models (LLMs) considered	6

Table 3: Experimental details and notations.

Similarity Method	Concept	Data Scope
Jaccard Index [50]	A similarity score calculated by the overlap (of items) between two lists relative to their total size	locations
TF-IDF [57]	Combines term frequency and inverse document frequency to assess word importance	location justifications
Cosine Similarity [46]	Measures semantic similarity between texts	location justifications
BLEU Score [53]	Measures n-gram similarity in text	location justifications

Table 4: Summary of similarity metrics and data scopes.

(1) Internal similarity: Are multiple responses generated for a given prompt by the same LLM similar?
(2) External similarity: Do different LLMs offer similar responses for a given prompt?

To address these two derived research questions, we formalize the process by which *internal* and *external* evaluations applied the similarity metrics described in Section 3.4.

4.0.1 Internal Evaluation. Note that for internal comparison, the set of responses contains n_s entries, where n_s denotes the number of responses sampled for a given prompt from a specific LLM. Each of the response entries R_i contains a list of n_c towns/cities along with their justifications, as requested in the prompt. Our internal similarity evaluation computes scores (Jaccard, TF-IDF, cosine similarity, BLEU) for each R_i, (with respect to the other responses in the set) as described in Table 4; this is repeated for all the prompts.

4.0.2 External Evaluation. External evaluation is performed at a higher granularity, where all response samples from an LLM are concatenated, and comparisons between LLMs are conducted. In this case, the set of responses contains n_L entries, where n_L denotes the number of LLMs under consideration. Each of the response entry R_i is constructed by combining the response samples across all queries for that LLM. Specifically, each response R_i contains a list of $n_c \times n_s$ towns/cities along with their justifications, where n_c is the number of locations requested in the prompt and n_s is the number of samples generated per query per LLM. Our external similarity evaluation computes scores (Jaccard, TF-IDF, cosine similarity, BLEU) for each R_i, (with respect to the other responses in the set) as described in the Table 4; this is repeated for all the prompts.

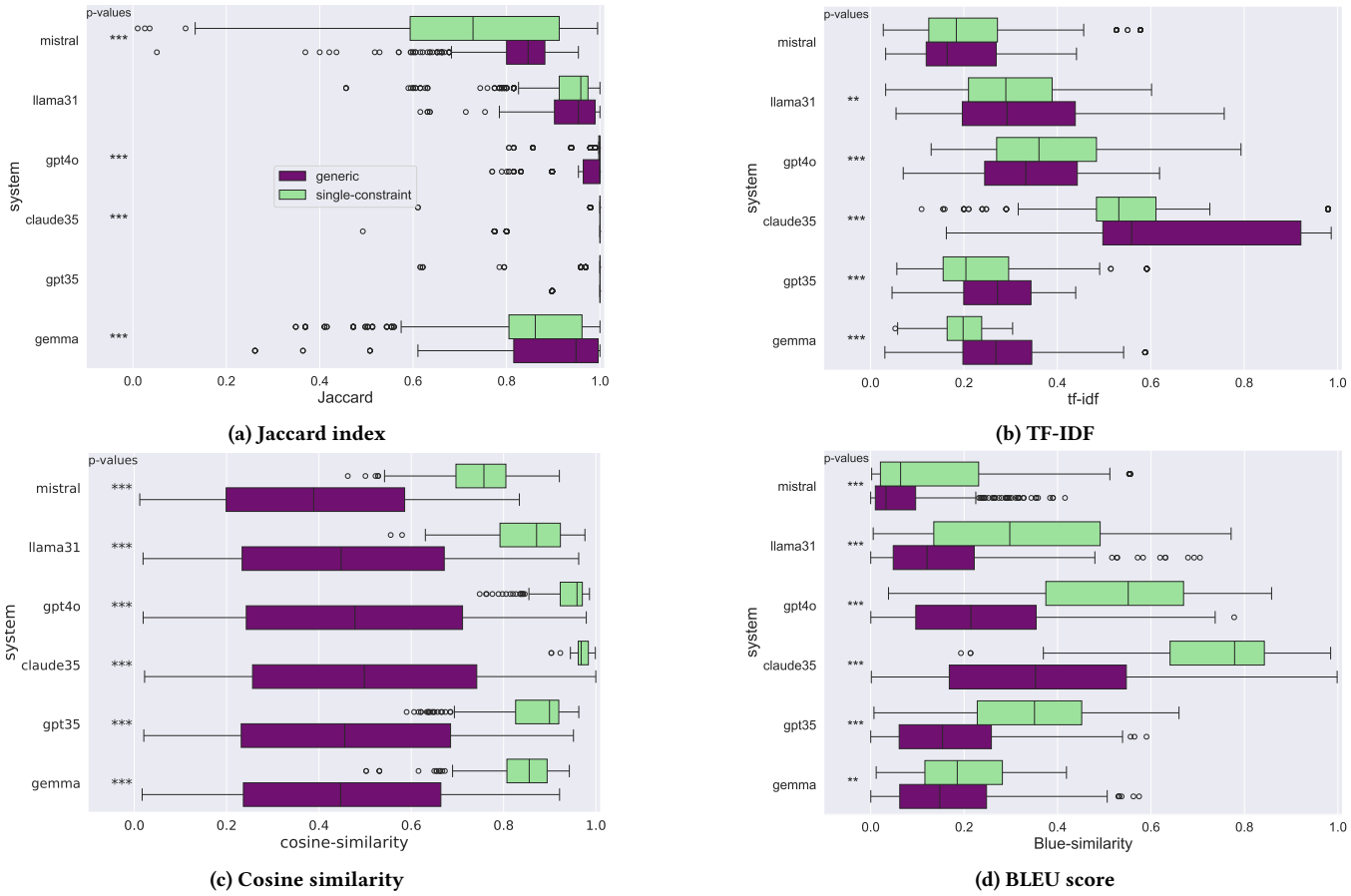
4.0.3 Statistical Significance. We conducted two-tailed t-tests to learn about differences in distributions between *single-constraint* and *generic* conditions. p-values < 0.05 were translated to significance level such that $p < 0.05$, $p < 0.01$, $p < 0.001$ were indicated by $*$, $**$, and $***$ respectively.

4.1 Are LLMs *Internally* Similar in the Cities/Towns Recommended?

Here, we are interested in addressing if the different responses generated for the same prompt, by the same LLM are similar or diverse. This is motivated by the attempt to study responses for the same query *in aggregate* to learn about potential trends emerging from LLMs.

Figure 1 demonstrates the difference between the evaluation measures for both *single-constraint* and *generic* conditions for each of the LLMs: Mistral, Llama 3.1, ChatGPT-4o, Claude 3.5, ChatGPT-3.5, and Gemma. To statistically study the difference between the observed measures for the two conditions, two-tailed t-tests are conducted and their corresponding p-values are indicated in a left-hand panel in each plot.

In Figure 1a, the p-values computed for the Jaccard scores of *single-constraint* and *generic* prompts show a significant difference for Mistral, GPT-4o, Claude 3.5, and Gemma in the pool of locations,



(a) Jaccard index **(b) TF-IDF**

(c) Cosine similarity **(d) BLEU score**

Figure 1: Internal comparison of *single-constraint* and *generic* conditions across LLMs using different similarity scores. p-value significance levels for the comparison between the two conditions are shown on the left side of each plot. Error bars reflect the variance of pair-wise scores comprising each distribution.

when comparing similarity scores across conditions. However, no significant difference was observed for Llama 3.1 and GPT-3.5. We also found that Mistral had the lowest median score, around 0.75 for *generic* prompts and above 0.8 for *single-constraint* prompts, indicating overall higher Jaccard scores and demonstrating high overlap in the cities generated by the LLMs for each of the conditions.

We analyzed the similarity of location justifications using TF-IDF scores in Figure 1b. All models, except Mistral, showed differences between conditions, indicating that the key words and word choices identified by TF-IDF varied across prompts. The figure also reveals that the highest median score is around 0.5, with many responses clustering around the 0.4 to 0.2 range, indicating small overlaps and thus limited similarity based on this measure.

Our analysis of the semantic similarity in city justifications is shown in Figure 1c. We observe clear differences between the conditions provided. In the *generic* condition, the distribution of semantic similarity is wide, with lower overall similarities, while in the *single-constraint* condition, similarities are much closer to 1, with a narrower range. This suggests that applying the *single-constraint* prompts reduced the variety of justifications given for the selected locations.

The final similarity measure used to evaluate the justifications of LLMs was BLEU, as shown in Figure 1d. Here, too, both conditions across all LLMs show significant differences in phrases. The similarity distributions reveal that the *single-constraint* condition had overall higher n-gram similarity, reflected in higher score ranges and a median score typically above that of the *generic* prompts. So applying the *single-constraint* made the wording of the justifications more consistent.

a) The lowest Jaccard median score was 0.8, indicating that LLMs exhibited a high degree of similarity in the cities and towns generated across repeated queries. This indicates that LLMs tend to produce similar sets of cities and towns when prompted multiple times, which may limit diversity in the options provided to the user.
b) TF-IDF revealed relatively low similarity in the *important* words used in the justifications. Additionally, while word choices became more similar in BLEU after applying the *single-constraint*, there remained notable lexical variation.

c) Despite the low TF-IDF and BLEU scores, the semantic consistency of justifications in the *single-constraint* condition was high, indicating that the underlying justification was similar, though expressed in different words. However, the limited range of justifications could restrict our understanding of the unique qualities of different locations, as people may base their choices on varied factors.

4.2 Are LLMs *Externally* Similar in the Cities/Towns Recommended?

Examining external similarity is important, as it reveals whether different LLMs converge on similar responses despite having differences in training, architectures, or data when responding to a specific prompt.

Figure 2 shows the differences between *single-constraint* and *generic* prompts across LLMs: Gemma, ChatGPT-3.5, Claude 3.5, ChatGPT-4o, Llama 3.1, and Mistral incorporating the measures described above. Here too, we present two-tailed t-tests on both conditions and their corresponding p-values are indicated in a left-hand panel in each plot.

In Figure 2a, the p-values computed for the Jaccard index show varying degrees of significance between the *single-constraint* and *generic* conditions across all LLMs. That together with the visualization suggest that *generic* locations are more similar across LLMs, while, unexpectedly, the *single-constraint* prompts result in less similarity.

In Figure 2b, the p-values calculated on TF-IDF scores on justifications for locations show that there is no statistically significant difference in scores between *single-constraint* and *generic* prompt as indicated by the lack of '*' across LLMs. Furthermore, the responses show considerable diversity in terms of TF-IDF as indicated by the low TF-IDF scores, reflecting distinct representative words that were generated for the same prompt across LLMs.

Figure 2c presents the cosine similarity scores of the justifications of cities between *single-constraint* and *generic* prompts. Here too, there was no difference in conditions, and both presented considerable overlap. However, different from TF-IDF, we find that semantic similarity around the justifications had a median ranging between $0.65 - 0.75$ offering a reasonable degree of similarity across LLMs.

In Figure 2d, the p-values computed for the BLEU scores between *single-constraint* and *generic* conditions indicate significant differences for Gemma and GPT-3.5. The overall low BLEU scores across LLMs indicate that the phrases used in the justifications/reasons were more diverse for a given prompt.

a) When comparing similarities of responses for city names, we find that the *single-constraint* condition presented lower similarity rates across responses compared to the *generic*.
b) The overall representative words (indicated by TF-IDF) as well as the phrasing of the responses (indicated by BLEU) suggest that the generated reasons are diverse on those fronts.
c) However, despite the above, the high cosine similarity indicates that the justifications behind each city tends to be semantically similar between *single-constraint* and *generic* across LLMs.

In summary:

a) The list of towns generated by LLMs tends to show significant overlap when queried multiple times. However, this list may differ when compared across different LLMs, as indicated by the Jaccard index.
b) The phrases and words used to describe and justify the locations vary, suggesting differences in language use both within the same LLM and across different LLMs, as shown by TF-IDF and BLEU scores.
c) Despite these differences, the underlying semantics of the justifications of locations remain fairly consistent within the same LLM and are not drastically different across LLMs, as reflected by cosine similarity scores.

5 RQ2: What Kind of Locations are Recommended?

Understanding whether LLMs disproportionately represent cities or towns with certain characteristics in their responses is essential for promoting fairness and inclusivity. LLMs are trained on vast datasets from the Web, which may carry inherent biases related to socio-economic, cultural, or geographical factors. If specific locations (or types of locations) are consistently over- or under-represented for the same queries, this can create a skewed distribution, reinforcing existing inequalities and limiting the diversity of experiences offered to users. In this section, we aim to identify and assess these biases through both *intrinsic* and *extrinsic* evaluations in the following subsections respectively. For a concise analysis we focus on the *single-constraint* condition throughout this section.

- Intrinsic evaluation: How do distributions of cities/towns in LLM-generated responses reflect opportunities based on the frequency of cities suggested?
- Extrinsic evaluation: How do distributions of cities/towns in LLM-generated responses reflect real-world possibilities based on external data sources?

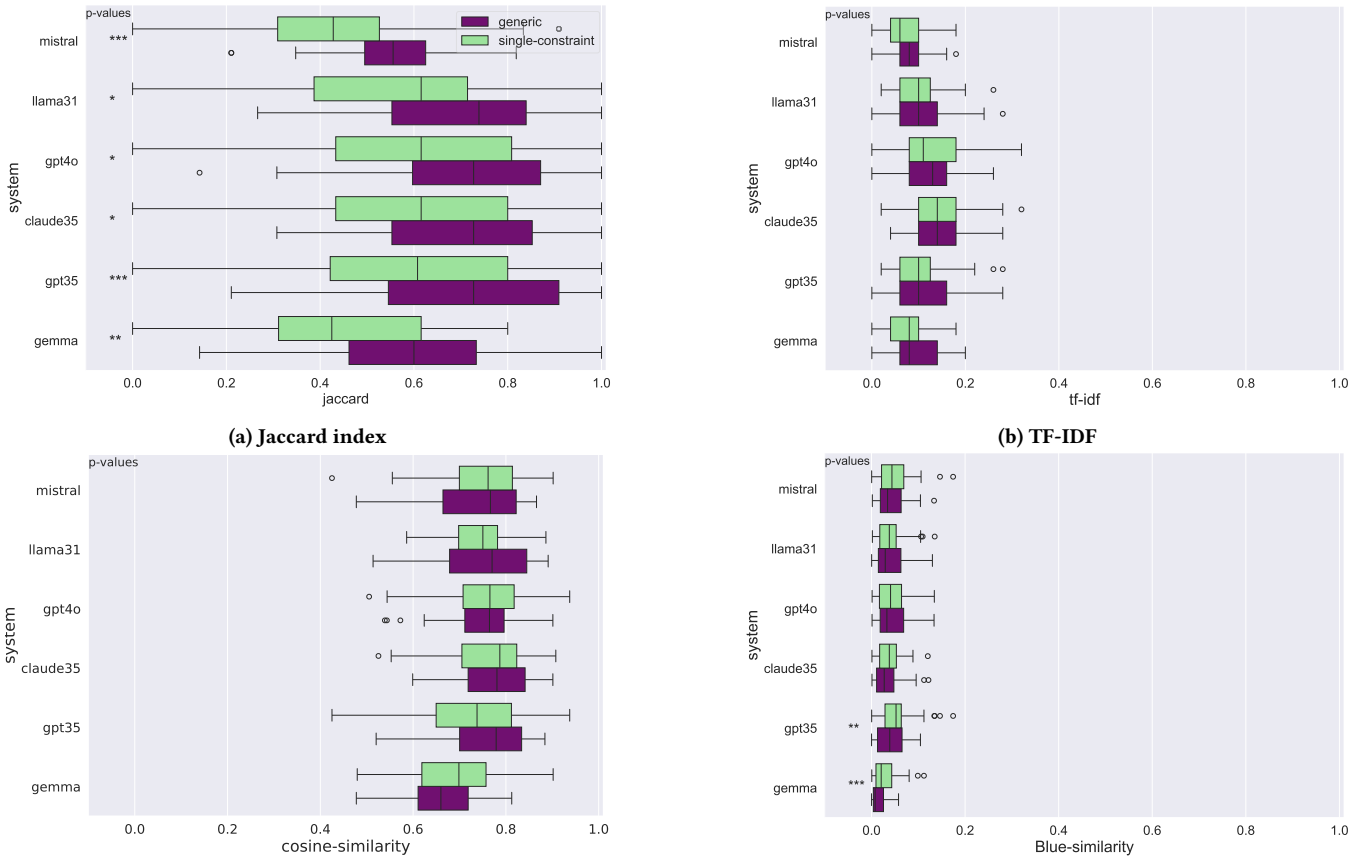
Intrinsic evaluation focuses on analyzing the distribution based on the generated responses whereas *extrinsic* evaluation compares LLM distributions with real-world data/distributions.

5.1 Intrinsic Evaluation of LLMs' Distributional Properties

We conduct intrinsic evaluation on the distributions generated by LLMs, focusing on named entities such as city or town locations. Each distribution is based on the frequency of these entities across 40 responses generated per LLM for each query, yielding a total of 200 cities per query per LLM (5 cities per response). Our goal is to assess the distributional properties of LLM responses using metrics that measure distributional inequality.

5.1.1 *Measuring Distributional Inequality of Cities/Towns.*

The metrics employed for this purpose are concentration ratio and Theil index. The concentration ratio [18] measures the dominance of the most frequent entities; it can be calculated as the cumulative proportion of occurrences attributed to the top K most frequently occurring entities. This metric is used in the field of economics to quantify market share and market concentration. A value closer to 0 indicates that the top K entities have no dominance, whereas a value closer to 1 indicates complete dominance,



(a) Jaccard index

(b) TF-IDF

(c) Cosine similarity

(d) BLEU score

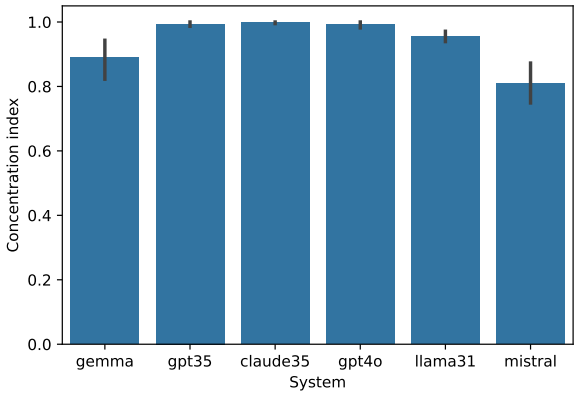
Figure 2: External comparison of *single-constraint* and *generic* conditions across LLMs using different similarity scores. p-value significance levels for the comparison between the two conditions are shown on the left side of each plot. Error bars reflect the variance of pair-wise scores comprising each distribution.

i.e., the top K entities account for all occurrences. Concentration ratio is sensitive only to the frequencies of the top K entities; it will not change if the frequencies of entities outside the top K change. In this analysis, we display the cumulative proportion of top 5 locations made by LLMs. We selected the Theil Index [17] as it is an inequality measure based on entropy, specifically focused on deviations from equal distributions, with values ranging from 0 (perfect equality) to infinity (extreme inequality). Unlike entropy, the Theil Index explicitly accounts for the number of unique locations produced. The formulae we employed are included in the supporting material. Theil index is complementary to concentration ratio as it analyzes the equality of the entire distribution whereas concentration ratio only focuses on the frequencies of the top K entities. Figure 3 presents the respective analyses of concentration ratio and Theil index in 3a and 3b across LLMs for all queries.

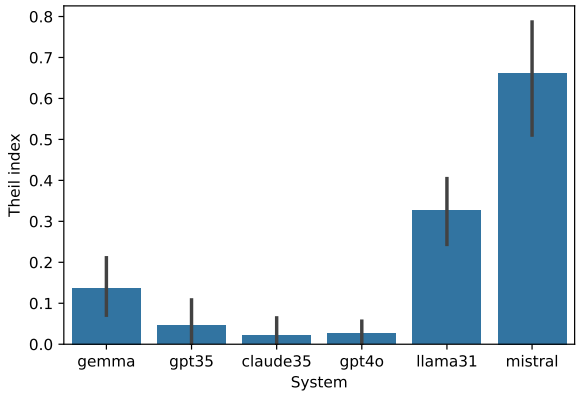
In Figure 3a, we see that the cumulative distributions of the LLMs for the top five locations approach 1 and, at a minimum, encompass 0.8 of the distribution. This suggests a strong preference for the same five locations across the majority of queries. Theil Index in Figure 3b provides a complementary perspective to this analysis. In addition to the preference for these five locations, systems that are *approaching* a concentration ratio of 1—such as Claude-35, GPT-4o, GPT-35, and Gemma—exhibit a nearly uniform distribution across these locations, resulting in a minimal or nonexistent distributional tail. On the other hand, systems with slightly lower concentration ratio (between $[0.8 - 1)$) also display heavy preference towards top 5 locations, such as Llama-31 and Mistral; but, they were less uniform and presented a longer tail.

We note that it is likely that observing 5 locations is tied to our prompt, which specifically requests 5 locations. Alternatively we assume the model would have repeated x locations based on the value of x in the prompt. We leave further investigation of this for future work.

The *same* set of locations tends to be repeatedly generated across multiple samples for most queries in LLMs. The size of this set is likely influenced by parameter biases introduced in the prompt (requesting for 5 locations in our experiment). This *limited* distribution, characterized by a strong preference for a few locations, reduces exposure to alternative places and, in aggregate, may contribute to a "rich get richer" effect.



(a) Concentration ratio represents distribution inequality across top 5 locations produced by LLMs.



(b) Theil Index present distribution inequality across the entire distribution of LLM locations.

Figure 3: Theil Index and concentration ratio of elicited responses by LLM. Error bars reflect the variance of concentration ratio and Theil index across respective distributions.

5.1.2 Measuring Distributional Inequality of Demographic Attributes. Given the strong preference for certain locations and the lower representation of others, we examine whether there are any demographic differences between the over-represented and under-represented cities within these distributions. Note that this section does *not* compare with real-world distributions; that is the focus of Section 5.2.

Demographic attributes: To obtain the demographic information for cities, we use the U.S. cities database.[8] The database is built using sources such as the U.S. Geological Survey and U.S. Census Bureau, providing information about over 109, 000 cities and towns from all 50 states pertaining to various fields such as population, income, age, race, gender, marital status, home value, education, disability, and more. In this section, our focus is on *demographic attributes* of historically underserved groups pertaining to race, gender, health and financial status, where we evaluate how well they are represented or potentially under-represented. Table 5 describes the *demographic attributes* that are investigated in this part of our analysis[9].

Metric employed: Since all demographic attributes are numerical, we utilize the skewness metric to assess the asymmetry of the data distributions. Skewness describes distribution asymmetry. **Positive skewness** (right-skewed) has a longer right tail with high outliers, making the mean greater than the median. **Negative skewness** (left-skewed) has a longer left tail with low outliers, where the median exceeds the mean. **Zero skewness** indicates a symmetrical distribution with equal mean and median.

It is important to note that skewness only reflects the direction and asymmetry of the distribution; it does not provide insight into whether the distribution accurately represents the underlying demographic as we are conducting an intrinsic analysis and not comparing distributions to external *ground truth* sources. For historically underserved communities, we would ideally prefer negative skewness, as this suggests that data points are concentrated around higher values, indicating more equitable outcomes.

Results: Figure 4 presents race and gender skewness corresponding to the attributes in Table 5 split by domain. Historically underserved races are found to be consistently positively skewed across LLMs and across domains indicating that more data points are cluttered around the lower values. The attribute `race_black` exhibits relatively low positive skewness, and in a few instances (for certain domains on specific LLMs), it is nearly centered around zero. In contrast, nearly all other underserved racial groups display significantly higher skewness values, with `race_pacific` showing some of the highest skewness. This shows that the percentage of historically underserved populations in the provided locations tends to cluster around lower values, resulting in a right-tailed distribution. This suggests a lack of inclusivity and diversity in the recommended locations, reflecting an uneven representation of these populations.

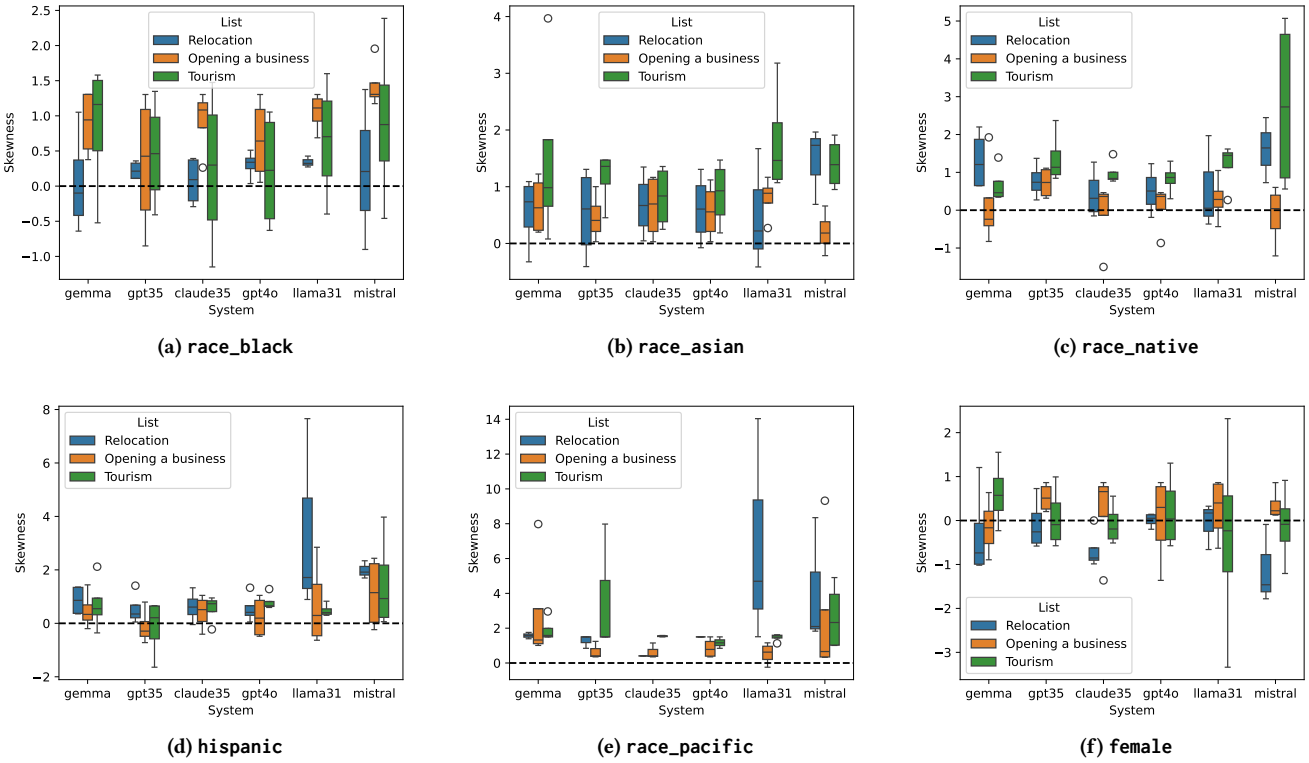
Figure 4f presents our analysis on `female`, which is another historically underserved group. We observe that for queries related to relocation, the distribution of female population in the recommended locations is mostly negatively skewed, which is a favorable outcome, as it indicates that a higher percentage of women are concentrated in these areas. However, the converse is true in the case of opening a business, where more positive skewness is observed. In case of tourism, the skewness values are mostly centered around zero, exhibiting no tilt in the distribution.

The skewness of attributes related to unemployment rate, percentage of disability, and poverty is shown in Figure 5. Similar to the race attributes, unemployment rate exhibits positive skewness across domains for all LLMs, with the exception of "opening a business" in the case of Gemma. For the percentage of disabled population and poverty, an interesting pattern emerges: in prompts related to relocation and opening a business, positive skewness is observed across all LLMs. However, in the case of tourism, the

[8]https://simplemaps.com/data/us-cities
[9]In the appendix, we analyzed additional *demographic attributes* that are associated with over-representation.

Demographic Category	Demographic Attributes
Race	`race_black` , `race_native` , `hispanic` `race_pacific`, `race_asian`
Gender	`female`
Financial	`unemployment_rate`, `poverty`
Health	`disabled`

Table 5: *Demographic categories* associated with their *demographic attributes*.



(a) `race_black` **(b) `race_asian`** **(c) `race_native`**

(d) `hispanic` **(e) `race_pacific`** **(f) `female`**

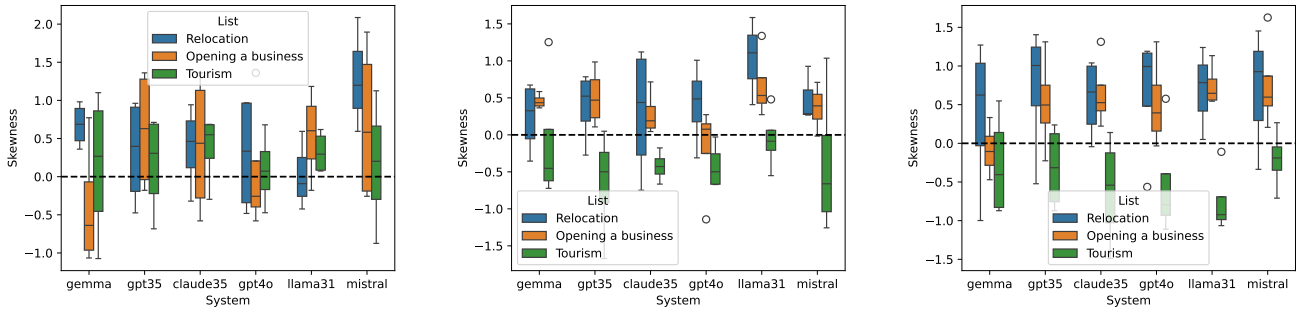
Figure 4: Skewness of attributes pertaining to historically underserved race and gender groups. Error bars describe the variance of an LLM distribution for a particular attribute.

skewness is notably negative. This suggests greater inclusivity for populations with disabilities and those in poverty within touristic responses compared to the other domains.

A majority of the attributes corresponding to historically underserved communities are positively skewed, indicating that the corresponding distributions contain more data points in the lower end of the spectrum, reflecting an uneven treatment of these groups. Tying this evidence with the finding in Section 5.1.1, the limited locations presented in Section 5.1.1 is unfavorable towards the historically underserved stakeholders shown in this part.

5.2 Extrinsic Evaluation of LLM Distributional Properties

Now we compare the responses generated by each LLM for the *single-constraint* condition with an external source, a list of U.S. cities or towns that meet the specified query criteria—referred to as the *database distribution*. We aim to determine whether the LLM-generated distributions are representative of the *types* of cities available. The objective is to assess how statistically similar these LLM-generated distributions are to the database and identify any misalignments between them. We evaluate the similarity of the distributions by analyzing the presence of specific *demographic attributes*. Discrepancies in these attributes may signal the inclusion or exclusion of certain *types* of locations, which could disadvantage demographically historically underserved groups by limiting their access to relevant opportunities. Additionally, this exclusion may hinder the visibility and growth of underrepresented cities, especially if LLM-generated recommendations are widely adopted.



(a) `unemployment_rate` (b) `poverty` (c) `disabled`

Figure 5: Skewness of attributes pertaining to unemployment, disability and poverty.Error bars describe the variance of an LLM distribution for a particular attribute.

Conversely, the over-representation of a select few cities could amplify a "rich-get-richer" dynamic, where only a handful of cities are positioned for future growth.

5.2.1 Forming the City Database Distributions. Forming comparable lists was conducted for each of the 12 queries from the *single-constraint* condition (see Table 1) based on the US city database and additional resources. Table 6 outlines the processes and databases used for this purpose.

We employed US city database, Walk Score[10], WY wildlife habitat[12], AR state parks,[13], AL public fishing ponds[14], TN historic sites[15] and consulted with US Census Bureau resources[16]

For constrains 1, 6, and 7, we sorted locations based on the attributes in the U.S. city database, forming a list based on the first quartile that exhibited the highest rates of that particular attribute; for instance, the highest percentage of the combined attributes of `age_20` and `age_30` for the constraint '*place with many people in their 20s and 30s*'. Constraints, 2, 3, and 5 extracted locations based on highest mobility score based on the Walk score database. Constrains 9-12 were based on extracting towns in the vicinity of the described constraint, within an empirically determined radius to form the list of close proximity towns.

5.2.2 Demographic Attributes. The experiment described below focuses on comparing distributions for similarity across their distributional properties. To this end we selected eight *demographic categories*, each based on one or more *demographic attributes*. Our comparison utilizes a one-tailed t-test, where the alternative hypothesis (H_1) posited that the sample mean of the LLM is either greater or smaller than the population mean of the *database distribution*. To this end, we identified the *demographic attributes* for which 'smaller' or'larger' values, respectively, could potentially introduce biases limiting opportunities for historically underserved users.

In selecting *demographic attributes*, as shown in Table 7, the terms 'smaller' and 'larger' are used to approximately indicate instances of under-representation or over-representation in LLM responses that might disadvantage historically underserved stakeholders. Within the financial category, higher values in LLM distributions indicate less affordable opportunities for users. Some attributes are classified under both 'smaller' and 'larger' as both under- and over-representation of those attributes may be of concern, based on the context. For instance, deviations in median age (`age_median`) and the proportion of never-married users were flagged as concerns on both accounts,[17] Additionally, fewer cities with divorced individuals, women, or people with disabilities under the family structure, gender, and health categories were noted as problematic, given that these groups tend to be more historically disadvantaged, and hence, are categorized as 'smaller' demographic attributes. Along the same lines, we checked for over-representation of white individuals (`race_white`) and under-representation of non-white groups. Finally, an over-representation of highly educated people, as well as cities with significantly longer commute times compared to relevant alternatives, may also limit opportunities.

5.2.3 Results. We compared the *database distributions* to their respective responses per LLM (across 12 different queries). We conducted *t*-tests across every *demographic category* across its *demographic attributes*. Our null hypothesis H_0 is that there are no demographic differences between the distribution of cities between the ones generated by an LLM to the *database distributions* which we evaluated through applying *t*-tests. *p*-values of each t-test were translated to the following $p < 0.05$, $p < 0.01$, and $p < 0.001$ correspond to $*$, $**$, and $***$, to indicate the distribution of *p*-values, and the likelihood that the LLM responses were drawn from the *database distributions*.

Figure 6 presents the results across eight demographic *categories*, comparing LLM outputs to *database distributions*. For most of the

[10], KS crime index[11]https://www.walkscore.com/professional/research.php
[12]https://wgfd.wyo.gov/Public-Access/WHMA
[13]https://koordinates.com/layer/102903-arkansas-state-park-state-park-locations/
[14]https://www.outdooralabama.com/where-fish-alabama/alabama-public-fishing-lakes-pfls
[15]https://www.tn.gov/historicalcommission/state-programs/state-historic-sites.html
[16]https://www.census.gov/library/stories/2020/05/america-a-nation-of-small-towns.html

[17]US cities often prioritize `never_married` individuals The major cities being designed without children in mind, BBC article while other places may show the opposite trend Why do people in cities stay single longer vs suburbs/rural marrying young?, Reddit post

	Constraint	Database	Process	State
1	***people in their 20s and 30s***	US cities	top quartile cities of combined attributes `age_20s`, `age_30s`	OR
2	***public transit friendly***	Walk Score US cities	include cities of `rider's paradise` `excellent transit`, `good transit scores`	MA
3	***walkable area***	Walk Score US cities	same as in 2	MD
4	***safe***	KS crime index	top quartile cities	KS
5	***good bike score***	Walk Score US cities	same as in 2	NJ
6	***people in retirement age***	US cities	top quartile cities of `age_over_65`	FL
7	***affordable area***	US cities	top quartile cities of combined attributes of below OH median income `['income_household_20_to_25',..,` `'income_household_50_to_75']`	OH
8	***small-town***	US Census Bureau US cities	include cities of `population_proper` less than $5,000$	MI
9	***towns near wildlife habitat***	WY wildlife habitat US cities	include cities within less than 4 miles to listed sites	WY
10	***towns near state parks***	AR state parks US cities	same as in 9	AR
11	***towns near public fishing ponds***	AL public fishing ponds US cities	same as in 9	AL
12	***towns near historical heritage sites***	TN historic sites US cities	same as in 9	TN

Table 6: This table describes the process to generate the *database distribution* datasets based on formal databases on U.S. cities. For instance, in order to extract the cities to address the first constraint, attributes of particular ages were combined, sorted, and formed by the second median (top quartile) of that list.

demographic categories—namely, *financial status, family structure, age, gender, health, education, and geography*—we observe similar patterns across the LLMs, as indicated by comparable p-value levels. In contrast, for the *race* category, Mistral showed the closest alignment with the *database distribution*, while GPT-3.5 exhibited the greatest deviation.

Next, we decomposed each of the *demographic categories* into their respective *demographic attributes*. This analysis allows us to assess the contribution of each attribute to the differences observed in Figure 6. We perform t-tests across LLMs for all the queries for every attribute of interest. Figures 7 and 8 illustrate this breakdown for the 'smaller' and 'larger' relations, respectively, by counting the number of tests with p-values below 0.05.

It is interesting to note from Figures 7 and 8 that for all *demographic attributes*, we observe that LLMs produce distributions that deviate from the *database distribution* at similar rates, indicating that demographic biases are consistent across LLMs. As Figure 7 presents the 'smaller' demographic attributes, note that a higher count indicates that the attribute is under-represented for as many prompts. LLM responses tend to consider less cities that has divorced individuals; however, it does not under-represent cities with individuals that never married. It also tends to under-represent locations with an older population (over 65) more significantly than those with a younger population (under 19). It under-represented cities that can be relevant to protected races; this is especially pronounced for `race_pacific` when compared to the other protected races. It also indicated 6-8 counts for `family_size` that reflect under-representation of smaller families. We also find under-representation of cities with higher unemployment rates, where lack of exposure to these cities may reduce opportunities

Demographic Category	'Small' Demographic Attributes	'Large' Demographic Attributes
Financial	unemployment_rate, poverty	home_value, rent_median, IH_150k_over, IH_100k_to_150k, IH_median
Family Structure	family_size, never_married, divorced	family_size, never_married
Age	age_median, age_over_65, age_over_80, age_under_10, age_10_to_19	age_median
Gender	female	
Race	race_black, race_asian, race_native, hispanic, race_pacific	race_white
Health	disabled	
Education		college_or_above
Geographic		commute_time

Table 7: Categories corresponding 'small' and 'large' *demographic attributes*.

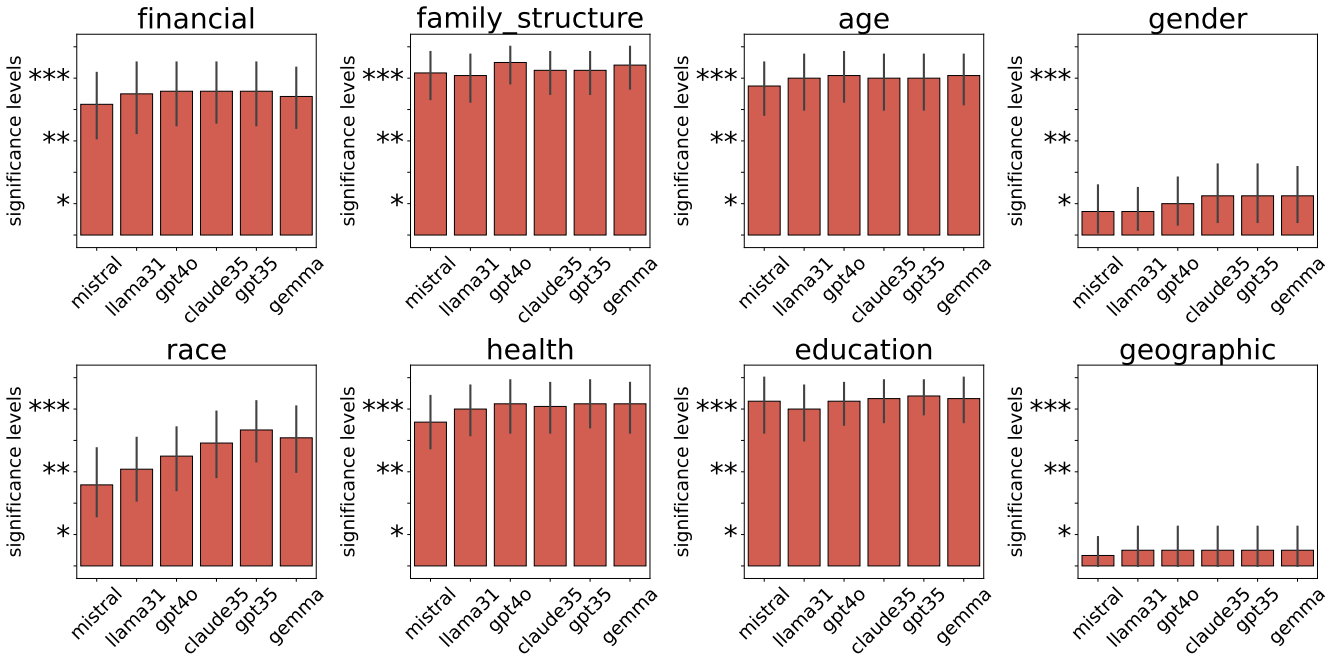


Figure 6: Demographic *category* comparisons between LLM responses to *database distributions*. The x-axes corresponds to LLM systems while the y-axes indicate p-value strength levels ($p < 0.05$, $p < 0.01$, and $p < 0.001$ correspond to $*$, $**$, and $***$). For instance, for in financial category there are distributional differences that are similar across LLM in their strength indicated by $**$.The error bars describe the variance of p-values within an LLM.

both to those places and to potential residents. One of the most significantly under-represented attributes is `disability`, where LLMs consistently under-represent cities in terms of percentage of disabled population.

On the other hand, as Figure 8 presents the "larger" demographic attributes, a higher count indicates that the attribute is over-represented for as many prompts. It over-represented financially more affluent cities, with higher home values, higher rent and greater household income. LLMs also over-represent cities with individuals that never married. It also significantly over-represents places with people who are college-educated in over 10 out of 12 counts.

Finally, we evaluated `age_median` for both 'smaller' and 'larger' categories to assess deviations from the *database distribution*. We found that in approximately 6-8 instances, the LLM outputs leaned towards younger populations, while in about 1-2 cases, the LLM reflected older residents relative to the age in the *database distribution*.

When comparing the location produced by LLMs to external database sources to learn about how inclusive these responses get, we find that *demographic attributes* that represent historically underserved stakeholders are less represented in these locations. This results in (1) limited exposure to cities of these *demographic attributes* (2) limited representation of users of these demographics. We also found over-representation of *demographic attributes* of stronger population, reinforcing the rich getting richer effect.

In summary we find both intrinsic and extrinsic biases in LLM responses on the types of locations provided in their responses, and demonstrate that these types of biases under-represent historically underserved populations.

6 Discussion

6.1 Impacts of LLM Geographic Biases on Communities

The findings of this study show that LLM-generated responses tend to under-represent historically underserved groups and communities with fewer financial resources.

Intrinsic evaluation representation: We found that cities recommended more frequently tended to under-represent historically underserved populations, as they were positively skewed in the inequality and concentration ratios we applied. This included underrepresentation of five non-White racial groups, people with disabilities, areas with higher unemployment rates, lower-income regions, and women, particularly in job-related queries.

Extrinsic evaluation representation: We find that these demographic attributes (except for Asian race), together with attributes indicating older or younger population and divorced individuals are under-represented when comparing them to U.S. city database. On the other hand, we found that attributes associated with stronger financial means as well as higher education we more represented compared to the same U.S. City database.

This finding has two significant implications: first, it reduces the visibility of towns and cities linked to these demographics, making it more difficult for potential residents, entrepreneurs, or tourists to discover and connect with these locations, thereby reinforcing a "poor getting poorer" dynamic. In this way, using LLMs in their current form can limit opportunities for less privileged individuals and hinder their social and economic mobility. If widely adopted, LLM will be shaping our *future* based on our *past* data as noted by both by Birhane, 2022 [13] and Vallor, 2024 [66]. At its best, it can limit growth opportunities for less visible cities, and at its worst, be detrimental. It may also limit the social and economic mobility of less visible people.

Second, the responses disproportionately favor a select group of cities, typically associated with financially well-off or highly educated demographics, reinforcing a "rich-get-richer" effect. While this may benefit these cities in the short term, it limits opportunities for all residents across the country in the long term, letting LLM shape the economic, cultural, and political futures of the places we live.

LLM Similarity: First, we found on the internal evaluation on text, there is relatively high overlap across responses within the same LLM around the cities recommended, and second, that similarity scores around the semantics of the free-form text around those locations had exhibited high scores, when evaluated on textual similarity. Fianlly, we also find that LLMs deviate from the external database at the same rate, when evaluated for representation. These are evidence for similarities, that LLMs may not offer meaningful differences that cater to diverse backgrounds.

More broadly, the similarities we observed in the LLMs responses may stem from the underlying data and the use of transformer-based architecture, which are likely to produce outputs with similar distributional properties, as discussed by [30]. This concern extends beyond the over- or under-representation of specific populations to a more fundamental risk: relying on systems that generate homogeneous responses could contribute to the development of a monoculture, as highlighted by [14].

6.2 Implications for Designers and Developers of LLM-Enabled Applications

The current setup of LLMs, under which our task was conducted, generates responses based on limited context. Specifically, the query only provides basic details such as state, domain, and a particular constraint, without including any personal information or specific needs. To make such technology more inclusive, the assumption is that responses should address a broader range of users, especially across diverse demographics. We note that while we demonstrated demographic differences based on available data, not all life experiences are easily quantifiable, which limits our study's interpretation of what truly constitutes "inclusive" technology.

In cases where the initial responses do not account for diverse life experiences, an alternative approach is to engage in *follow-up questions*. Leveraging the conversational nature of LLMs can yield more personalized and relevant responses, tailored to the user's specific context. If follow-up questions are not an option, another alternative is to indicate that the response may be *incomplete*, as shown by [36]. In such cases, such alternatives could involve offering caveats about the limitations of the answer and providing links to multiple relevant sources or websites, similar to how search engines present a diversity of perspectives.

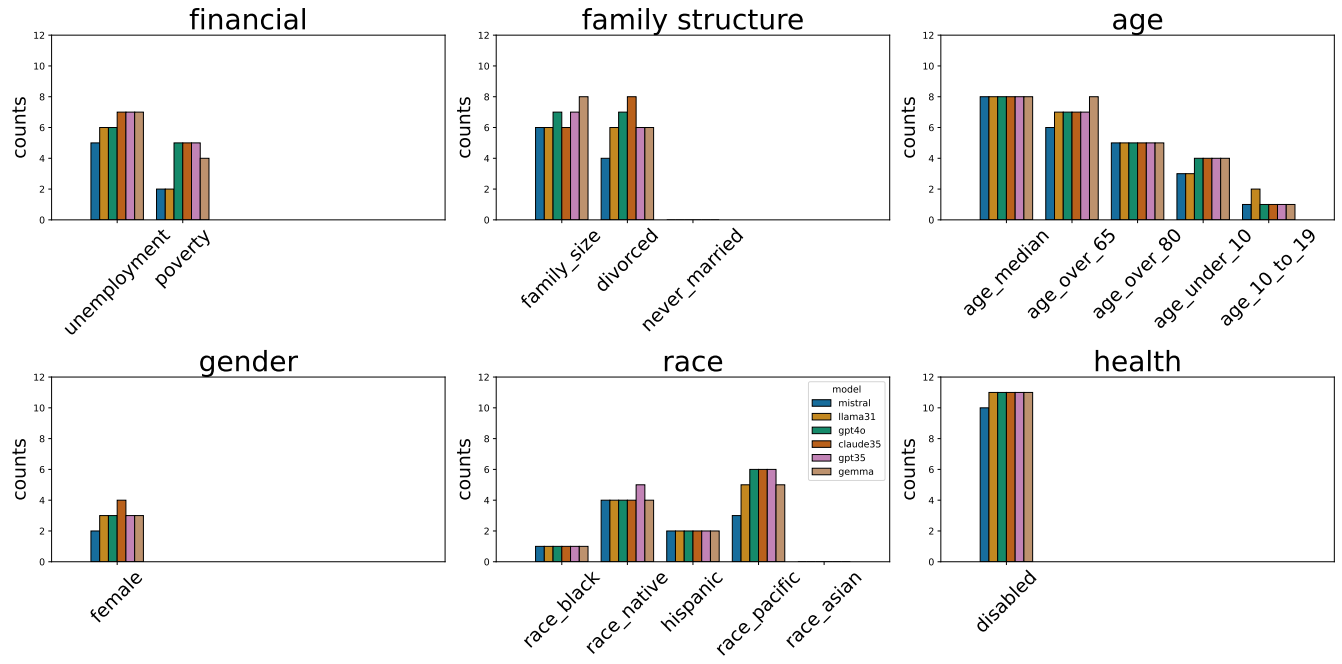


Figure 7: Demographic *category* comparisons between LLM responses to *database distributions* for 'smaller' demographic attributes. The x-axes corresponds to *demographic attributes* that were evaluated under a certain *demographic category*. Each *demographic attributes* indicates a respective value of an LLM system. These values correspond to counts (indicated on y-axes) from a total of 12 distributional comparisons per each LLM (as there are 12 queries). For instance, across 12 comparisons on the health category, about 11 comparisons under-represented cities that can accommodate individuals with disabilities, across all LLMs.

6.3 Biases in Data Representation for LLMs

The vast amount of online data used to train LLMs is known to contain biases and is not fully representative of global populations nor U.S. residents. As a result, certain demographic groups with less political and economic power are underrepresented in the training data (data bias) and are even less likely to be adequately represented in the model outputs (algorithmic bias) [21]. This amplifies exclusion and exacerbates existing inequalities on a much larger scale. As these systems become more widely adopted, it becomes increasingly important to evaluate their algorithmic biases and explore ways to mitigate them, to ensure that LLMs promote the most responsible, ethical and representative information possible.

7 Limitations

There are several limitations and threats to validity to the study we conducted:

Limits of generalizability. The queries used in our study drew on Reddit posts, which reflect the narratives and needs of a specific group of people active on the platform. This focus may not capture the full spectrum of perspectives and needs of those who do not engage with Reddit. Additionally, the study is centered on the U.S., addressing users from a particular cultural and geographical background. The domains we found, *i.e.*, relocation, opening a business, and tourism, as well as the constraints we employed, may not apply for other cultures. Moreover, since the experiment was conducted solely in English and not applied to other languages, the choices of platform, language, and cultural context limit the extent to which our findings can be generalized to more diverse populations.

Limitations of recommendations. A potential limitation of our experiment lies in its focus on requesting recommendations, which may conflict with our goal of using the tool for broader information exploration. Recommendations inherently imply a ranking based on certain metrics of relative value, rather than offering a comprehensive set of options. This framing can limit the inclusiveness of responses.

On the other hand, it is unclear whether the criteria used in these responses is equally relevant to all stakeholders—what may be considered ideal for one group might not hold the same value for another. Therefore, in a limited context setting, even if the recommendations favor certain places, the expectation remains that the tool should also include options that cater to the needs and preferences of a more diverse range of users. However, many of these models do not currently employ personalization, nor do they attempt to learn about the specific needs of the user during the interaction. As a result, the recommendations may not be as tailored as users might assume.

Evaluating inclusivity. Our study assesses inclusivity by examining the representation of various demographic groups. However, the specific set of demographic attributes evaluated here is neither

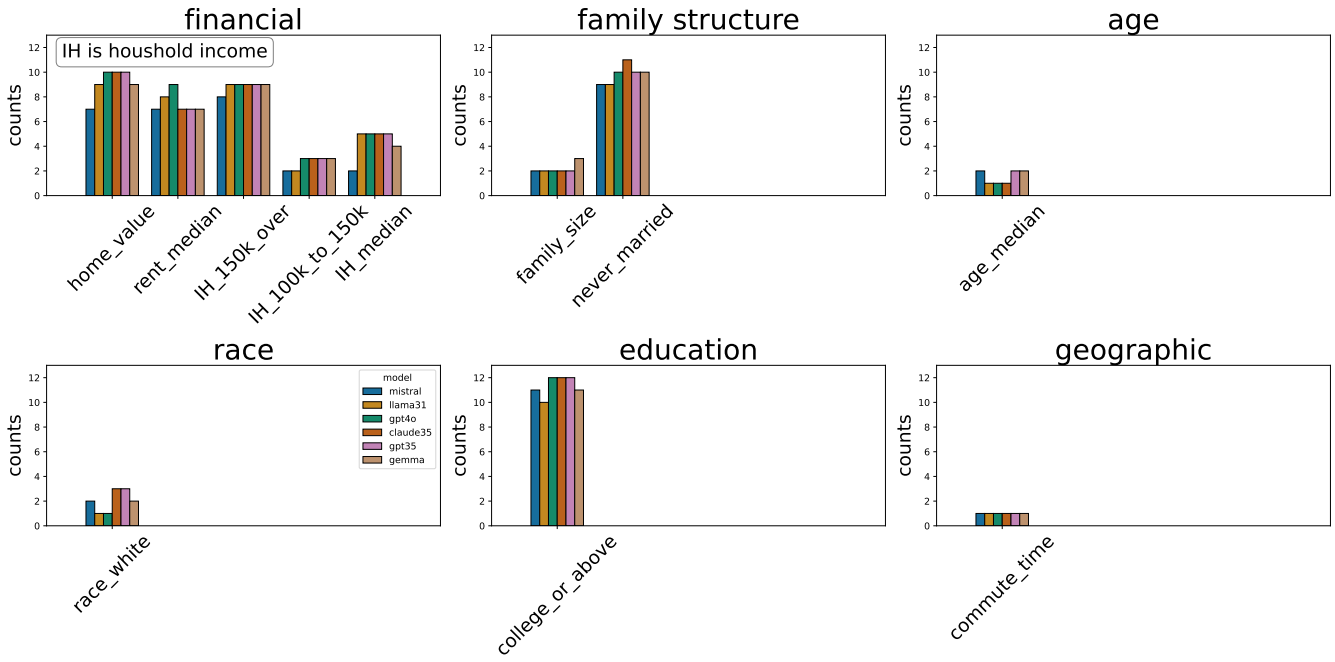


Figure 8: Demographic *category* comparisons between LLM responses to *database distributions* for 'larger' demographic attributes. The x-axes corresponds to *demographic attributes* that were evaluated under a certain *demographic category*. Each *demographic attributes* indicates a respective value of an LLM system. These values correspond to counts (indicated on y-axes) from a total of 12 distributional comparisons per each LLM (as there are 12 queries). For instance, across 12 comparisons on the education category, about 10-12 comparisons over-represented cities that have more individuals with college education or higher, across all LLMs.

exhaustive nor complete. There are likely additional demographic factors we could have included, and we recognize that our ability to assess inclusivity was limited to the demographic data available. Inclusivity, however, should be understood more broadly, encompassing the diverse life experiences of individuals—something that cannot be fully captured through demographics alone or the common attributes found in demographic datasets.

8 Conclusions

In this research, we audited LLMs to investigate patterns in their responses. The audit focused on analyzing the distribution of LLM outputs to uncover insights regarding response patterns, the extent of similarity between different LLMs, and how these similarities may result in the exclusion of certain groups or entities. As LLMs are increasingly used for information-seeking, the content generated—or omitted—can profoundly impact how individuals make decisions based on their recommendations. This is particularly critical in the context of recommendations related to cities or locations for relocation, tourism, or business ventures, where the inclusion or exclusion of certain cities or towns may carry significant economic, cultural, and political consequences for the communities. Our results demonstrate that LLM-generated responses may not adequately cater to certain demographics and that various LLMs display similar biases in this regard. Identifying these gaps is an essential first step in providing a foundation for efforts to make these systems more inclusive.

Acknowledgments

This work was supported by Notre Dame–IBM Technology Ethics Lab award. This work was supported by the National Science Foundation (NSF) under Grant No. CMMI-2326378. The authors would like to thank Tomo Lazovich for their contribution for the project.

References

[1] Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412* (2024).

[2] Ahmed Agiza, Mohamed Mostagir, and Sherief Reda. 2024. Analyzing the Impact of Data Selection and Fine-Tuning on Economic and Political Biases in LLMs. *arXiv preprint arXiv:2404.08699* (2024).

[3] Ibrahim Ahmad, Shiran Dudy, Resmi Ramachandranpillai, and Kenneth Church. 2024. Are Generative Language Models Multicultural? A Study on Hausa Culture and Emotions using ChatGPT. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, Vinodkumar Prabhakaran, Sunipa Dev, Luciana Benotti, Daniel Hershcovich, Laura Cabello, Yong Cao, Ife Adebara, and Li Zhou (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 98–106. doi:10.18653/v1/2024.c3nlp-1.8

[4] Ibrahim Said Ahmad, Shiran Dudy, Resmi Ramachandranpillai, and Kenneth Church. 2024. Are Generative Language Models Multicultural? A Study on Hausa Culture and Emotions using ChatGPT. *arXiv preprint arXiv:2406.19504* (2024).

[5] Shahriar Akter, Grace McCarthy, Shahriar Sajib, Katina Michael, Yogesh K Dwivedi, John D'Ambra, and Kathy Ning Shen. 2021. Algorithmic bias in data-driven innovation in the age of AI. 102387 pages.

[6] Muhammad Ali, Swetasudha Panda, Qinlan Shen, Michael Wick, and Ari Kobren. 2024. Understanding the Interplay of Scale, Data, and Bias in Language Models: A Case Study with BERT. arXiv:2407.21058 [cs.CL] https://arxiv.org/abs/2407.21058

[7] Jiafu An, Difang Huang, Chen Lin, and Mingzhu Tai. 2024. Measuring Gender and Racial Biases in Large Language Models. *arXiv preprint arXiv:2403.15281* (2024).

[8] Anthropic. 2024. *Claude-3.5: 'anthropic/claude-3.5-sonnet:beta'*. https://www.anthropic.com [Language Model].

[9] Raghav Awasthi, Shreya Mishra, Dwarikanath Mahapatra, Ashish Khanna, Kamal Maheshwari, Jacek Cywinski, Frank Papay, and Piyush Mathur. 2023. HumanELY: Human evaluation of LLM yield, using a novel web-based evaluation tool. *medRxiv* (2023), 2023–12.

[10] Ansar Aynetdinov and Alan Akbik. 2024. SemScore: Automated Evaluation of Instruction-Tuned LLMs based on Semantic Textual Similarity. *arXiv preprint arXiv:2401.17072* (2024).

[11] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).

[12] 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 *Conference on fairness, accountability, and transparency*. ACM, 610–623.

[13] Abeba Birhane. 2022. Automating ambiguity: Challenges and pitfalls of artificial intelligence. *arXiv preprint arXiv:2206.04179* (2022).

[14] Rishi Bommasani, Kathleen A Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. 2022. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? *Advances in Neural Information Processing Systems* 35 (2022), 3663–3678.

[15] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669* (2024).

[16] Shiqi Chen, Siyang Gao, and Junxian He. 2023. Evaluating factual consistency of summaries with large language models. *arXiv preprint arXiv:2305.14069* (2023).

[17] Pedro Conceição and Pedro Ferreira. 2000. The young person's guide to the Theil index: Suggesting intuitive interpretations and exploring analytical applications. UTIP working paper.

[18] David Dranove, David Besanko, Mark Shanley, and Scott Schaefer. 2017. *Economics of strategy*. John Wiley & Sons.

[19] Yucong Duan, Fuliang Tang, Kunguang Wu, Zhendong Guo, Shuaishuai Huang, Yingtian Mei, Yuxing Wang, Zeyu Yang, and Shiming Gong. 2023. Ranking of large language model (llm) regional bias. (2023).

[20] Shiran Dudy, Ibrahim Said Ahmad, Ryoko Kitajima, and Agata Lapedriza. 2024. Analyzing Cultural Representations of Emotions in LLMs through Mixed Emotion Survey. *arXiv preprint arXiv:2408.02143* (2024).

[21] Shiran Dudy and Steven Bedrick. 2020. Are Some Words Worth More than Others?. In *First Workshop on Evaluation and Comparison of NLP Systems*. ACL, 131–142.

[22] Marcos Fernández-Pichela, Juan C Pichela, and David E Losadaa. 2024a. Search Engines, Large Language Models or Both? Evaluating Information Seeking Strategies for Answering Health Questions. arXiv:2407.12468v2.

[23] Luciano Floridi and Josh Cowls. 2022. A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design* (2022), 535–545.

[24] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166* (2023).

[25] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics* (2024), 1–79.

[26] Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554* (2023).

[27] Simret Araya Gebreegziabher, Yukun Yang, Elena L Glassman, and Toby Jia-Jun Li. 2024. Supporting Co-Adaptive Machine Teaching through Human Concept Learning and Cognitive Theories. *arXiv preprint arXiv:2409.16561* (2024).

[28] Google. 2024. *Gemma: 'google/gemma-7b-it'*. https://ai.google.dev/gemma/docs [Language Model].

[29] Ziniu Hu, Ahmet Iscen, Chen Sun, Kai-Wei Chang, Yizhou Sun, David A Ross, Cordelia Schmid, and Alireza Fathi. 2023. AVIS: autonomous visual information seeking with large language model agent. In *37th International Conference on Neural Information Processing Systems*. 867–878.

[30] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987* (2024).

[31] Isaac L Johnson, Yilun Lin, Toby Jia-Jun Li, Andrew Hall, Aaron Halfaker, Johannes Schöning, and Brent Hecht. 2016. Not at home on the range: Peer production and the urban/rural divide. In *CHI conference on Human Factors in Computing Systems*. ACM, 13–25.

[32] Mahammed Kamruzzaman, Hieu Minh Nguyen, and Gene Louis Kim. 2024. " Global is Good, Local is Bad?": Understanding Brand Bias in LLMs. *arXiv preprint arXiv:2406.13997* (2024).

[33] Shivani Kapania, Ruiyi Wang, Toby Jia-Jun Li, Tianshi Li, and Hong Shen. 2024. " I'm categorizing LLM as a productivity tool": Examining ethics of LLM use in HCI research practices. *arXiv preprint arXiv:2403.19876* (2024).

[34] Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *arXiv preprint arXiv:2406.14805* (2024).

[35] Jhanvee Khola, Shrujal Bansal, Khushi Punia, Rishika Pal, and Rahul Sachdeva. 2024. Comparative Analysis of Bias in LLMs through Indian Lenses. In *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*. IEEE, 1–6.

[36] Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. " I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *ACM Conference on Fairness, Accountability, and Transparency*. ACM, 822–835.

[37] Harsh Kumar, Mohi Reza, Jeb Mitchell, Ilya Musabirov, Lisa Zhang, and Michael Liut. 2024. Understanding Help-Seeking Behavior of Students Using LLMs vs. Web Search for Writing SQL Queries. *arXiv e-prints* (2024), arXiv–2408.

[38] Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics* 10 (2022), 163–177.

[39] Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. MEDIQ: Question-Asking LLMs for Adaptive and Reliable Medical Reasoning. *arXiv preprint arXiv:2406.00922* (2024).

[40] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. ACL, 74–81.

[41] Siyang Liu, Trish Maturi, Bowen Yi, Siqi Shen, and Rada Mihalcea. 2024. The Generation Gap: Exploring Age Bias in the Value Systems of Large Language Models. arXiv:2404.08760 [cs.CL] https://arxiv.org/abs/2404.08760

[42] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634* (2023).

[43] Valerio Lorini, Javier Rando, Diego Saez-Trumper, and Carlos Castillo. 2020. Uneven coverage of natural disasters in Wikipedia: The case of flood. *arXiv preprint arXiv:2001.08810* (2020).

[44] Yiwei Luo, Kristina Gligorić, and Dan Jurafsky. 2024. Othering and Low Status Framing of Immigrant Cuisines in US Restaurant Reviews and Large Language Models. In *International AAAI Conference on Web and Social Media*, Vol. 18. AAAI, 985–998.

[45] Meta AI. 2024. *LLaMA-3.1: 'meta-llama/llama-3.1-405b-instruct'*. https://ai.meta.com/llama [Language Model].

[46] What Is Data Mining. 2006. *Introduction to data mining*. Springer.

[47] Mistral AI. 2024. *Mistral: 'mistralai/mistral-nemo'*. https://www.mistral.ai [Language Model].

[48] Mazda Moayeri, Elham Tabassi, and Soheil Feizi. 2024. WorldBench: Quantifying Geographic Disparities in LLM Factual Recall. In *Conference on Fairness, Accountability, and Transparency*. ACM, 1211–1228.

[49] Jann Railey Montalan, Jian Gang Ngui, Wei Qi Leong, Yosephine Susanto, Hamsawardhini Rengarajan, William Chandra Tjhi, and Alham Fikri Aji. 2024. Kalahi: A handcrafted, grassroots cultural LLM evaluation suite for Filipino. *arXiv preprint arXiv:2409.15380* (2024).

[50] Allan H Murphy. 1996. The Finley affair: A signal event in the history of forecast verification. *Weather and forecasting* 11, 1 (1996), 3–20.

[51] OpenAI. 2023. *GPT-4o: 'openai/gpt-3.5-turbo'*. https://www.openai.com [Language Model].

[52] OpenAI. 2023. *GPT-4o: 'openai/gpt-4o'*. https://www.openai.com [Language Model].

[53] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *40th annual meeting of the Association for Computational Linguistics*. ACL, 311–318.

[54] Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. 2024. LLM Targeted Underperformance Disproportionately Impacts Vulnerable Users. *arXiv e-prints* (2024), arXiv–2406.

[55] Raphael Poulain, Hamed Fayyaz, and Rahmatollah Beheshti. 2024. Bias patterns in the application of LLMs for clinical decision support: A comprehensive study. *arXiv preprint arXiv:2404.15149* (2024).

[56] Faisal Rahutomo, Teruaki Kitasuka, Masayoshi Aritsugi, et al. 2012. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology*. University of Seoul, South Korea.

[57] Anand Rajaraman. 2011. Mining of massive datasets.

[58] Rajesh Ranjan, Shailja Gupta, and Surya Narayan Singh. 2024. A Comprehensive Survey of Bias in LLMs: Current Landscape and Future Directions. *arXiv preprint arXiv:2409.16430* (2024).

[59] Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. *arXiv preprint arXiv:2104.06683* (2021).

[60] Abel Salinas, Parth Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. 2023. The Unequal Opportunities of Large Language Models: Examining Demographic Biases in Job Recommendations by ChatGPT and LLaMA. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Boston, MA, USA) *(EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 34, 15 pages. doi:10.1145/3617694.3623257

[61] Abel Salinas, Parth Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. 2023. The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama. In *3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, 1–15.

[62] Shreya Shankar, JD Zamfirescu-Pereira, Björn Hartmann, Aditya G Parameswaran, and Ian Arawjo. 2024. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. *arXiv preprint arXiv:2404.12272* (2024).

[63] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1033, 17 pages. doi:10.1145/3613904.3642459

[64] Annalisa Szymanski, Simret Araya Gebreegziabher, Oghenemaro Anuyah, Ronald A Metoyer, and Toby Jia-Jun Li. 2024. Comparing Criteria Development Across Domain Experts, Lay Users, and Models in Large Language Model Evaluation. *arXiv preprint arXiv:2410.02054* (2024).

[65] Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2024. Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks. *arXiv preprint arXiv:2410.20266* (2024).

[66] Shannon Vallor. 2024. *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*. Oxford University Press.

[67] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM collaborative annotation through effective verification of LLM labels. In *CHI Conference on Human Factors in Computing Systems*. ACM, 1–21.

[68] Yuanchun Wang, Jifan Yu, Zijun Yao, Jing Zhang, Yuyang Xie, Shangqing Tu, Yiyang Fu, Youhe Feng, Jinkai Zhang, Jingyao Zhang, et al. 2024. A Solution-based LLM API-using Methodology for Academic Information Seeking. *arXiv e-prints* (2024), arXiv–2405.

[69] Youfu Yan, Yu Hou, Yongkang Xiao, Rui Zhang, and Qianwen Wang. 2024. KNOWNET: Guided Health Information Seeking from LLMs via Knowledge Graph Integration. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[70] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739* (2023).

[71] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[72] Yuhang Zhou, Yuchen Ni, Xiang Liu, Jian Zhang, Sen Liu, Guangnan Ye, and Hongfeng Chai. 2024. Are Large Language Models Rational Investors? *arXiv preprint arXiv:2402.12713* (2024).