Information Research - Vol. 31 No. 1 (2026)

Myth, reality, or in between: Unveiling potential geographical biases of ChatGPT

Manjula Wijewickrema

DOI: https://doi.org/10.47989/ir31146885

Abstract

Introduction. This research examines how geographically biased training data influence the nature of content in ChatGPT responses and to assess the potential occurrence of various geographically biased responses from the users' perspective.

Method. ChatGPT was tested with geographically oriented prompts on ninety-eight countries. The responses were analysed for opinions, facts, and neutral directive sentences, as well as their qualitative and quantitative characteristics. A user survey was conducted on identifying potential geographical biases in ChatGPT responses.

Analysis. The Wilcoxon signed-rank test and the Permutation test were employed, in addition to descriptive analysis. R programming language within RStudio were utilised for the data analysis.

Results. Central and Western European countries exhibited more opinion and fact sentences in their responses, respectively. Qualitative responses had greater meaning consistency than quantitative ones. Sentence type depended on qualitative or quantitative nature, not prompt geography. ChatGPT 3.5 was the most used version, with no reports of geographically offensive, racially biased, or religiously biased responses. Views on geographical bias varied by region, though certain trends emerged.

Conclusion. ChatGPT generates responses of similar lengths irrespective to regions. Qualitative responses are generally more consistent or reliable in terms of their meanings. Most users do not perceive geographical biases, though concerns arise in East Asia and South America.

Introduction

Using chatbots as an Artificial Intelligence (AI) solution for personalised information retrieval is a modern trend, built on developments in fields like machine learning and natural language processing, which have made a substantial impact on information technology over the past decades. These developments typically appear as automated dialogue systems and utilise training datasets to deliver personalised information for user requests. Chat Generative Pre-trained Transformer (ChatGPT) is a recent development within the same family of AI technologies, introduced by OpenAI (https://openai.com/). ChatGPT is widely recognised as a prominent example of a chatbot that uses Large Language Models (LLMs) to communicate with humans. Furthermore, it combines supervised and reinforcement learning methods, such as Reinforcement Learning from Human Feedback (RLHF), with LLM to advance accuracy and efficiency (Wu et al., 2023). The ChatGPT generative model is trained on a diverse set of sources including news articles, academic research, websites, and books (Haleem et al., 2022; Ray, 2023; Wu et al., 2023). However, these sources may not include subscription-based content unless provided by the owners (OpenAI, n.d.). ChatGPT can be used for numerous tasks in addition to using as a personalised information assistant. For example, job application assistance, content generation, language learning and translations, coding assistance, drafting messages, research assistance and idea generation (Bin-Hady et al., 2023; Kalla et al., 2023). While this study focuses exclusively on ChatGPT, it acknowledges the existence of several other AI chatbots such as Google Bard, Bing Chat, Perplexity AI, Jasper, and YouChat. These systems represent a growing ecosystem of conversational AI tools, each with distinct design features and data sources.

Along with ChatGPT's multiple strengths, it also presents certain limitations. Some of these drawbacks include a tendency to create inaccurate information, potential bias in its training data, risk of using sensitive data, challenges in handling simultaneous queries, lack of in-depth information, and the need for fine tuning its AI model (Cao, 2023; Lund & Wang, 2023). Among the other limitations, the current research focuses on a potential bias in its training data. Potential biases in responses generated by ChatGPT to geographically oriented queries have not been thoroughly examined. This study seeks to address this problem by exploring the existence of such biases. Previous studies have emphasised the negative impact of ChatGPT caused by geographically imbalanced training data (Hosseini & Horbach, 2023; Kim et al., 2024; Ray, 2023), but the lack of assessments on this topic hinders understanding the depth of this issue. The absence of information on a topic confined to a specific geographic region could be a key issue that arises from possible scarcity of information in the training data. Limited or skewed information about some regions and cultures could mislead people in other parts of the world and may also lead to their isolation from the rest. This situation may result in the generation of offensive ideas about cultures due to lack of sufficient information to understand them properly. In fact, the geographically imbalanced training data is a common issue in recommender systems research and AI, often leading to negative impacts on various professions and communities in underrepresented regions (Gómez et al., 2021; Park, 2024). Inability to handle multiple languages equally and the potential unavailability of datasets in less dominant languages could lead ChatGPT to respond less effectively in those languages. Therefore, users can be misled by geographically outdated and detrimental information. Finally, this biased information could damage users' trust, because of providing culturally, racially, or religiously insensitive and insignificant information.

Therefore, the purpose of this research was to determine whether potential geographical biases exist in the responses generated by ChatGPT and, to inform ChatGPT users about the regions affected while examining the nature and extent of these biases. To achieve this, the study employed two distinct strategies: a response analysis for a set of geographically oriented test prompts and a user survey. The response analysis evaluated the characteristics of responses generated for the geographically oriented test prompts. The ChatGPT user survey took a different approach, directly exploring users' practical experience with ChatGPT, with a focus on potential geographical biases

and the nature of these biases, if they exist. Consequently, a holistic approach from both system performance and users' experience was expected to examine these latent issues.

An adequate understanding of regions with insufficient or inaccurate information could assist users determine how carefully they should evaluate the retrieved information. Identifying these regions will be a significant contribution of this research for encouraging users to evaluate ChatGPT's responses rationally. From developer's point of view, a comprehensive knowledge on the nature of the geographical information issues would support them to advance the system further.

Geographical bias in AI

Bias in AI refers to an unfair inclination or prejudice embedded in algorithmic decision-making, which may favour or disfavour certain individuals or groups (Fenwick & Molnar, 2022). In practical terms, bias manifests as systematic differences in how AI systems treat particular events, places, people, or objects. Within this framework, geographical bias is a specific form of bias in which such disparities arise in responses based on the geographic orientation of queries. This can lead to AI systems producing inconsistent, inaccurate, or culturally insensitive responses across different regions, even when identical inputs are provided.

AI bias is commonly categorised into three types: pre-existing bias, technical bias, and emergent bias (Friedman et al., 2013; Rana et al., 2023). Geographical bias can result from any or all of these types. For instance, pre-existing bias may stem from training datasets that overrepresented certain regions while underrepresenting others; technical bias can arise when model architectures are optimised primarily for dominant geographies; and emergent bias may appear as users in different regions engage with the model in varied ways over time. A complementary classification highlights three sources of AI bias: Data, Design, and Human-AI Interaction (AlMakinah et al., 2024). Geographical bias rooted in data occurs when training datasets are unbalanced across global regions, resulting in skewed outputs for some locations. In the design stage, insufficient localisation with linguistic, cultural, or contextual aspects can introduce additional disparities. Furthermore, in human-AI interaction, user behaviour and feedback from underrepresented regions may reinforce regional biases, further shaping responses in unintended ways. LLMs like ChatGPT may inherently replicate patterns found in their training data, including any regional inequalities present therein. This can lead to systematic differences in response quality, completeness, or relevance across different geographies. Therefore, geographical bias in AI should be understood as a multifaceted issue involving structural inequalities, design limitations, and interaction dynamics. However, to maintain a manageable research scope, the present study focuses specifically on potential biases stemming from the training data of ChatGPT.

While detecting potential biases generated by AI systems, it is equally important to examine existing AI bias management strategies. Rana et al. (2023) proposed an agile framework to address bias in AI systems, incorporating key components such as data collection and pre-processing, algorithmic transparency and explainability, evaluation metrics, regular audits and monitoring, stakeholder inclusivity, and ethical guidelines and governance. The focus on biases in training data is directly connected to the data collection and pre-processing component of this framework. It also underscores the importance of careful pre-processing to identify and address data-related biases in order to prevent unfair outcomes.

Recent advancements have introduced several innovative approaches for bias detection. For instance, Tellez et al. (2023) developed self-diagnostic deep learning models capable of detecting underlying defects and biases within themselves. Sallami and Aïmeur (2024) introduced FairFrame, a novel framework designed to detect and mitigate bias in textual data. FairFrame employs LLMs and leverages sophisticated few-shot prompting techniques for bias mitigation, representing a pioneering use of LLMs in this domain. Similarly, Iqbal and Ismail (2025) proposed a generalisable

AI-based framework for bias detection applicable across domains, with a detailed methodology for evaluating the influence of gender bias on machine learning model outcomes. Moreover, broader discussions around bias detection and mitigation strategies in machine learning models and LLMs are found in recent literature (Agarwal et al., 2023; Chen et al., 2024; Ferrara, 2023; Gomez & Benavides, 2024; Katare et al., 2022; Mergen et al., 2025; Moon & Ahn, 2025), which collectively emphasise the need for robust and adaptive frameworks to ensure fairness and accountability in AI systems.

Literature review

The purpose of this review is to closely study the literature that examines the biases in ChatGPT, including political, racial, geographic, gender, and religious biases, and to understand the contexts and extent of their influence. Specifically, the review aims to explore what has been revealed about geographic biases and from which perspectives these biases have been investigated. This includes research published in journal articles and conference proceedings between 2021 and 2024, focusing on recent studies relevant to the topic. The literature review section is primarily organised around empirically studied biases in ChatGPT responses, grouped into thematic categories such as gender, political orientation, religion, ethnicity, and most centrally, geography. This thematic structure was adopted to provide a broader context for understanding potential biases in ChatGPT beyond geographically oriented queries. By presenting findings from a range of bias categories, the review offers a comprehensive backdrop to situate the focus of the study on geographical bias. Furthermore, the literature on geographical bias extends beyond comparisons between countries or global regions; it also addresses disparities within counties of a single country. In addition, geography is considered not only in terms of physical location, but also as classified by economic levels, population density, and other contextual factors.

An empirical study to reveal the capabilities and limitations of ChatGPT, with a particular focus on environmental justice issues in the United States, found that ChatGPT was unable to provide information on the topic inquired about for 83 percent of the counties in the selected sample (Kim et al., 2024). The counties with lower population densities, a higher proportion of white population, and lower income levels did not receive localised information from ChatGPT. This finding highlighted the potential for regional disparities in ChatGPT's training data, even within a single country. These disparities indicated that certain counties, including regions that are less urbanized or economically underprivileged, may be underrepresented in the data used to train the model. Although research specifically focusing on the differences in output of chatbots, based on geographical-orientated prompts is rare, one can find occasional discussions with an attention of the abundance of information in training sets. These discussions point toward the growing recognition of such differences in AI-driven technologies. For example, Deldjoo (2024) highlighted the potential biases in ChatGPT-based recommender systems due to their over-representation of content from some geographical regions. This over-representation could have skewed the responses provided, leading to a lack of diversity in recommendations and an underrepresentation of content from some geographies. Kim and Lee (2023) found that ChatGPT provide relatively longer responses to four test prompts related to transport issues and solutions in the United States compared to the responses given for the same prompts in Canada. The researchers attributed this difference to potential geographic biases in the training data. The same study demonstrated that the word counts of responses across the five occasions for each prompt were approximately similar in each occasion for all test cases. This research shares similarities with the current study, as both involved running the same test prompts on multiple occasions and analysing the similarity of responses in terms of meaning and length.

Ray (2023) has discussed numerous limitations of ChatGPT revealed by the existing literature in the field. This review emphasised the negative aspects of gender and racial, ideological, sensationalism, exclusionary, commercial, cognitive, attention, source, novelty, authority, recency,

availability, and hindsight biases in addition to cultural and linguistic bias of ChatGPT. Further, this study has explained this latter type of ChatGPT bias as a result of predominantly trained data from the Internet.

There are studies aimed at investigating biases in ChatGPT, often focus on specific subject domains. Rozado (2023) tested ChatGPT for its potential bias in politics. The researcher has administered 15 different politically oriented tests and almost all tests showed a preference of ChatGPT for left-leaning viewpoints. However, ChatGPT had not shown any political preference and remained with neutral answers for the explicit questions raised about ChatGPT's own political viewpoint. Research conducted by Motoki et al. (2023) had submitted queries to ChatGPT related to Political Science, while minimising concerns about the randomness of the generated text. To address the potential issues of randomness, the researchers used ChatGPT to answer the same queries 100 times, with the query order randomised in each round of response collection. The findings revealed a significant and systematic political favouritism of ChatGPT for the Labour party in the United Kingdom, Democrats in the United States, and the President Lula da Silva in Brazil. Further, Motoki et al. (2023) argued that political bias in LLMs might be difficult to detect and challenging to eradicate from the model compared to eliminating gender or racial bias.

Authors including Gross (2023) and Kaplan et al. (2024) have explored gender bias in ChatGPT. The potential gender bias of ChatGPT was studied by Gross (2023) and reported a heavy gender bias which could have led to undesirable effects. Experimental evidence for gender bias in recommendation letters generated by ChatGPT had been presented by Kaplan et al. (2024). The study compared recommendation letters written using distinct test prompts based on popular male and female names in the United States. Consequently, the researchers noticed significant differences in language between letters generated for male and female names across the prompts. This revealed the potential for AI systems like ChatGPT to reproduce many of the gender-based biases that had been identified in studies of human-written reference letters. Furthermore, this research served as an ideal example of how biased training data, like recommendation letters, could have directed to similar biases being reproduced in ChatGPT's responses.

Moreover, several studies have explored ChatGPT's potential biases in a range of practical contexts. The potential for religious bias with a focus on anti-Muslim topics, had been discussed by Abid et al. (2021). This was examined after analysing the performance of GPT-3 in areas such as prompt completion, analogical reasoning and story generation. Lippens (2024) examined systemic bias in ChatGPT by evaluating a collection of curricula vitae for fake profiles. The study demonstrated how ethnic and gender identity influenced the evaluations made by the chatbot. ChatGPT versions 3.5 and 4 had been tested for their potential cognitive bias (i.e. deviation from generating rational ideas) through human evaluation and linguistic comparison (Castello et al., 2024). Accordingly, both 3.5 and 4 versions exhibited cognitive bias, and the answers generated by these two versions deviated from human-like responses. However, version 4 displayed a slight improvement in performance over version 3.5, though it was not significant. Duncan and Mcculloh (2023) investigated the potential for biased information produced by ChatGPT version 4 using public data from media sources. The results indicated a clear tendency of biased responses. A qualitative analysis conducted by Kocoń et al. (2023) revealed a potential bias of ChatGPT due to the rules assigned on human trainers by OpenAI. The presence of various types of biases in ChatGPT has been mentioned in a number of research publications within the state-of-the-art in AI (Afjal, 2023; Dwivedi et al., 2023; Hosseini and Horbach, 2023; Liu et al., 2023; Tan Yip Ming et al., 2023).

This literature review highlights that while studies by Kim et al. (2024), Deldjoo (2024), and Kim and Lee (2023) identified geographical biases in ChatGPT across various disciplines, Rozado (2023) and Motoki et al. (2023) focused on political bias. Additionally, Gross (2023) and Kaplan et al. (2024) examined gender bias, whereas Abid et al. (2021) investigated religious bias. Lippens (2024)

addressed both ethnic and gender bias, and Castello et al. (2024) explored cognitive bias in ChatGPT. The review further demonstrates that while the types and nature of biases in the training datasets of LLMs, including ChatGPT, have been identified, important gaps remain. In particular, the existing research has largely overlooked the exploration of geographic biases in relation to the context of executed queries and the features of responses received. For instance, both the qualitative and quantitative natures of the test prompts, as well as the behaviour of the response sentences can be considered. Moreover, no prior study has thoroughly explored the prevalence of geographic biases or sought to understand user perceptions of geographic biases in ChatGPT.

Specific objectives of the research:

- To study the influence of geographical related prompts on the length consistency and nature of sentences in ChatGPT responses.
- To evaluate the meaning consistency and reliability of responses generated by ChatGPT across geographical related prompts and query types such as quantitative and qualitative oriented.
- To examine how the nature of sentences in responses varies across different query types which are oriented quantitatively and qualitatively.
- To study ChatGPT users' opinions on its general usability and the potential presence of geographical biases.

Methods

Test prompts and reponses

The current research constructed seven test prompts that queried geographically oriented information of ninety-eight countries (see Table 1). Within each region, countries were chosen randomly from an alphabetical list of all countries in that region. Each region included five countries, except for North America which officially includes only three countries.

Region	Countries
Central Europe	Poland, Slovakia, Slovenia, Switzerland, Germany
Eastern Europe	Moldova, Georgia, Ukraine, Azerbaijan, Russia
Southern Europe	Italy, Albania, Greece, Malta, Montenegro
Northern Europe	Sweden, Lithuania, Latvia, Estonia, Norway
Western Europe	Belgium, United Kingdom, France, Ireland, Andorra
East Africa	Mauritius, Zambia, Madagascar, Seychelles, Djbouti
Southern Africa	Botswana, Namibia, South Africa, Zimbabwe, Lesotho
West Africa	Ghana, Niger, Cabo Verde, Guinea-Bissau, Mali
North Africa	Tunisia, Morocco, Algeria, Egypt, Libya
Central Africa	Central African Republic, Chad, Congo, Equatorial Guinea, Angola
South Asia	Pakistan, Nepal, Bangladesh, Maldives, Sri Lanka
Central Asia	Tajikistan, Kyrgyzstan, Uzbekistan, Kazakhstan, Turkmenistan
Southeast Asia	Laos, Philippines, Singapore, Cambodia, Brunei
East Asia	Japan, China, South Korea, North Korea, Taiwan
Middle East	Kuwait, Syria, Jordan, Saudi Arabia, Lebanon
Oceania	Micronesia, Kiribati, Australia, New Zealand, Tonga
Caribbean	Jamaica, Saint Kitts and Nevis, Barbados, Dominican Republic, Haiti
South America	Paraguay, Argentina, Guyana, Chile, Brazil
Central America	Nicaragua, El Salvador, Costa Rica, Panama, Belize
North America	United States of America, Mexico, Canada

Table 1. Countries and the regions they belong to.

The research queried the output of seven test prompts from ChatGPT version 3.5. All test prompts used in this study were geographically oriented, meaning they inquired about aspects that are dependent on a country's geographic context. The research focused on analysing responses to two distinct types of queries: quantitative and qualitative. Quantitative queries were defined as those that explicitly request information involving measurable or numerical aspects of a subject. These queries typically seek data-driven responses, often involving statistics, rankings, or other forms of quantifiable evidence. In contrast, qualitative queries were those that inquire into descriptive, interpretive, or conceptual aspects of a subject. Accordingly, quantitative responses were the answers given for the quantitative queries, often including numbers, percentages, or rankings. These responses are typically more objective in tone and structure (OpenAI, 2023; Park et al., 2022). On the other hand, qualitative responses were the answers generated for the qualitative queries that tend to be more elaborative and subjective, often incorporating explanatory language. Test prompts 2, 3, 5, and 7 were designed to elicit quantitative responses, while prompts 1, 4, and 6 were intended to generate qualitative responses. These three qualitative prompts were aligned with specific geographic dimensions: culture (prompt 1), language (prompt 4), and safety (prompt 6). Additionally, prompt 6 was used to examine the nature of more opinion-based responses. Among the quantitative prompts, prompt 2 and 7 targeted aspects of physical geography, prompt 3 addressed economic geography, and prompt 5 focused on environmental aspects.

1. What is the most popular cultural event in <country>?

2. How large is the land area of the capital of <country>?
3. What was the <country>'s GDP in 2021?
4. What are the languages used by the indigenous people in <country>?
5. What is the average annual rainfall in <country>?
6. Do you recommend <country> as a safe country to live?
7. What are the precise geographical coordinates of <country>?

Each test prompt was executed twice in separate, newly initiated ChatGPT sessions to eliminate any influence from conversational context. Both executions were conducted on the same day to avoid the impact of model updates or backend changes. The responses generated by ChatGPT for each test prompt and country, were recorded under the respective region. Each test prompt was executed twice for each country. This selection aimed to examine the differences in responses generated by ChatGPT for the same prompt within the same country. Therefore, a total of 196 responses were available for analysis. Each response was analysed at a sentence level. These sentences were then organised into three major categories: opinions, facts, and neutral directives. This research assumed that each sentence belongs to only one of the three categories. The sentences that did not seem to belong to any of the three categories were grouped under the closest category to which they were most likely to belong. Convenience of organising the sentences was the reason for this selection. The three categories of sentences were interpreted as follows:

> **Opinion**: a view, belief, or judgement that reflects an individual's thoughts or feelings about a particular subject. Opinions are subjective and can vary between individuals. Example: "in my view, summer is the best season of the year." This sentence reflects a personal preference that can vary from person to person, making it a clear opinion.
>
> **Facts**: pieces of information that are objectively verifiable and can be proven true through evidence, observations, or measurements. These are not influenced by personal feelings, interpretations, or biases.
> Example: "water boils at 100 degrees Celsius at sea level." This statement can be objectively verified through scientific measurement, making it a fact.
>
> **Neutral directives**: instructions presented in an objective, impartial, non-emotional, and unbiased form.
> Example: "read the following sentences carefully." This instruction is clear, objective, and free from emotional or biased language, making it a neutral directive.

This study separately recorded the number of sentences in each response that reflected opinions, facts, and neutral directives. In addition, it examined whether the two responses given for the same test prompt contradicted or significantly differed from each other.

Survey

The current research also conducted a global survey with a view to reveal the opinion of ChatGPT users across diverse geographical locations worldwide. However, the sample selected for the study was limited to academics representing different regions of the world. Practical challenges of including diverse professional groups in the study was the reason for this limitation. Three universities from each country listed in Table 1 were randomly selected based on the Quacquarelli Symonds World University Rankings (https://www.topuniversities.com/university-rankings) to locate suitable respondents. All academics from the Computing or Information Technology departments of each university were selected for inclusion in the sample. Academics from both departments were selected when both streams were represented. If the selected universities lacked departments or faculties in Computing or Information Technology, the next randomly

selected universities were considered to find suitable respondents. However, the number of universities from each country was limited to three due to time constraints for conducting the research. Examining the opinions of individuals who use ChatGPT and are knowledgeable about the topic being discussed is important for the current study, which is why academics from disciplines likely to be familiar with the topic were selected. Consequently, there were 5583 potential respondents in the sample.

The email address of each sample member was recorded from their respective university profile pages to send out an email. Then, an online questionnaire was designed in Google Forms to collect responses. The questionnaire primarily focused on users' opinions of potential geographical biases in ChatGPT. The first three questions inquired about the respondent's country of present affiliation, their use of chatbots, and the names of those chatbots. The next two questions specifically focused on ChatGPT, inquiring about the version of ChatGPT they were using and their opinion on its impact. Questions 6 to 18 addressed the presence of biases, regardless of the topic, and potential geographical biases in the responses of ChatGPT. These biases were examined along three dimensions: physical geography, context (e.g. culture, race, religion), and nature (e.g. stereotypes, completeness, ambiguity). An email was sent to each sample member, inviting them to participate in the survey. Of all the emails sent, 65 were bounced as undeliverable due to non-existent addresses. The sample members were given three weeks to respond, with a reminder sent after the third week to request non-respondents to complete the survey.

This study used standard methods to impute missing responses from the survey participants. For the analysis of ordinal-scale responses, each response was first coded with a numerical value (i.e. Strongly Disagree=1, Disagree=2, Neutral=3, Agree=4, Strongly Agree=5). Then, the Random Forest algorithm was applied as the imputation method subject to 10 imputations. This approach was chosen primarily to handle mixed data types with complex relationships, as the study scaled ordinal responses into a numerical scale for data analysis and accounted for potential interactions among responses (Stekhoven & Bühlmann, 2012; Tang & Ishwaran, 2017).

The current research employed the Wilcoxon signed-rank test and the Permutation test as statistical methods, in addition to descriptive analysis, to examine the collected data. Further, R programming language version 4.1.1 within the RStudio integrated development environment version 1.4 were utilised for the data analysis of both survey and sentence level analysis of ChatGPT responses.

Results

Analysis of responses generated by ChatGPT

First, the differences in the average number of sentences with different natures (i.e., opinion, factual, and neutral directive) between the first and second responses were examined separately for each region. These differences were found to vary, but not significantly across most of the world regions. Second, the consistency of the responses generated for the same test prompt was examined by comparing the meaning between the two responses produced. These consistencies were analysed separately for prompts designed to generate quantitative and qualitative responses. The results showed a generally high level of consistency between the two responses for qualitative test prompts. Two of the quantitative test prompts showed notably higher consistency, while the other quantitative test prompts demonstrated significantly lower consistency in responses. To address the third objective, the amount of opinion, factual, and neutral directive sentences in responses was defined using new density variables: opinion-dense, fact-dense, and neutral-dense. Although regional differences in density variables were minimal, notable variations appeared across different prompt types such as quantitative and qualitative. The fact-dense had higher values for quantitative type of responses. Neutral-dense was low for test prompts with quantitative type of responses, while it was higher for prompts with qualitative type of responses. Finally, to

address the fourth specific objective, the results from the ChatGPT user survey were analysed to reveal user perspectives on the existence of potential geographical biases in ChatGPT-generated responses.

This research compared the difference in the average number of opinion, fact, and neutral directive sentences between the first and second responses for each region separately. Given the relatively small number of responses per region—seven responses corresponding to seven prompts—a non-parametric test was selected, in particular the Wilcoxon signed-rank test. Table 2 presents the Wilcoxon signed-rank test results comparing the average number of opinion, fact, and neutral directive sentences between the first and second responses across all regions.

Region	Average opinion sentences		Average fact sentences		Average neutral directive sentences	
	Test statistic	**p-value**	**Test statistic**	**p-value**	**Test statistic**	**p-value**
Central Europe	15	0.04217	10	0.5002	4	0.593
Eastern Europe	2	0.6547	7	0.4652	4	0.593
Southern Europe	1	0.3173	10	0.06789	3	1
Northern Europe	3	0.1797	8	0.2733	2	0.593
Western Europe	5	0.285	15	0.04311	3	1
East Africa	2	0.6547	10	0.06789	5	0.285
Southern Africa	2	0.6547	9	0.1441	3	1
West Africa	3	0.1797	9	0.1408	1	0.285
North Africa	10	0.06789	7	0.4652	2	0.593
Central Africa	8.5	0.1975	17	0.173	5	1
South Asia	6	0.1025	10	0.06789	4	0.593
Central Asia	4	0.5637	10	0.06789	3	1
Southeast Asia	0	0.3173	7	0.4652	4	0.593
East Asia	3.5	0.5807	12	0.2249	3	0.1797
Middle East	6	0.715	12	0.2249	2	0.6547
Oceania	2.5	0.3573	7	0.4652	2	0.593
Caribbean	2	0.6547	1.5	0.1975	0	0.3173
South America	4	0.593	2	0.2733	0	0.1797
Central America	0	0.1797	0	0.06789	3	0.1797
North America	2	0.6547	5	1	3	1

Table 2. Wilcoxon signed-rank test results for significant difference between two responses.

There are no significant differences in the average number of sentences between the two responses for all three categories: opinion, fact, and neutral directive across regions, except in the Central and Western European regions. The average number of opinion and fact sentences differs significantly between the two responses for Central and Western Europe regions respectively. The average number of opinion and fact-based sentences was higher in the first response than in the second response for both the Central and Western European regions.

Test prompts 1, 4 and 6 were identified as addressing qualitative geographical aspects, while test prompts 2, 3, 5 and 7 focused on quantitative aspects. This research assessed the meaning consistency between the two responses generated by the system for each test prompt. Qualitative responses were deemed meaning consistent if they conveyed similar meanings, while quantitative responses were considered meaning consistent if they provided identical values. The meaning consistency between two responses was scored as 1 for consistency and 0 for inconsistency, allowing for a quantifiable measure of meaning consistency. This approach enabled the study to determine the overall meaning consistency of responses across all countries for each test prompt. Consequently, the meaning consistency for the two responses given to qualitative test prompts is apparently high, with 244 consistent instances out of 294. However, the meaning consistency for

the two responses is notably higher for the second and seventh quantitative test prompts, with 184 consistent instances out of 196. In contrast, other quantitative test prompts show considerably lower consistency, with only 44 instances out of 196 occasions.

This research defined the meaning consistency rate of a region as the ratio of meaning consistent response pairs (or instances) to total response pairs (or occasions) for all countries within a geographic region, averaged across seven test prompts. Figure 1 illustrates the variation in meaning consistency rates across different regions.

Accordingly, East Asian, South Asian, and North American regions report the highest meaning consistency rate between the two responses. Conversely, Central European and Caribbean regions report the lowest meaning consistency rate between the two responses.

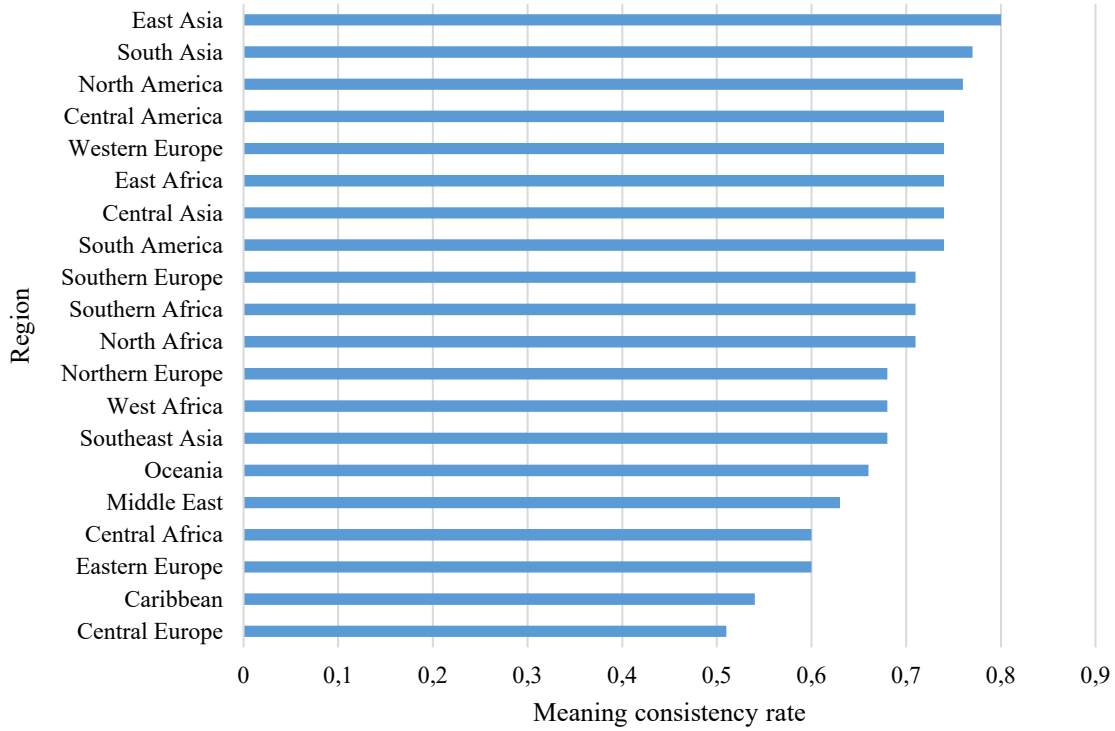


Figure 1. Meaning consistency rate of two responses.

Figure 1 does not show continent-specific bias toward meaning consistency at the middle or lower levels of the meaning consistency rate. However, Asian and American continents display relatively higher meaning consistency rates between the two responses.

Furthermore, the research defined the following measures to compare the density of opinion, fact, and neutral directive sentences across each region.

> Opinion-dense: proportion of opinion sentences to all sentences in a response. A response with high opinion-density is subjective and heavily focused on personal views and beliefs.
> For example, if a response comprises 5 opinion sentences out of a total of 20 sentences, the resulting opinion-density is calculated as 0.25.
>
> Fact-dense: proportion of factual sentences to all sentences in a response. Fact sentences usually provide objective information that can be verified by evidence or

research. A response with a higher density of factual information is highly informative.
For example, if a response comprises 10 factual sentences out of a total of 20 sentences, the resulting fact-density is calculated as 0.5.

Neutral-dense: proportion of neutral directive sentences to all sentences in a response. Neutral directive sentences are usually unbiased and impartial.
For example, if a response comprises 15 neutral directive sentences out of a total of 20 sentences, the resulting neutral-density is calculated as 0.75.

These scores were calculated separately for responses for each test prompt and each country before averaging them for the corresponding region. Finally, the two scores obtained were averaged again, as each case included two responses. The heat maps in figures 2, 3, and 4 compare the average density scores for opinions, facts, and neutral directive sentences, respectively.

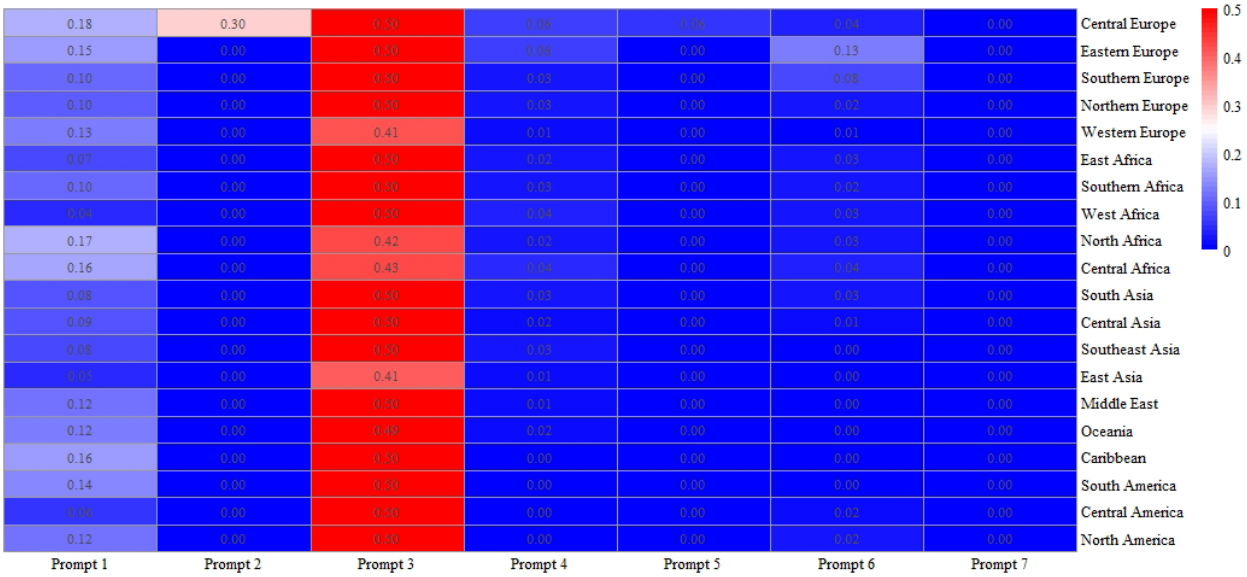


Figure 2. Average opinion-dense for each test prompt (opinion-dense increases from blue to red).

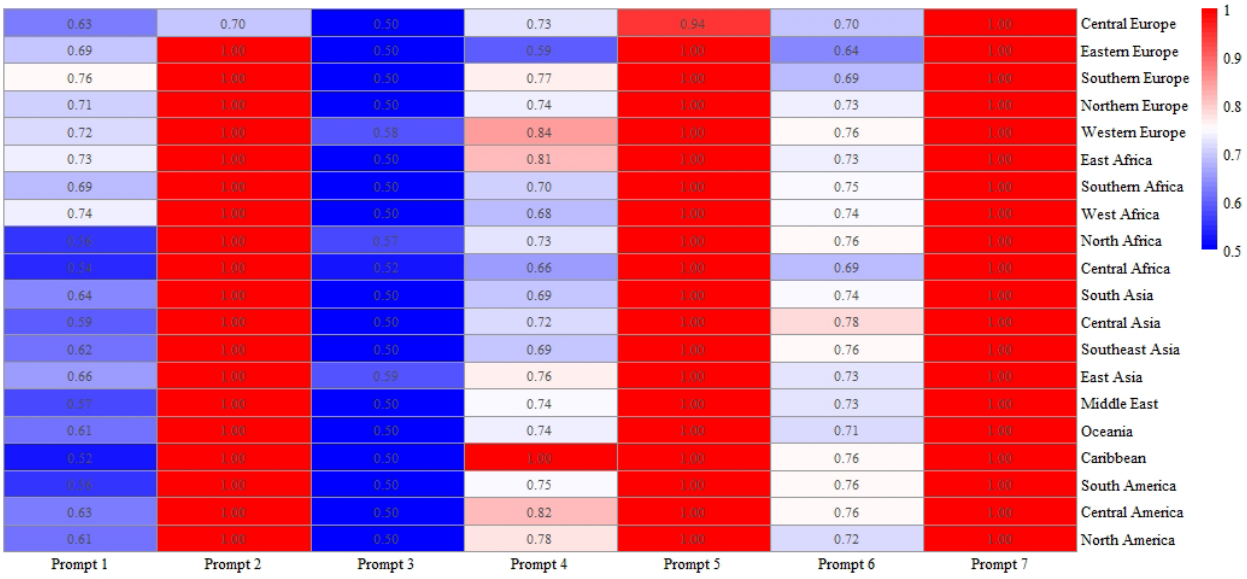


Figure 3. Average fact-dense for each test prompt (fact-dense increases from blue to red).

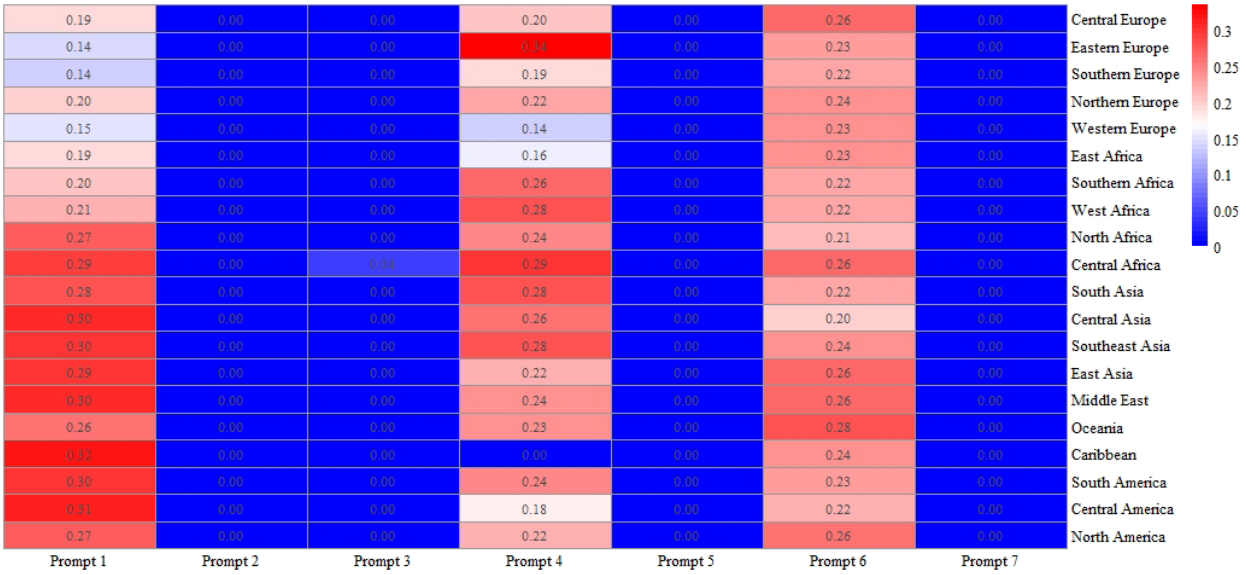


Figure 4. Average neutral-dense for each test prompt (neutral-dense increases from blue to red).

According to figures 2, 3 and 4, fact-dense is relatively higher in all regions for the majority of test prompts compared to that of opinion and neutral dense. Although there is no significant variation in the three average densities across different regions, a notable difference can be observed across different prompts. The fact-dense generally has higher values for prompts 2, 5 and 7, which represent responses of a quantitative nature. In contrast, neutral-dense is low for test prompts that receive quantitative responses, while it is higher for prompts seeking qualitative responses. All Asian and American regions, Central and North African regions, the Middle East, and Oceania show lower fact-dense but higher neutral-dense for prompts expecting more qualitative type responses. Overall, the density of the three sentence types varies more for the fact-based sentences, while the variation is minimal for the opinion-based sentences. Moreover, higher density variations are observed in responses of a qualitative nature, which is common to both fact and neutral sentences.

Survey results

The current research conducted a survey to collect users' opinions on potential geographical biases of ChatGPT. Table 3 shows the percentage of responses received from the online questionnaire (Appendix 1).

Sample information	**Number**
Emails sent	5583
Emails bounced	65
Effective sample	5518
Completed responses	115
Response rate	2.1%

Table 3. Response rate

Responses from countries in regions such as West Africa, North Africa, Central Africa, Central Asia, Caribbean and Central America, which affected the geographical diversity of the responses. North American and South Asian regions reported the highest response rates, while East Asian and South American regions reported low response rates.

Figure 5 depicts the geographical distribution of the number of respondents as ratios across five ordinal usage levels for AI chatbots. Not all regions studied are included in the figure, as six regions did not receive any responses for any of the five usage levels. According to Figure 5, a higher number of respondents use AI chatbots either "sometimes" (37 percent of users from all responded regions) or "very often" (34 percent of users from all responded regions), with these options being considerably popular across all regions except East Asia. In East Asia, an equal number of respondents use chatbots "rarely", "very often", and "always" (33 percent each). There are relatively few users who "never" (2.6 percent of users from all responded regions) or "rarely" (10.7 percent of users from all responded regions) use AI chatbots in each region, indicating high demand for chatbots worldwide. Respondents from Eastern and Southern Europe, East and Southern Africa, Southeast Asia, and South America use AI chatbots at least "sometimes". A considerable number of respondents from East Africa, Southeast Asia, East Asia, and North America use AI chatbots "always" (32.5 percent of users from all these four regions) implying the use of chatbots almost every time they perform a relevant task.

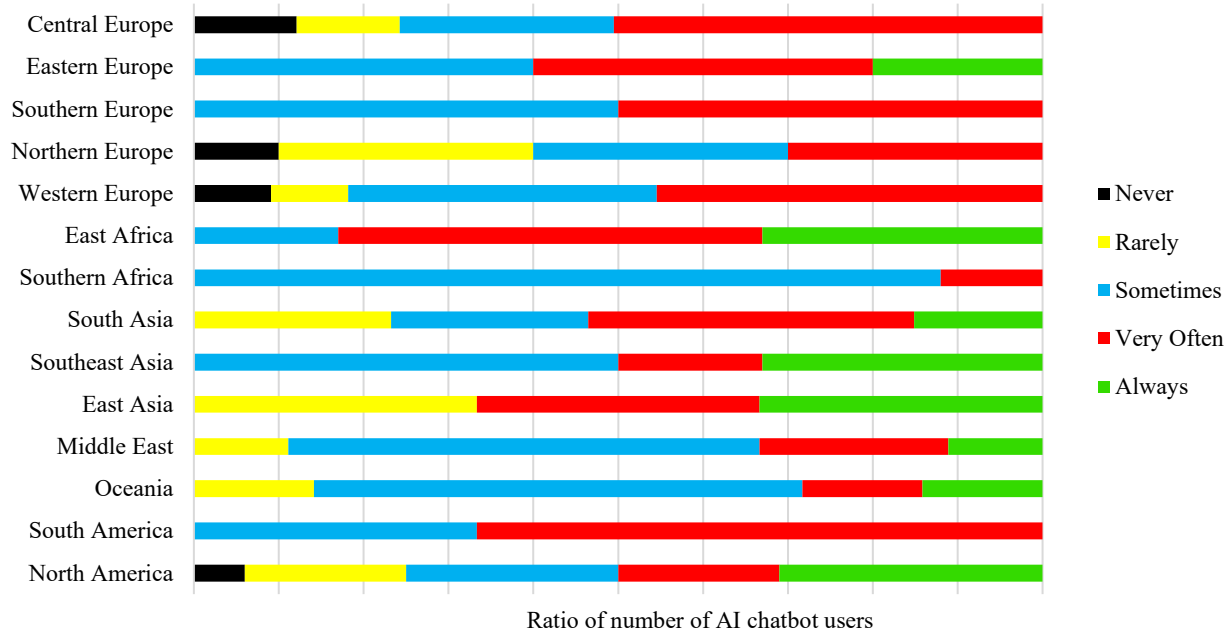


Figure 5. Number of respondents in regions as ratios of their AI chatbot usage levels.

Figure 6 illustrates the average number of respondents from 14 regions based on the preferred chatbots. The figure illustrates that ChatGPT is the most widely used AI chatbot among respondents across all regions. In addition to ChatGPT, Bing Chat is somewhat popular in all regions except in Central and Eastern Europe. Moreover, Google Bard is somewhat popular in all regions except in Central and Eastern Europe, while Perplexity AI is somewhat popular in Southern Africa, East Asia, Middle East, and North America. In contrast, ChatGPT has attracted a relatively higher and approximately similar number of users across all regions. Moreover, the results indicate that ChatGPT holds a dominant position over other AI chatbots in the Central and Eastern European regions.

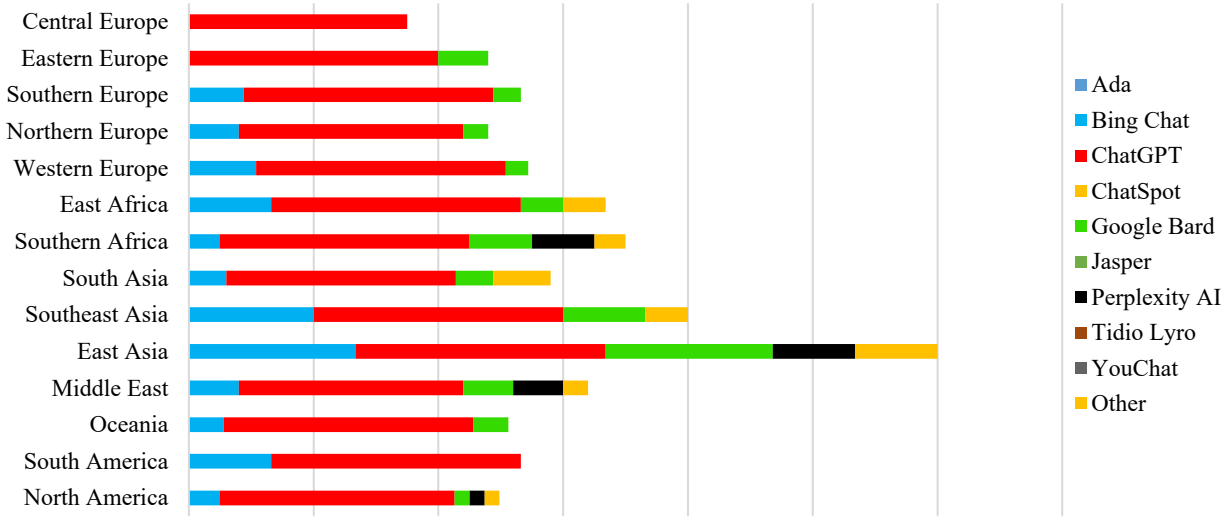


Figure 6. Preferred AI chatbots for average respondents representing 14 regions.

Figure 7 shows the regional distribution of average respondents according to the version of ChatGPT they use. Accordingly, all respondents from East Africa use ChatGPT 3.5 free version. Notably, respondents from Africa, the Middle East and South America show a preference for the free version, whereas there is a higher demand for the advanced, paid version in Europe, North America, and East Asia.

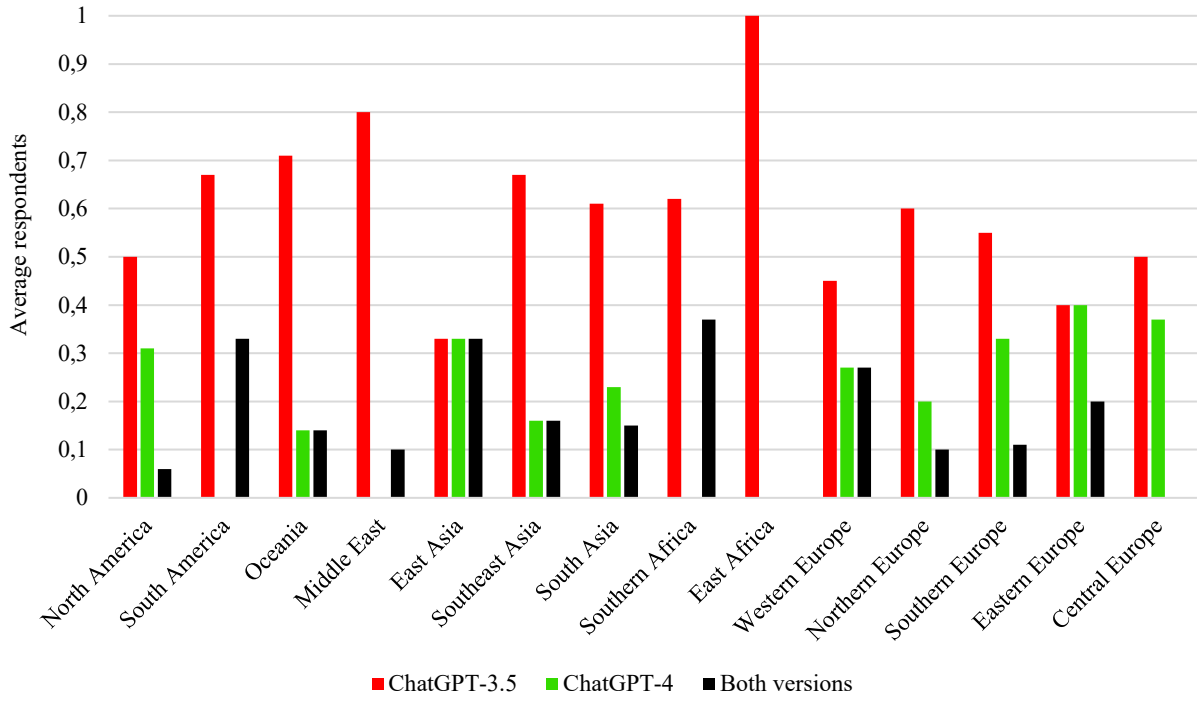


Figure 7. Average respondents for each ChatGPT version across regions.

There were missing responses for some questions in the questionnaire. Specifically, questions 7, 8, 9, 12, 13 and 17 each had 1.8 percent of missing responses, while questions 10, 11, 14, 15, 16, and 18 each had 0.9 percent of missing responses. Therefore, the research employed a missing value imputation method using the Random Forest algorithm to generate values for the missing

responses. The permutation test was then conducted to detect significant differences in responses across different regions. Table 4 gives the results of the permutation test.

Question number	p-value	Test statistic	99% Confidence interval	
5	0.3211933	14.70729	0.3189986	0.3233937
6	0.06088667	20.90943	0.05976743	0.06201986
7	0.1622033	17.6273	0.1604734	0.1639440
8	0.2638033	15.67317	0.2617329	0.2658812
9	0.5979533	11.37515	0.5956443	0.6002593
10	0.73814	9.687954	0.7360670	0.7402054
11	0.43741	13.24669	0.4350764	0.4397455
12	0.9776367	5.141169	0.9769320	0.9783261
13	0.95016	6.147552	0.9491278	0.9511779
14	0.3498	14.39263	0.3475579	0.3520468
15	0.7345233	9.758489	0.7324413	0.7365980
16	0.23968	16.07248	0.2376749	0.2416934
17	0.54774	11.94918	0.5453969	0.5500816
18	0.3018967	15.08138	0.2997392	0.3040604

Table 4. Permutation test results for differences of responses across all regions.

The p-values for all questions are greater than 0.05. Therefore, based on the permutation test, there is no statistically significant difference in responses across all geographical regions for any of the questions from numbers 5 to 18.

The current research also determined the mean response scores, based on numerical coding from 1 to 5, for each question from numbers 5 to 18 (Appendix 1) across each region. These scores are illustrated in the heat map in Figure 8. Accordingly, respondents from East Asia, the Middle East and South America display relatively higher levels of agreement with the statements in questions 5 to 18 compared to respondents from other regions. This reflects a tendency among respondents from these regions to perceive the presence of geographical biases in ChatGPT responses. Furthermore, regardless of the region, question 5 received a relatively high number of positive responses, a pattern not observed for the other questions. This suggests that respondents recognise ChatGPT's significant influence on scholars globally. However, these behaviours can be more accurately understood with the Standard Deviation (SD) values provided in Table 5, which consider all questions collectively across each region.

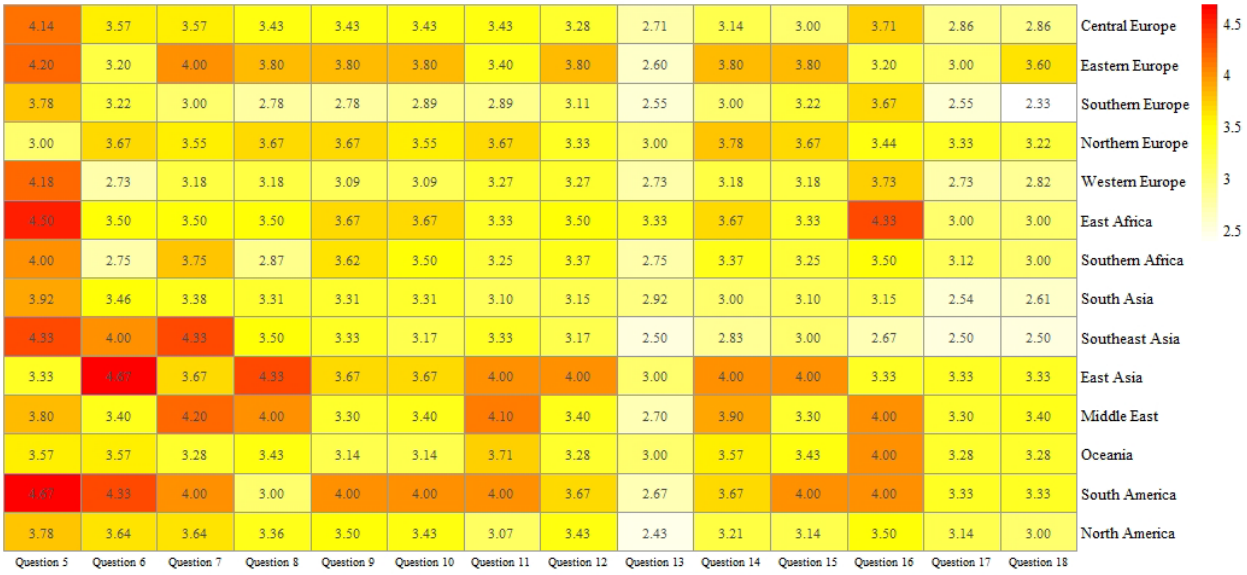


Figure 8. Mean response scores for each question across regions (mean increases from white to red).

Region	Mean for all questions	SD for all questions
Central Europe	3.32	0.917
Eastern Europe	3.57	0.878
Southern Europe	2.98	0.934
Northern Europe	3.47	1.05
Western Europe	3.17	0.831
East Africa	3.56	0.949
Southern Africa	3.29	0.834
South Asia	3.16	1.26
Southeast Asia	3.22	0.949
East Asia	3.74	0.989
Middle East	3.58	0.913
Oceania	3.4	0.872
South America	3.76	0.932
North America	3.3	0.981

Table 5. Mean response scores and standard deviations for all questions across regions.

The South Asian region reports the highest standard deviation for responses received across all questions. This indicates that although the responses tended to cluster around “neutral”, there was a higher relative dispersion, reflecting greater diversity in responses. This observation also holds true for the North European region. In contrast, responses from the Western European and South African regions appear concentrated around the “neutral” response, as indicate by their low standard deviations. The highest mean scores across all questions are reported by the East Asian and South American regions. Additionally, the standard deviations for these two regions are mid-

range compared to those of other regions. This suggests that respondents from these regions are moderately likely to agree with the arguments presented in the questions.

Discussion and Conclusions

This study compared the differences between the first and second responses for each test prompt to assess the influence of geographical regions on response lengths. Only Central and Western European regions reveal statistically significant differences between the two responses, specifically in terms of the average number of opinion and fact-based sentences, respectively. This finding suggests that, for the same query, ChatGPT usually generates responses of similar lengths, maintaining a comparable number of sentences across the three categories: opinions, facts, and neutral directives. However, in Central Europe, the first response is likely to contain significantly more opinion sentences compared to the second response for the same query. A similar trend is observed in Western Europe, where the first response contained a significantly higher number of fact-based sentences (Appendix 2).

The meaning consistency between the first and second responses generated by ChatGPT is higher in qualitative type of responses compared to quantitative type of responses. This may indicate a relatively higher possibility of receiving confident responses of a qualitative nature. However, in addition to the different methods employed by this study to determine meaning consistency between responses in qualitative and quantitative types, the potential abundance of errors in the quantitative training data and the ChatGPT language models' potential capability to handle qualitative data compared to quantitative data may also contribute to the consistency differences observe. These conjectures need to be examined further to obtain concrete conclusions. Responses generated for the queries involving East Asian, South Asian, and North American regions demonstrated a higher level of meaning consistency, indicating more reliable outputs. In contrast, responses related to the Central European and Caribbean regions showed weaker meaning consistency. In other words, ChatGPT responses concerning regions in Asia and the America tend to be more reliable in terms of answer's meaning consistency compared to those involving other parts of the world.

Facts provided in responses to most prompts are relatively more frequent compared to that of opinions and neutral directive sentences in responses to the same prompts. Therefore, the responses generated by ChatGPT are rich in factual information, which supports users in improving their knowledge, as factual information is neither neutral ideas nor opinion-based. Usually, opinion sentences are likely to be biased, as ideas in them are generated by the system itself. Therefore, with its high density of factual information, ChatGPT would be able to provide more reliable, testable, and informative ideas, while controlling geographical biases. Moreover, the number of sentences—reflecting the richness of these factual, opinion, and neutral information in responses— does not vary significantly on the basis of the name of different countries being included in the prompts. In contrast, the difference in the richness of these three types of sentences is notable across different prompt types. The richness of factual sentences is usually higher in responses with a quantitative nature. One potential reason for this could be the influence of user feedback. For instance, users expecting quantitative responses may value the precision of answers more highly than the precision included in qualitative responses, because quantitative responses are often measurable. Consequently, the system is more likely to receive feedback from users, particularly for quantitative responses. These frequent feedbacks received by ChatGPT may reinforce the model, making factual sentences more prominent in quantitative responses. For instance, consider ChatGPT initial response to the query, "What is the literacy rate in Africa?", if similar feedback is submitted by multiple users, the patterns of feedbacks can be aggregated and analysed during future training phases. As a result, the updated versions of the model may include more facts such as country wise literacy rates.

Qualitative responses are rich in neutral directive sentences, though one might question why these responses are not richer in opinions than neutral directives. On the one hand, this could be viewed negatively, as a potential performance issue of the ChatGPT model in generating opinions. On the other hand, it could be seen positively, as a deliberate strategy to limit the number of opinions and maintain a neutral tone, thereby avoiding unnecessary influence on users through its opinions. This trend is particularly notable in Asia, America, Central and North Africa, the Middle East, and Oceania when it comes to qualitative responses. The density of fact-based ideas often varies more compared to the other two sentence types, implying a potential uneven distribution of training data for addressing distinct types of questions.

The survey reveals that respondents' opinions on their use of AI chatbots are generally positive, regardless of their geographical regions, collectively reinforced by wider research (Decoupes et al., 2025; Gondwe, 2023; Liu et al., 2024). This suggests a widespread acceptance and enthusiasm for emerging AI technologies among scholars in Computing and Information Technology. Respondents from Eastern and Southern Europe, East and Southern Africa, Southeast Asia, and South America show a strong preference for using AI chatbots, highlighting their growing significance in these regions.

Among all the AI chatbots considered, ChatGPT has emerged as the most widely used tool by the respondents. Several factors could explain this popularity. For example, the experimentally verified accuracy of the language models employed in ChatGPT (Haltaufderheide and Ranisch, 2024; Kung et al., 2023; Samaan et al., 2023) and its capability to maintain context-awareness during interactions may be a prominent reason. Additionally, its user-friendly accessibility through both web and mobile interfaces, free availability, potential ability to manage high user demands, established trust and recognition, and integration with Application Programming Interfaces (APIs) may further contribute to its widespread adoption. Meanwhile, Google Bard and Bing Chat are also utilised by respondents, albeit to a lesser extent. The free version of ChatGPT is significantly popular among respondents from low-income regions, with the exception of Oceania. In contrast, respondents from high and middle-income regions are more likely to use both free and paid versions of ChatGPT. This trend may reflect the respondents' spending capacity and willingness to invest in resources based on their income levels (Daepp and Counts, 2025; Hassan and Aziz, 2025). Further, respondents acknowledge that ChatGPT has had a substantial impact on scholars.

The permutation test results further demonstrate that there is no statistically significant difference in responses to survey questions 5 through 18 across the geographical regions considered. This indicates a high degree of similarity in the answer options chosen by respondents for each question, regardless of their geographic location. Treating all geographical biases and their influence as equivalent by the users worldwide, may have contributed to this outcome. Nevertheless, respondents across all regions notably reject the notion that ChatGPT exhibits geographically offensive, racially biased or religiously biased tendencies. This reflects a positive perception of ChatGPT's ability to deliver balanced and impartial responses, even on extremely sensitive topics that could impact harmony among diverse nations, races and religions. Therefore, under certain constraints, ChatGPT's training data tends to produce balanced responses regardless of geographic context. This argument is supported by the findings of Georgiou (2025), which demonstrate its ability to generate positive sentiments about all countries.

The results for Mean and Standard Deviation suggest that respondents frequently selected the 'neutral' option for most questions. This may indicate a lack of experience with potential geographical issues of ChatGPT or limited exposure to geographically oriented prompts. However, the relatively large Standard Deviations observed in responses from Northern Europe, South Asia, and North America suggest a diverse range of opinions regarding the potential existence of geographical biases in ChatGPT. In contrast, respondents from East Asia and South America show a stronger tendency to agree with the presence of geographical biases highlighted in the survey,

implying a more critical perspective in these regions. All in all, while there is no significant consensus on the existence of geographical biases highlighted by users from certain regions, except East Asia and South America, notable differences emerge regarding biases associated with certain characteristics. In particular, users express concerns about accuracy, relevance, stereotypes, and language biases of ChatGPT to some extent. These concerns highlight the importance of continued evaluation and improvement of ChatGPT to ensure fair and reliable responses across diverse contexts.

The findings of this research will help users better understand potential geographical and cultural biases in ChatGPT's training data. This awareness allows for more informed use of AI, especially when handling sensitive topics. Developers can use the insights to identify and address biased datasets, improving fairness and trust in AI systems. Information professionals may also guide users in evaluating geographically influenced responses, while social scientists can explore broader societal impacts, supporting the development of more inclusive AI technologies.

One promising avenue for future research is to compare the responses generated by ChatGPT with those from other competing AI chatbots. This comparative analysis could help identify which chatbot performs more effectively across different types of queries, particularly those involving regional, cultural, or linguistic nuances. Such an extension would not only highlight strengths and weaknesses of performance in ChatGPT, but also provide valuable insights into how different AI models handle geographical diversity, ultimately guiding users in selecting the most appropriate chatbot for their specific needs. The accuracy, completeness, and abomination of sentences included in the responses were not assessed in the current research. However, future research could focus on assessing these characteristics, particularly in factual content. In addition to fact ideas, relative opinions could also be evaluated, as they are measurable too. Other types of content may be excluded from evaluation, assuming they have minimal or no influence on the overall meaning of a response.

Limitations

This study focused exclusively on ChatGPT due to several reasons. Expanding to multiple chatbots would have increased complexity and compromised depth. Methodological differences, such as using static or real-time data, varying content filtering, and the differences in response style and structure (Waisberg et al., 2024), make fair comparison difficult. Varying response lengths generated by different systems further complicate evaluation. Survey results also showed strong user preference for ChatGPT, citing its advantages over competitors (Chalyi, 2024; Ray, 2023), ability to handle original queries (Plevris et al., 2023), consistent performance across domains (Lee et al., 2024), and clarity of responses (Raman et al., 2024). This study did not assess geographical bias in terms of accuracy, as many existing studies already address ChatGPT's accuracy (Cao et al., 2023; Cappellani et al., 2024; Chalyi, 2024; Hake et al., 2024; Kuşcu et al., 2023; Samaan et al., 2023). Instead, it focused on meaning consistency, which is less studied (Elazar et al., 2021) but closely linked to accuracy. Consistency reflects response stability across regions and avoids the challenges of defining objective truths for open-ended prompts. Future research could combine both accuracy and consistency using expert evaluation. Due to time and resource limits, this study used a limited number of carefully selected prompts focused on geography-related reasoning. Similar studies have also used limited prompts. For instance, Georgiou (2025) used one, Renshaw et al. (2025) used two, and some others used 15 targeted prompts (Plevris et al., 2023; Rudolph et al., 2023). This shows that prompt quality and relevance, not quantity alone, are key to meaningful analysis. Future research could involve multiple coders to reduce subjective bias and improve classification validity. This also enables inter-coder reliability assessment using measures like Cohen's kappa, Scott's pi, Fleiss' kappa, inter-rater correlation, and concordance correlation coefficient. This study did not account for confounding factors like regional literacy, internet access, or cultural attitudes toward AI, as its focus was on biases linked to ChatGPT's training data.

Including such variables would require a different framework and datasets. However, selecting university-based computing and information technology scholars helped minimize variation in user capability and access, allowing clearer focus on geographic response disparities.

About the author

Manjula Wijewickremas is Senior Assistant Librarian at the Library of Sabaragamuwa University of Sri Lanka, Belihuloya. His main research interests are in information retrieval and scientometrics. He can be contacted at manju@lib.sab.ac.lk

References

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-Muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 298–306). https://doi.org/10.1145/3461702.3462624

Afjal, M. (2025). ChatGPT and the AI revolution: A comprehensive investigation of its multidimensional impact and potential. *Library Hi Tech*, *43*(1), 353–376. https://doi.org/10.1108/LHT-07-2023-0322

Agarwal, R., Bjarnadottir, M., Rhue, L., Dugas, M., Crowley, K., Clark, J., & Gao, G. (2023). Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. *Health Policy and Technology*, *12*(1), 100702. https://doi.org/10.1016/j.hlpt.2022.100702

AlMakinah, R., Goodarzi, M., Tok, B., & Canbaz, M. A. (2024). Mapping artificial intelligence bias: A network-based framework for analysis and mitigation. *AI and Ethics*, *5*, 1995–2014. https://doi.org/10.1007/s43681-024-00609-0

Bin-Hady, W. R. A., Al-Kadi, A., Hazaea, A., & Ali, J. K. M. (2023). Exploring the dimensions of ChatGPT in English language learning: A global perspective. *Library Hi Tech*. https://doi.org/10.1108/LHT-05-2023-0200

Cao, J. J., Kwon, D. H., Ghaziani, T. T., Kwo, P., Tse, G., Kesselman, A., Kamaya, A., & Tse, J. R. (2023). Accuracy of Information Provided by ChatGPT Regarding Liver Cancer Surveillance and Diagnosis. *American Journal of Roentgenology*, *221*(4), 556–559. https://doi.org/10.2214/AJR.23.29493

Cao, X. (2023). A new era of intelligent interaction: Opportunities and challenges brought by ChatGPT (0). *Geographical Research Bulletin*, *2*, 162-165. https://doi.org/10.50908/grb.2.0_162

Cappellani, F., Card, K. R., Shields, C. L., Pulido, J. S., & Haller, J. A. (2024). Reliability and accuracy of artificial intelligence ChatGPT in providing information on ophthalmic diseases and management to patients. *Eye*, *38*(7), 1368–1373. https://doi.org/10.1038/s41433-023-02906-0

Castello, M., Pantana, G., & Torre, I. (2024). Examining cognitive biases in ChatGPT 3.5 and ChatGPT 4 through human evaluation and linguistic comparison. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas, Chicago, USA, September 30-October 2, 2024* (Vol. 1, pp. 250–260). https://aclanthology.org/2024.amta-research.21.pdf

Chalyi, O. (2024). An evaluation of general-purpose AI chatbots: A comprehensive comparative analysis. *InfoScience Trends*, 1(1), 52–66. https://doi.org/10.61186/ist.202401.01.07

Chen, F., Wang, L., Hong, J., Jiang, J., & Zhou, L. (2024). Unmasking bias in artificial intelligence: A systematic review of bias detection and mitigation strategies in electronic health record-based models. *Journal of the American Medical Informatics Association*, 31(5), 1172–1183. https://doi.org/10.1093/jamia/ocae060

Daepp, M. I., & Counts, S. (2025). The emerging generative artificial intelligence divide in the United States. *Proceedings of the International AAAI Conference on Web and Social Media*, 19, 443-456. https://doi.org/10.1609/icwsm.v19i1.35825

Decoupes, R., Interdonato, R., Roche, M., Teisseire, M., & Valentin, S. (2025). Evaluation of Geographical Distortions in Language Models. In D. Pedreschi, A. Monreale, R. Guidotti, R. Pellungrini, & F. Naretto (Eds.), *Discovery science* (Vol. 15243, pp. 86–100). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-78977-9_6

Deldjoo, Y. (2024). Understanding biases in ChatGPT-based recommender systems: Provider fairness, temporal stability, and recency. ACM *Transactions on Recommender Systems*. https://doi.org/10.1145/3690655

Duncan, C., & Mcculloh, I. (2023). Unmasking bias in Chat GPT responses. In *ASONAM '23: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, (pp. 687–691). https://doi.org/10.1145/3625007.3627484

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., & Goldberg, Y. (2021). Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9, 1012–1031. https://doi.org/10.1162/tacl_a_00410

Fenwick, A., & Molnar, G. (2022). The importance of humanizing AI: Using a behavioral lens to bridge the gaps between humans and machines. *Discover Artificial Intelligence*, 2(1), 14. https://doi.org/10.1007/s44163-022-00030-8

Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1), paper 3. https://doi.org/10.3390/sci6010003

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. In N. Doorn, D. Schuurbiers, I. Van De Poel, & M. E. Gorman (Eds.), *Early engagement and new technologies: Opening up the laboratory* (Vol. 16, pp. 55–95). Springer Netherlands. https://doi.org/10.1007/978-94-007-7844-3_4

Georgiou, G. P. (2025). ChatGPT exhibits bias toward developed countries over developing ones, as indicated by a sentiment analysis approach. *Journal of Language and Social Psychology*, 44(1), 132–141. https://doi.org/10.1177/0261927x241298337

Gomez, A. A. R., & Benavides, M. L. C. (2024). Framework for bias detection in machine learning models: A fairness approach. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (pp. 1152–1154). https://doi.org/10.1145/3616855.3635731

Gómez, E., Shui Zhang, C., Boratto, L., Salamó, M., & Marras, M. (2021). The winner takes it all: Geographic imbalance and provider (un)fairness in educational recommender systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 1808-1812). https://doi.org/10.1145/3404835.3463235

Gondwe, G. (2023). CHATGPT and the Global South: How are journalists in sub-Saharan Africa engaging with generative AI? *Online Media and Global Communication*, 2(2), 228–249. https://doi.org/10.1515/omgc-2023-0023

Gross, N. (2023). What ChatGPT tells us about gender: A cautionary tale about performativity and gender biases in AI. *Social Sciences*, 12(8), 435. https://doi.org/10.3390/socsci12080435

Hake, J., Crowley, M., Coy, A., Shanks, D., Eoff, A., Kirmer-Voss, K., Dhanda, G., & Parente, D. J. (2024). Quality, accuracy, and bias in ChatGPT-based summarization of medical abstracts. *The Annals of Family Medicine*, 22(2), 113–120. https://doi.org/10.1370/afm.3075

Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 2(4), 100089. https://doi.org/10.1016/j.tbench.2023.100089

Haltaufderheide, J., & Ranisch, R. (2024). The ethics of ChatGPT in medicine and healthcare: A systematic review on Large Language Models (LLMs). *Npj Digital Medicine*, 7(1), 183. https://doi.org/10.1038/s41746-024-01157-x

Hassan, M. R., & Aziz, M. S. A. (2025). Artificial Intelligence Comic Strip (AICS) generators: A review of subscription models, pricing, and user satisfaction. *International Journal on Perceptive and Cognitive Computing*, *11*(1), 95-102. https://doi.org/10.31436/ijpcc.v11i1.553

Hosseini, M., & Horbach, S. P. J. M. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research Integrity and Peer Review*, 8(1), article number 4. https://doi.org/10.1186/s41073-023-00133-5

Iqbal, R., & Ismail, S. (2025). Unbiased AI for a sovereign digital future: A bias detection framework. *Procedia Computer Science*, 254, 118–126. https://doi.org/10.1016/j.procs.2025.02.070

Kalla, D., Smith, N., Samaah, F., & Kuraku, S. (2023). Study and analysis of ChatGPT and its impact on different fields of study. *International Journal of Innovative Science and Research Technology*, 8(3), 827–833. https://ssrn.com/abstract=4402499

Kaplan, D. M., Palitsky, R., Arconada Alvarez, S. J., Pozzo, N. S., Greenleaf, M. N., Atkinson, C. A., & Lam, W. A. (2024). What's in a name? Experimental evidence of gender bias in recommendation letters generated by ChatGPT. *Journal of Medical Internet Research*, 26, e51837. https://doi.org/10.2196/51837

Katare, D., Kourtellis, N., Park, S., Perino, D., Janssen, M., & Ding, A. Y. (2022). Bias detection and generalization in AI algorithms on edge for autonomous driving. In *2022 IEEE/ACM 7th Symposium on Edge Computing* (SEC) (pp. 342–348). https://doi.org/10.1109/SEC54971.2022.00050

Kim, J., & Lee, J. (2023, March 7). How does ChatGPT introduce transport problems and solutions in North America? *Findings*. https://doi.org/10.32866/001c.72634

Kim, J., Lee, J., Jang, K. M., & Lourentzou, I. (2024). Exploring the limitations in how ChatGPT introduces environmental justice issues in the United States: A case study of 3,108 counties. *Telematics and Informatics*, *86*, 102085. https://doi.org/10.1016/j.tele.2023.102085

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion*, *99*, 101861. https://doi.org/10.1016/j.inffus.2023.101861

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., & Maningo, J. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Kuşcu, O., Pamuk, A. E., Sütay Süslü, N., & Hosal, S. (2023). Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? *Frontiers in Oncology*, *13*, 1256459. https://doi.org/10.3389/fonc.2023.1256459

Lee, G. U., Hong, D. Y., Kim, S. Y., Kim, J. W., Lee, Y. H., Park, S. O., & Lee, K. R. (2024). Comparison of the problem-solving performance of ChatGPT-3.5, ChatGPT-4, Bing Chat, and Bard for the Korean emergency medicine board examination question bank. *Medicine*, *103*(9), e37325. https://doi.org/10.1097/MD.0000000000037325

Lippens, L. (2024). Computer says 'no': Exploring systemic bias in ChatGPT using an audit approach. *Computers in Human Behavior: Artificial Humans*, *2*(1), 100054. https://doi.org/10.1016/j.chbah.2024.100054

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology*, *1*(2), 100017. https://doi.org/10.1016/j.metrad.2023.100017

Liu, Z., Janowicz, K., Currier, K., & Shi, M. (2024). Measuring geographic diversity of foundation models with a natural language–based geo-guessing experiment on GPT-4. *AGILE: GIScience Series*, *5*, 1–7. https://doi.org/10.5194/agile-giss-5-38-2024

Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Library Hi Tech News*, *40*(3), 26–29. https://doi.org/10.1108/LHTN-01-2023-0009

Mergen, A., Çetin-Kılıç, N., & Özbilgin, M. F. (2025). Artificial Intelligence and Bias Towards Marginalised Groups: Theoretical Roots and Challenges. In J. Vassilopoulou & O. Kyriakidou (Eds.), *International Perspectives on Equality, Diversity and Inclusion* (pp. 17–38). Emerald Publishing. https://doi.org/10.1108/S2051-233320250000012004

Moon, D., & Ahn, S. (2025). Metrics and Algorithms for Identifying and Mitigating Bias in AI Design: A Counterfactual Fairness Approach. *IEEE Access*, 13, 59118–59129. https://doi.org/10.1109/ACCESS.2025.3556082

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: Measuring ChatGPT political bias. *Public Choice*, 198, 3–23. https://doi.org/10.1007/s11127-023-01097-2

OpenAI (2023). *GPT-4 technical report*. https://arxiv.org/html/2303.08774v5?utm_source=chatgpt.com

OpenAI (n.d.). *How ChatGPT and our foundation models are developed*. Retrieved from https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed (Archived at https://web.archive.org/web/20250719180532/https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed)

Park, S., Ryu, S., & Choi, E. (2022). Do language models understand measurements? arXiv preprint https://arxiv.org/pdf/2210.12694

Park, S. (2024). AI chatbots and linguistic injustice. Journal of Universal Language, 25(1), 99-119. https://doi.org/10.22425/jul.2024.25.1.99

Plevris, V., Papazafeiropoulos, G., & Jiménez Rios, A. (2023). Chatbots put to the test in math and logic problems: A comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. AI, 4(4), 949–969. https://doi.org/10.3390/ai4040048

Raman, R., Calyam, P., & Achuthan, K. (2024). ChatGPT or Bard: Who is a better Certified Ethical Hacker? *Computers & Security*, 140, 103804. https://doi.org/10.1016/j.cose.2024.103804

Rana, S. A., Azizul, Z. H., & Awan, A. A. (2023). A step toward building a unified framework for managing AI bias. *PeerJ Computer Science*, 9, e1630. https://doi.org/10.7717/peerj-cs.1630

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Renshaw, A., Lourentzou, I., Lee, J., Crawford, T., & Kim, J. (2025). Comparing the spatial querying capacity of large language models: OpenAI's ChatGPT and Google's Gemini Pro. *The Professional Geographer*, 77(2), 186–198. https://doi.org/10.1080/00330124.2024.2434455

Rozado, D. (2023). The political biases of ChatGPT. *Social Sciences*, 12(3), 148. https://doi.org/10.3390/socsci12030148

Rudolph, J., Tan, S., & Tan, S. (2023). War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning & Teaching*, 6(1), 364–389. https://doi.org/10.37074/jalt.2023.6.1.23

Sallami, D., & Aïmeur, E. (2024). Fairframe: A fairness framework for bias detection and mitigation in news. *AI and Ethics*, 5, 2467–2483. https://doi.org/10.1007/s43681-024-00568-6

Samaan, J. S., Yeo, Y. H., Rajeev, N., Hawley, L., Abel, S., Ng, W. H., Srinivasan, N., Park, J., Burch, M., Watson, R., Liran, O., & Samakar, K. (2023). Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. *Obesity Surgery*, 33(6), 1790–1796. https://doi.org/10.1007/s11695-023-06603-5

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Tan Yip Ming, C., Rojas-Carabali, W., Cifuentes-González, C., Agrawal, R., Thorne, J. E., Tugal-Tutkun, I., Nguyen, Q. D., Gupta, V., de-la-Torre, A., & Agrawal, R. (2023). The potential role of large language models in Uveitis Care: Perspectives after ChatGPT and Bard launch. *Ocular Immunology and Inflammation*, 32(7), 1435–1439. https://doi.org/10.1080/09273948.2023.2242462

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6), 363–377. https://doi.org/10.1002/sam.11348

Tellez, N., Serra, J., Kumar, Y., Li, J. J., & Morreale, P. (2023). Gauging Biases in Various Deep Learning AI Models. In K. Arai (Ed.), *Intelligent Systems and Applications*, 544, 171–186. Springer International. https://doi.org/10.1007/978-3-031-16075-2_11

Waisberg, E., Ong, J., Masalkhi, M., Zaman, N., Sarker, P., Lee, A. G., & Tavakkoli, A. (2024). Google's AI chatbot "Bard": A side-by-side comparison with ChatGPT and its utilisation in ophthalmology. *Eye*, 38(4), 642–645. https://doi.org/10.1038/s41433-023-02760-0

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Copyright

Appendices

Appendix 1: Survey questionnaire

A survey to assess users' opinions on potential biases in ChatGPT

This survey explores the users' opinions on potential geographical biases in ChatGPT. Specifically, the survey aims to examine biases along three directions: physical geography, context (e.g. culture, race, religion, etc.), and nature (e.g. stereotype, completeness, ambiguity, etc.).

This survey is anonymous. The record of your survey responses does not contain any identifying information about you. Your responses will be kept in absolute confidence and will be used for academic purposes only.

This survey needs approximately 10 minutes to complete. Thank you very much for your contribution.

(1) Please select your country of present affiliation:

Choose

(2) How frequently do you use AI Chatbots?

- ◯ Always
- ◯ Very Often
- ◯ Sometimes
- ◯ Rarely
- ◯ Never

(3) What are the AI Chatbots that you prefer to use?

- [ ] Bing Chat
- [ ] ChatGPT
- [ ] Perplexity AI
- [ ] YouChat
- [ ] Google Bard
- [ ] Jasper
- [ ] Tidio Lyro
- [ ] Ada
- [ ] ChatSpot
- [ ] Other

The following sections specifically focus on ChatGPT. Proceed to conclude and submit the survey now if you lack any experience with ChatGPT.

(4) Which version of ChatGPT do you use?

- ◯ ChatGPT-3.5 (also called Free version)
- ◯ ChatGPT-4 (also called Plus version/ paid version)
- ◯ Both versions

(5) ChatGPT has had a substantial impact on scholars.

- ◯ Strongly Agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly Disagree

(6) ChatGPT generates biased responses, regardless of the topic.

- ◯ Strongly Agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly Disagree

Questions (7) to (13) specifically address potential geographical biases in ChatGPT.
Note: the term 'some geographical regions' in the following questions pertains to the country where you are currently located, or any other country or countries around the world, or even a combination of both.

(7) Responses generated by ChatGPT are less accurate for some geographical regions.

- ◯ Strongly Agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly Disagree

(8) Responses generated by ChatGPT are not highly relevant for some geographical regions.

- ◯ Strongly Agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly Disagree

(9) ChatGPT generates incomplete responses for some geographical regions.

- ◯ Strongly Agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly Disagree

(10) ChatGPT generates vague responses for some geographical regions.

- ◯ Strongly Agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly Disagree

(11) ChatGPT generates stereotype responses for some geographical regions. Stereotype: an often unfair and untrue belief that many people have about all people or things.

- ◯ Strongly Agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly Disagree

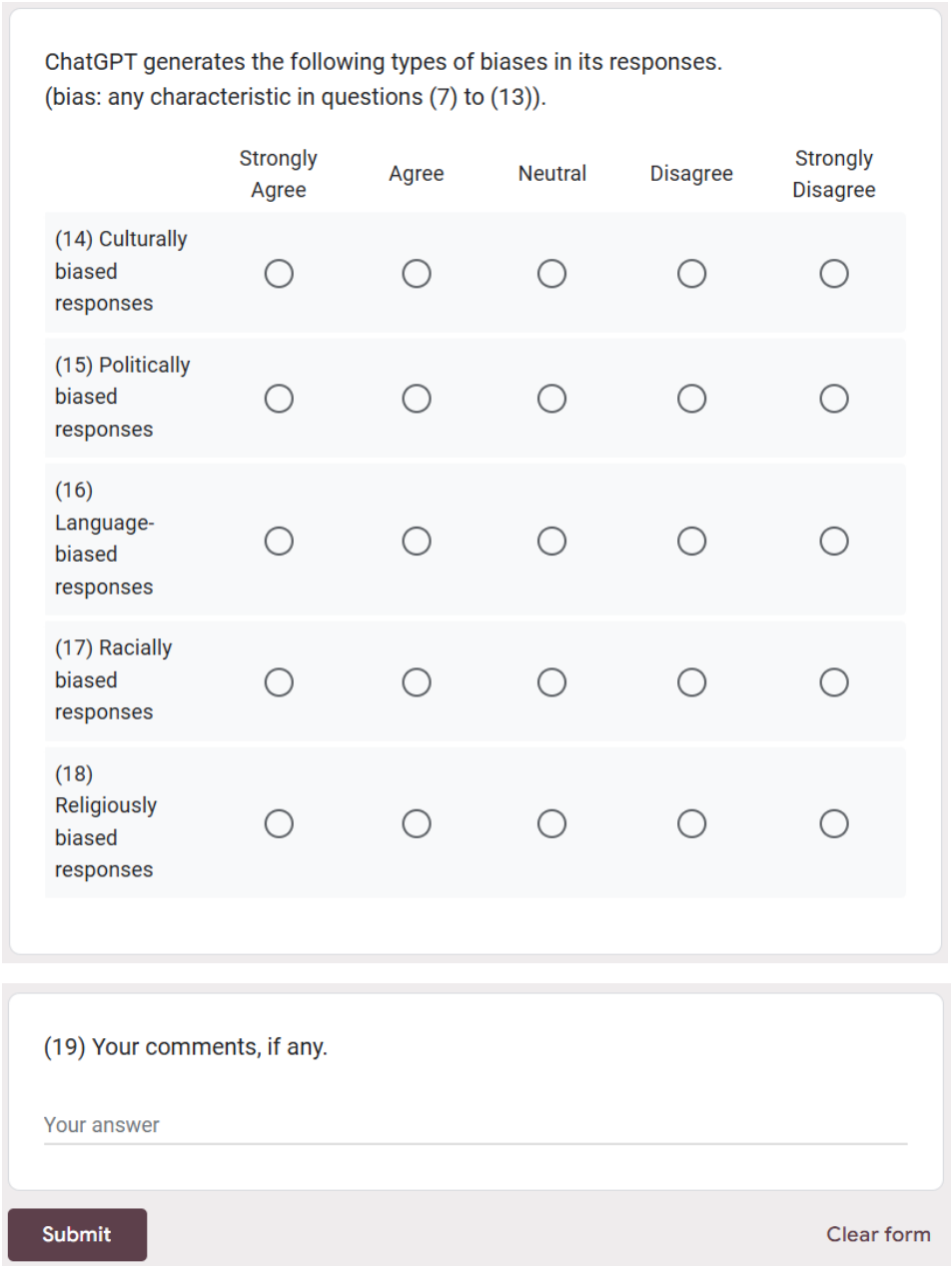
(12) ChatGPT generates preconceived responses for some geographical regions. Preconceived: ideas formed before having evidence.

- ◯ Strongly Agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly Disagree

(13) ChatGPT generates abominable responses for some geographical regions. Abominable: unpleasant and causes a feeling of disgust.

- ◯ Strongly Agree
- ◯ Agree
- ◯ Neutral
- ◯ Disagree
- ◯ Strongly Disagree

Questions (14) to (18) delve deeper into the geographical bias.



Appendix 2: Average length of each idea/sentence type

Region	Avg. opinion sentences		Avg. fact sentences		Avg. directive sentences		Avg. Similarity
	Res 1	Res 2	Res 1	Res 2	Res 1	Res 2	
Central Europe	3.8	3.2	10.8	15	1.8	6.8	3/5
Eastern Europe	3.2	3.8	14	17.6	2.2	4.8	4/5
Southern Europe	2.6	2.6	20.4	17.6	1.6	5.8	4/5

Northern Europe	3.6	1	18.4	19	2.8	7.8	4/5
Western Europe	4	2.4	19.8	15.4	3.2	3.8	5/5
East Africa	3	0.6	19	15.8	4	5	5/5
Southern Africa	3.2	2	19.2	15	4.4	5.6	4/5
West Africa	1.8	0.2	17	16	3.2	6.4	4/5
North Africa	3.8	2.8	9.8	11.6	4.6	5.8	4/5
Central Africa	4	2.8	10.6	12.4	5.8	6.6	3/5
South Asia	2.4	1.6	15.8	15.4	6.8	6.6	3/5
Central Asia	2	3	16.8	15.4	8.2	8.4	5/5
Southeast Asia	1.6	2.4	15	15.4	6.8	8	4/5
East Asia	0.8	2	17.6	17.4	8	7.8	5/5
Middle East	2.2	3.6	13.8	13.4	6.4	8.2	1/5
Oceania	2.4	2.8	14.2	11.4	4.6	6	2/5
Caribbean	4.2	3.6	11.8	14.6	7.4	8.2	1/5
South America	4.4	3.2	14.8	16.6	8.2	8.6	3/5
Central America	1.4	2	17.4	18.6	9	9	4/5
North America	4	2.7	18.3	17.3	6.3	9.3	3/3

Table 1. Composition of responses to test prompt 1.

Region	**Avg. opinion**		**Avg. fact**		**Avg. directive**		**Avg. Similarity**
	Res 1	**Res 2**	**Res 1**	**Res 2**	**Res 1**	**Res 2**	
Central Europe	1	0.6	2	1.6	0	0	2/5
Eastern Europe	0	0	2	2	0	0	2/5
Southern Europe	0	0	2	2	0	0	5/5
Northern Europe	0	0	2	2	0	0	5/5
Western Europe	0	0	2	2	0	0	5/5
East Africa	0	0	2	2	0	0	5/5
Southern Africa	0	0	2	2	0	0	5/5
West Africa	0	0	2	2	0	0	5/5
North Africa	0	0	2	2	0	0	5/5
Central Africa	0	0	2	2	0	0	5/5
South Asia	0	0	2	2	0	0	5/5
Central Asia	0	0	2	2	0	0	5/5
Southeast Asia	0	0	2	2	0	0	5/5
East Asia	0	0	2	2	0	0	5/5
Middle East	0	0	2	2	0	0	5/5

Oceania	0	0	2	2	0	0	4/5
Caribbean	0	0	2	2	0	0	5/5
South America	0	0	2	2	0	0	5/5
Central America	0	0	2	2	0	0	5/5
North America	0	0	2	2	0	0	3/3

Table 2. Composition of responses to test prompt 2.

Region	**Avg. opinion**		**Avg. fact**		**Avg. directive**		**Avg. Similarity**
	Res 1	**Res 2**	**Res 1**	**Res 2**	**Res 1**	**Res 2**	
Central Europe	1	1	1	1	0	0	0/5
Eastern Europe	1	1	1	1	0	0	1/5
Southern Europe	1	1	1	1	0	0	1/5
Northern Europe	1	1	1	1	0	0	0/5
Western Europe	0.8	1	1.6	1	0	0	2/5
East Africa	1	1	1	1	0	0	0/5
Southern Africa	1	1	1	1	0	0	0/5
West Africa	1	1	1	1	0	0	0/5
North Africa	1	0.8	1.2	1.2	0	0	1/5
Central Africa	0.8	1	1.2	1	0.2	0	0/5
South Asia	1	1	1	1	0	0	0/5
Central Asia	1	1	1	1	0	0	0/5
Southeast Asia	1	1	1	1	0	0	0/5
East Asia	1	0.8	1.4	1.2	0	0	1/5
Middle East	1.2	1	1.2	1	0	0	0/5
Oceania	1.2	0.8	1	1	0	0	1/5
Caribbean	1	1	1	1	0	0	0/5
South America	1	1	1	1	0	0	1/5
Central America	1	1	1	1	0	0	1/5
North America	1	1	1	1	0	0	2/3

Table 3. Composition of responses to test prompt 3.

Region	Avg. opinion		Avg. fact		Avg. directive		Avg. Similarity
	Res 1	Res 2	Res 1	Res 2	Res 1	Res 2	
Central Europe	1.2	0.6	13	7.6	6.6	0.8	1/5
Eastern Europe	1.4	1.4	15.6	10.2	8.4	6	3/5
Southern Europe	0.4	0.4	12.2	8.2	5	1	4/5
Northern Europe	0.6	0.6	15.6	13	4.2	4.4	3/5
Western Europe	0.4	0.2	16	12	1.6	2.8	5/5
East Africa	0.4	0.4	14.6	12.4	6	4.2	3/5
Southern Africa	0.6	0.6	15	15.2	6.2	5.2	5/5
West Africa	1	1	16.8	15.8	6.2	6.8	5/5
North Africa	0.8	0	11.2	10	3.2	3.8	5/5
Central Africa	1.2	0.8	18.6	13.2	6.4	7.6	5/5
South Asia	1	0.4	17.4	15.4	6	6.8	5/5
Central Asia	1	0	16.4	13.2	5.2	5.4	5/5
Southeast Asia	0.6	0.6	16.4	13.6	6.4	5.4	4/5
East Asia	0.2	0.4	13.6	11.8	3.6	3.6	5/5
Middle East	0.6	0	12.8	8.2	5	2.2	5/5
Oceania	0.2	0.4	11.4	9.8	4	2.8	5/5
Caribbean	0	0	6.6	6.4	0	0	5/5
South America	0.2	0	17.4	14.6	5.2	5.2	5/5
Central America	0	0	15.6	15.8	4.2	2.8	5/5
North America	0	0	19.3	13	4.2	2.6	3/3

Table 4. Composition of responses to test prompt 4.

Region	Avg. opinion		Avg. fact		Avg. directive		Avg. Similarity
	Res 1	Res 2	Res 1	Res 2	Res 1	Res 2	
Central Europe	0.6	0	4.8	7.2	0	0	3/5
Eastern Europe	0	0	6.4	7.2	0	0	2/5
Southern Europe	0	0	6.4	6	0	0	1/5
Northern Europe	0	0	6.4	6.2	0	0	2/5
Western Europe	0	0	6.2	6	0	0	2/5
East Africa	0	0	7	5.8	0	0	3/5
Southern Africa	0	0	7.8	6.6	0	0	1/5
West Africa	0	0	8	6	0	0	0/5
North Africa	0	0	7.2	6.2	0	0	2/5

Central Africa	0	0	7.6	6.8	0	0	1/5
South Asia	0	0	8.6	7.8	0	0	4/5
Central Asia	0	0	8.2	7	0	0	1/5
Southeast Asia	0	0	7.8	8.2	0	0	2/5
East Asia	0	0	9	9.8	0	0	2/5
Middle East	0	0	8.2	8.8	0	0	1/5
Oceania	0	0	9.4	10.6	0	0	1/5
Caribbean	0	0	8.8	9	0	0	1/5
South America	0	0	7.8	10.8	0	0	2/5
Central America	0	0	9	10.4	0	0	2/5
North America	0	0	8	11.6	0	0	0/3

Table 5. Composition of responses to test prompt 5.

Region	**Avg. opinion**		**Avg. fact**		**Avg. directive**		**Avg. Similarity**
	Res 1	**Res 2**	**Res 1**	**Res 2**	**Res 1**	**Res 2**	
Central Europe	1.2	0.8	24.2	16	8.2	5.8	5/5
Eastern Europe	5.4	2.8	22.4	16.2	9	5.4	4/5
Southern Europe	4.8	0.8	23.2	17.4	8	5.4	5/5
Northern Europe	1.6	0.4	28.2	18.6	8.4	6.6	5/5
Western Europe	0.2	0.2	29	18.8	8.8	6.2	5/5
East Africa	0.6	1	30.8	18.8	8.2	5.8	5/5
Southern Africa	0.6	0.8	22.6	19.8	6.4	6.2	5/5
West Africa	1	0.8	20.6	20.8	6.4	6.2	5/5
North Africa	1.2	0.8	29.4	20.6	7.2	6.4	3/5
Central Africa	1.2	1	24	19	9.4	7.2	2/5
South Asia	1.2	0.6	26.6	16.2	6.8	5.6	5/5
Central Asia	1	0	27.8	16	6.4	4.6	5/5
Southeast Asia	0	0	23.6	13.4	7	4.6	4/5
East Asia	0.2	0	18.8	13.2	6.8	4.8	5/5
Middle East	0.2	0	17.2	15	5.8	5.8	5/5
Oceania	0	0.4	17.8	18.2	6.4	7.6	5/5
Caribbean	0	0.2	14.4	21.4	5.6	5.6	3/5
South America	0	0.4	15.8	20.6	5.4	5.6	5/5
Central America	0	1	16.6	19	5.6	4.6	4/5
North America	0.3	0.6	14.6	18	6	5.6	2/3

Table 6. Composition of responses to test prompt 6.

Region	Avg. opinion		Avg. fact		Avg. directive		Avg. Similarity
	Res 1	Res 2	Res 1	Res 2	Res 1	Res 2	
Central Europe	0	0	6	6	0	0	4/5
Eastern Europe	0	0	6	6	0	0	5/5
Southern Europe	0	0	6	6	0	0	5/5
Northern Europe	0	0	6	6	0	0	5/5
Western Europe	0	0	6	6	0	0	2/5
East Africa	0	0	6	6	0	0	5/5
Southern Africa	0	0	6	6	0	0	5/5
West Africa	0	0	6	6	0	0	5/5
North Africa	0	0	6	6	0	0	5/5
Central Africa	0	0	6	5.8	0	0	5/5
South Asia	0	0	6	6	0	0	5/5
Central Asia	0	0	6	6	0	0	5/5
Southeast Asia	0	0	6	6	0	0	5/5
East Asia	0	0	6	6	0	0	5/5
Middle East	0	0	6	6	0	0	5/5
Oceania	0	0	6	6	0	0	5/5
Caribbean	0	0	6	6	0	0	4/5
South America	0	0	6	6	0	0	5/5
Central America	0	0	6	6	0	0	5/5
North America	0	0	6	6	0	0	3/3

Table 7. Composition of responses to test prompt 7.