Biased Geolocation in LLMs: Experiments on Probing LLMs for Geographic Knowledge and Reasoning

Mila Stillman, Anna Kruspe

Hochschule München, Lothstraße 64, München

Abstract

Geographic biases in Large Language Models (LLMs) are evident. Research has shown that both the training data and outputs of LLMs are skewed towards western and economically affluent countries, resulting in the underrepresentation of certain regions. In addition, LLMs are prone to hallucinations, which can lead to the generation of incorrect or fabricated information. In this paper, we present an additional analysis of the geographic knowledge and geospatial reasoning of LLMs through two experiments carried out in English on a global scale. Specifically, we evaluate the geospatial capabilities of four open-source LLMs, namely: Llama 2-7B, Llama 3 (8B and 70B) and Phi-3-mini-4k, and demonstrate that geographic knowledge within LLMs is unevenly distributed across different regions of the world. This imbalance could lead to unfair treatment of certain areas and impact various applications that use geographic knowledge, including mobility and remote sensing applications, that aim to use LLMs for data analysis and decision-making.

Keywords

Large Language Models, Bias, Geospatial data, Reasoning, Fairness

1. Introduction

The use of Large Language Models (LLMs) is appealing in many applications, including geospatial applications. For instance, in remote sensing, merging satellite image data with natural language is valuable for geospatial vision-language question answering [1]. Furthermore, a geographic "cognitive map" could be useful for orienting and finding paths in mobility applications and user interfaces [2]. During disaster events, an effective integration of textual data, such as social media data, into GeoAI frameworks is invaluable for disaster management and response [3, 4]. Finally, with the efficient integration of LLMs in geospatial workflows, automation and simplification of many geospatial tasks would become possible [5, 6, 7].

Language models (LMs) have demonstrated the ability to encode geographic and spatial information, as is evident in previous Language Models (LMs) and in recently released LLMs [8, 9, 10, 11, 12]. However, LLMs also exhibit geospatial biases, representing certain populations, languages, and countries better than others [13, 14, 15]. Biases were found worldwide in objective and subjective subjects [16, 17], factual accuracy [18], and inaccuracy due to geopolitical favoritism [19]. Moreover, gaps in the geospatial knowledge of LLMs have been identified and are further described in the related work section. In this study, we test the geographic knowledge of LLMs through two separate experiments. In the first experiment, we probed random uniformly distributed geocoordinates from a large number of countries and asked the LLMs to provide the countries to which these geocoordinates belong. Knowing that LLMs tokenize textual information as embeddings and, therefore, do not possess polygon information of geographic entities, we do not expect LLMs to perform well on this task. Instead, we hypothesize that the accuracy of this task would not be equal for different regions of the world. In the second experiment, we ask the LLMs to provide a trip itinerary for a round-the-world trip, using the countries from the first experiment as starting points, to test the LLMs' geographic reasoning abilities on a global scale. A trip around the world is a complex task, which requires a proper understanding of the Earth's

GeoExT 2025: Third International Workshop on Geographic Information Extraction from Texts at ECIR 2025, April 10, 2025, Lucca, Italy

mila.stillman@hm.edu (M. Stillman); anna.kruspe@hm.edu (A. Kruspe)

0009-0008-8976-8855 (M. Stillman); 0000-0002-2041-9453 (A. Kruspe)

geography and advanced planning abilities using a fixed number of stops in different countries. Here, we test the ability of LLMs to combine those reasoning skills and analyze the effect of different starting points on performance.

2. Related Work

Current LLMs use Transformers [20] as their backbone architecture. They are black boxes and are not interpretable without additional components [21]. These models are trained on large textual datasets, or training corpora. These often contain inherent selection biases, leading to models that misrepresent both the underlying data and the real-world phenomena they aim to capture [13]. For instance, crowd-sourced geographic data on platforms such as OpenStreetMap [22] and Wikipedia[23, 24], have systemic biases and an uneven distribution of geographic information. In addition, geographic information from geotagged social media posts is mainly skewed towards urban and wealthy regions [25, 26, 4].

Geographic knowledge in LLMs is an increasingly studied field, especially since spatial and temporal representations have been demonstrated in LLMs [9]. Furthermore, 18.7% of the Common Crawl corpus used to train LLMs have been estimated to contain geospatial information such as addresses and geocoordinates [27]. Research reveals the potential to use LLMs in applications, such as extracting geocoordinates of cities and locations [28, 11], trip itineraries [11], and the use of GIS agents to automate geospatial tasks [29]. However, investigations into the geographic knowledge of these models show limitations, distance distortions, and inequality between regions [30, 14, 31, 28, 11, 32]. [33] researched the geographic knowledge of ChatGPT by measuring its results while taking a Geographic Information Systems (GIS) exam. The research shows that the GPT-3.5 and GPT-4 models achieve test scores of 66% and 88.3%, respectively. Studies testing GPT-4's capabilities in route planning and geocoded information retrieval reveal a certain level of success on geospatial tasks. However, there are some limitations in abstract reasoning, suggesting that memorization could play a role in task performance [32, 11]. [34] shows that LLMs demonstrate confidence in simple route planning tasks using cognitive maps; however, the authors suggest that this confidence is likely attributed to memorized routes rather than a genuine understanding of cognitive maps, route planning strategies or inference capabilities. Furthermore, the authors indicate that LLMs tend to fail due to hallucinations, constructing too long routes in some cases, or getting stuck in loops. The authors in [35] tested the factuality of LLMs on a global scale using data according to the World Bank and found significant biases for countries with lower income levels and in certain regions. This paper is a continuation of the work on identifying and quantifying existing geographic biases in LLMs. Future mobility applications, such as route planning and navigation, as well as the customization of LLMs to different geographic locations, require precise geographic knowledge. Moreover, any potential bias or discriminatory treatment of certain locations by LLMs could pose harm to individuals and therefore requires a careful examination.

3. Methodology

Our experimental setup includes probing and comparing the output of four LLMs for extraction of georeferenced information. In the first experiment, instead of asking LLMs to indicate the geocoordinates of known locations, e.g., cities or points of interest, as has been done in previous research, we conduct a reverse geocoding experiment by selecting geocoordinates in a randomized manner and ask the LLMs to 'guess' their location. This task is not common in texts; however, intelligent models that have an embedded geographic component should be able to make these educated guesses, similarly to human geographers. The models are downloaded from Huggingface [36], namely Llama 2 [37] with 7B parameters, Llama 3 [38] with 8B and 70B parameters, and Phi 3 mini[39] with 4k parameters. The use of the four models is subject to their respective license agreements. The probing experiments are conducted using the `transformers` library from Hugging Face [40]. We choose open-source models due to the ability to run these locally without additional API costs. All experiments are run using two Nvidia RTX A6000 GPUs. The exact models used are as follows: Llama-2-7b-chat-hf, Meta-Llama-3.1-

8B-Instruct, Meta-Llama-3.3-70B-Instruct, and Phi-3-mini-4k-instruct. For Llama 2, the chat model was selected due to its additional human-feedback training, which improved its performance on benchmark datasets, compared to the original model. The instruction tuned models are otherwise used due to their improved performance on common industry benchmarks compared to base or chat models. The prompts are adjusted to fit the geographic context via a system prompt. Programming is excluded in this task, since models tend to use code to download and process geospatial data. Instead, to extract potential inherent biases, we probe the implicit geographic knowledge that LLMs already possess. The answer 'not available' is acceptable when the LLMs are unable to provide an answer. The prompt is adjusted to fit the template provided by each model, while the prompt text remains the same. The number of tokens is limited to 1000. The temperature values are the default values: for Llama 2 and Llama 3 models the temperature value is 0.6. For Phi-3 the temperature was set to 0.0 as suggested for inference in the model's card on Hugging Face. The Llama 3-70B model was quantized to 8-bit for memory optimization using `bitsandbytes` from transformers [40]. The text of the system prompt is as follows:

```
Your role is an expert geographer who is familiar with the
    geocoordinate system and world geography. Provide short and
    concise answers to the question you are asked. Programming is not
     allowed. If you do not know the answer, answer with 'not
    available'.
```

In both experiments, we used the 177 countries from the `naturalearth_lowres` dataset using the `Geopandas` [41] library in Python. In the first experiment, for each country, we generate random uniformly distributed geocoordinates within its bounding box. We use polygon data from the `'naturalearth_lowres'` dataset to keep only points that fall on land within the countries' polygons, removing any in the oceans, until reaching exactly 20 points per country. For each pair of generated geocoordinates, we ask the LLMs to indicate the country to which they belong. The text of the user prompt is provided below:

```
You will be given a set of geocoordinates in the form of (lat,long).
    Provide the country to which these geocoordinates belong. First,
    identify the country name, then provide only the country code in
    ISO 3166 format.

Here are the geocoordinates:
```

In both experiments, the LLMs are asked to identify the country name and provide the country code in ISO-3166 format. This approach to prompting allowed us to improve robustness and avoid differences in country names from abbreviations and other toponyms.

In the second experiment, we examine the ability of LLMs to plan routes on a global scale. We prompt the LLMs to plan a trip itinerary around the world. In other words, define a route that will circumnavigate the Earth and return to the starting point, by using any means of travel. Curating such a trip requires reasoning capabilities, understanding geography on a global scale, applying a circular shape to the route, while comprehending that the Earth is spherical. To assess potential biases in this task, we ran this experiment using a different starting point each time, that is, starting from each of the 177 countries in the dataset. The number of stops was limited to 12 countries. The LLMs are also asked to provide the geocoordinates at each stop. To prevent ambiguities associated with the phrase "around the world," which could imply visiting various locations globally without necessarily completing a full circle around the Earth, the experiment is repeated with revised wording, changing the phrase "a trip around the world" to "a trip circumnavigating the Earth". An example of the desired geocoordinate format was provided to encourage its adoption, due to the models' tendency to provide various formats of geocoordinates. Given the example, this tendency was reduced, but not fully eliminated. In this experiment, the user prompt is as follows:

```
Your task is to plan a trip around the world. Use a maximum of 12
    stops. You can use any means of travel. For each stop: first,
    identify the country name, then provide only the country code in
    ISO 3166 format. Aditionally, provide the geocoordinates at each
    stop in the format (latitude,longitude) in decimal degrees only (
    e.g., (20.600, 78.800)).

You start from:
```

In both experiments, the prompt is designed to make geospatial predictions in a zero-shot manner. Multiple prompt variations were explored to achieve concise results that convey the necessary information. Further improvements in the form of prompt engineering are left for future work. The code is available at github.com/Milast/geo_biases_llms.

During post-processing, the `pycountry` library is used to identify country codes in ISO-3166 format within the free text given by the LLMs. The rate of missing, or incorrect, values in the output of the countries by the LLMs in the first experiment is less than 2% for all models. We consider this error rate acceptable when working with generated text. This is partially caused by the option given to the models to answer with 'not available', as well as due to the LLMs providing false or no country codes in rare cases. The country Kosovo was manually added since it does not have an ISO-3166 code. The geocoordinates generated in the second experiment in some cases did not match the coordinates of the provided country or were beyond the limits of longitude and latitude values. Further inspection and analysis of false geocoding is left for future work.

4. Results

4.1. First experiment

First, we analyze whether the LLMs are able to correctly identify the countries. When asking LLMs for both country name and country code, all models had a high preference towards providing an answer rather than answering with the option 'not available'. The percentage of missing answers due to the optional 'not available' answer or incorrect country codes is provided in Table 1. The Llama 2 model exhibited the most biased behavior, as well as proved to be the least knowledgeable. The continents with the best performance are Europe and North America, followed by Oceania and South America. Asia had a relatively low score, and all models performed the worst in recognizing countries in Africa. This goes in line with previous research that studied distorted distances and geographic biases in this region. Both Antarctica and Seven Seas (open ocean) continents in the dataset have a single country each and demonstrated low accuracies. This is expected since those are remote areas that most likely do not appear often in the training data of LLMs. Surprisingly, the larger model of Llama 3 with 70B parameters performed worse in this task than the Llama 3 with 8B parameters and the much smaller Phi-3-mini-4k model. As a post-processing step, we analyze whether the country with which the LLMs respond belongs to the same continent as the correct country. Here, we notice a better performance for all continents. The Phi-3-mini-4k and Llama 3-8B models demonstrate high accuracies in Africa compared to the other two models. This suggests that LLMs do possess a certain understanding of geolocation on a continent level. The results for the three models are presented in Figure 1, Figure 2, and Table 1.

The Llama 2 model has the lowest diversity of countries given as output and the lowest number of countries which were correctly identified at least once. Interestingly, the Llama 3 model with 70B parameters has a lower diversity of countries than the both the Llama3-8B and the Phi-3 model. For Llama 2, countries such as Indonesia, South Africa, Turkey and Germany are frequently used. In Llama 3-8B, the most frequently used countries were South Africa and Mozambique, and the Llama 3 70B model frequently used China and Italy. Surprisingly, the Phi-3-mini-4k model most frequently used South Africa, Turkey, Ghana and Kenya. The top 20 countries and their frequency of occurrence for

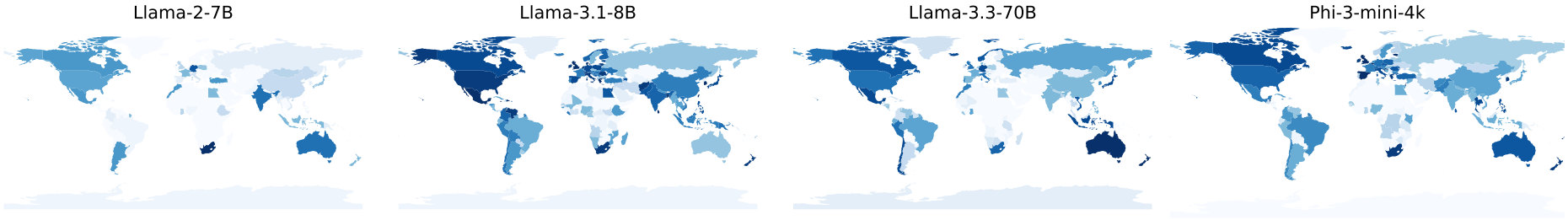


Figure 1: Number of correctly identified countries from a set of random geocoordinates within each country, for four Large Language Models: Llama 2-7B, Llama 3-8B, Llama 3-70B and Phi-3-mini-4k. Values range between 0 (the lightest shade), to 20 (the darkest shade.)

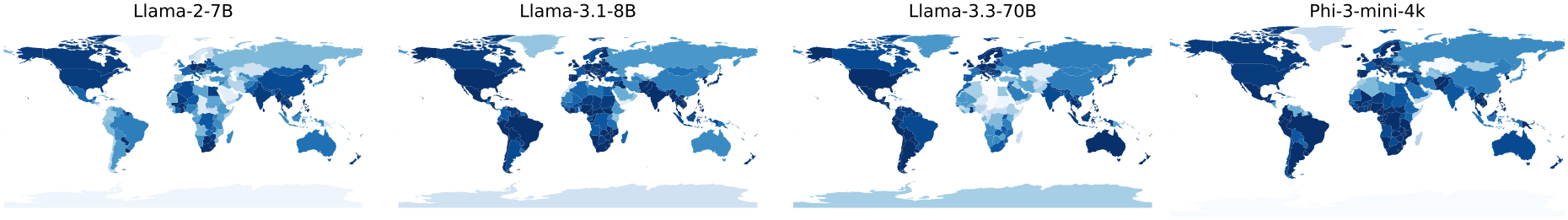


Figure 2: Number of correctly identified continents from a set of random geocoordinates within each country, for four Large Language Models: Llama 2-7B, Llama 3-8B, Llama 3-70B and Phi-3-mini-4k. Values range between 0 (the lightest shade), to 20 (the darkest shade.)

		Llama 2-7B	Llama 3-8B	Llama 3-70B	Phi-3-mini-4k
# of unique countries correctly identified (at least once)		39	130	87	98
# of unique countries mentioned by the LLM		74	158	125	137
% of total missing values		0.48	0.452	0.819	1.78
# of correctly identified countries (mean)	Africa	1.039	3.333	1.549	1.882
	Antarctica	1.00	1.0	2.0	0.0
	Asia	1.383	7.851	2.638	5.553
	Europe	1.256	13.256	4.795	11.103
	North America	2.0	10.056	6.278	6.056
	Oceania	3.857	9.571	7.0	8.143
	Seven seas (open ocean)	0.0	0.0	0.0	0.0
	South America	1.231	9.385	6.077	5.308
	Global	1.395	8.062	3.576	5.791
# of correctly identified continents (mean)	Africa	12.784	15.706	9.216	16.412
	Antarctica	1.0	4.0	7.0	0.0
	Asia	12.085	15.489	12.978	14.574
	Europe	13.333	19.0	17.795	18.513
	North America	9.778	17.0	15.111	14.5
	Oceania	14.142	17.143	17.571	18.286
	Seven seas (open ocean)	0.0	3.0	0.0	0.0
	South America	13.692	18.462	19.692	16.538
	Global	12.395	16.627	13.74	16.09

Table 1
Results for the first experiment for the four LLMs

	Llama 2-7B	Llama 3-8B	Llama 3-70B	Phi-3-mini-4k
Pearson correlation coefficient (r) (GDP)	0.263	0.265	0.539	0.149
p-value (GDP)	0.0004	0.0004	9.44×10^{-15}	0.047
Pearson correlation coefficient (population estimate)	0.303	0.177	0.611	0.134
p-value (population estimate)	4.1×10^{-5}	0.019	1.79×10^{-19}	0.075

Table 2
Pearson correlation coefficient and p-values for GDP and population estimate of countries and the number of times they were mentioned by the four LLMs

each model are presented in Table 3 in Appendix A. We also analyze the relationship between the frequency of countries provided by the LLMs and their Gross Domestic Product (GDP) and population estimate. The Llama 3-8B and the Phi 3 models show a weak correlation with population estimate, while both Llama 3-70B and Llama 2-7B show a moderate correlation. The correlation with GDP is most prominent in the Llama 3-70B model. We calculate the Pearson correlation coefficient and find significant p-values for both GDP and population estimate in all models except Phi-3, with p-values of 0.047 and 0.075 for GDP and population estimate, respectively. The Pearson correlation coefficients calculated and the corresponding p-values are summarized in Table 2. Further analysis of correlation with travel data or other training data was left for future work.

Figure 3 indicates the GDP and population estimate as a function of how often they were mentioned by the LLMs.

4.2. Second experiment

In the second experiment, the LLMs generate trip itineraries for a trip around the world. We calculate and analyze the average total distance traveled using the Haversine Distance formula [42] and find that

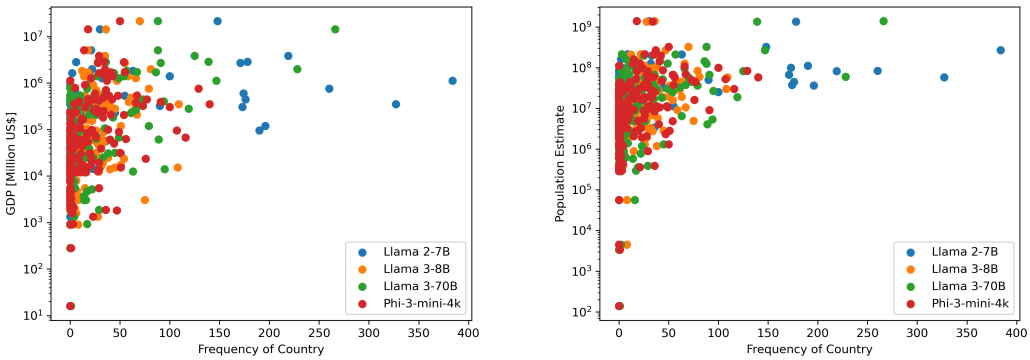


Figure 3: GDP (in Millions of USD) and Population estimation of the countries compared to the frequency in which they were mentioned by the LLMs in the first experiment.

it is much longer in the Llama 3-70B and Phi-3 models, with an average distance traveled resembling the Earth's circumference of around 40,000 km or higher. In comparison, the Llama 2 and Llama 3-8B models' average distances were close to 30,000 km or shorter. Moreover, the choice of wording, or lexical choice, has an effect on the average travel distance. For example, in the Phi 3 model the distance is shorter with the word 'circumnavigate', while for Llama 2 and Llama 3-8B it is longer. For Llama 2 and Phi 3, the average distance traveled from African countries is the shortest, followed by Asia and Europe. The distance traveled by Llama 3-70B is robust to changes in language and to starting countries on different continents.

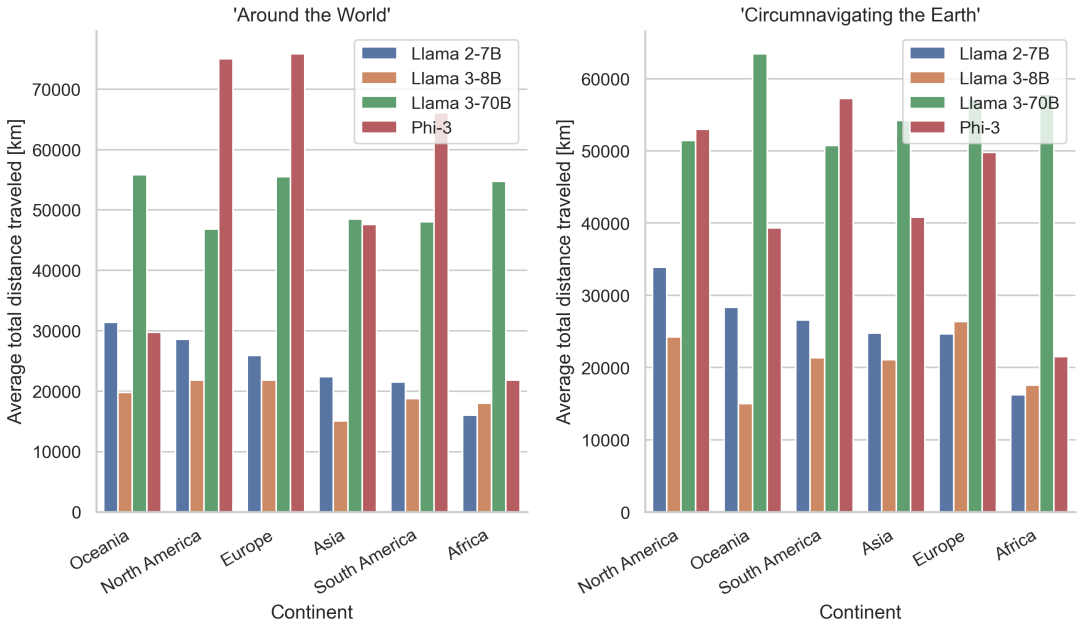


Figure 4: Average distance traveled using countries in different continents as starting points, per continent for the four LLMs. Left: using the lexical choice "around the world". Right: using the lexical choice "circumnavigating the Earth".

Furthermore, the percentage of countries chosen for the trip around the world that belong to the same continent as that of the starting country is found to be significantly higher in Africa for Phi-3 and Llama 2, irrespective of the syntactic variation and without repeating countries. In terms of the shape of the trips, the qualitative results suggest that Llama 3-70B can best capture a rounded trip. Examples of trip shapes can be found in the Appendix 6. The average distance traveled and the percentage of stops within the same continent as the starting point are presented in Figure 4 and Figure 5, respectively.

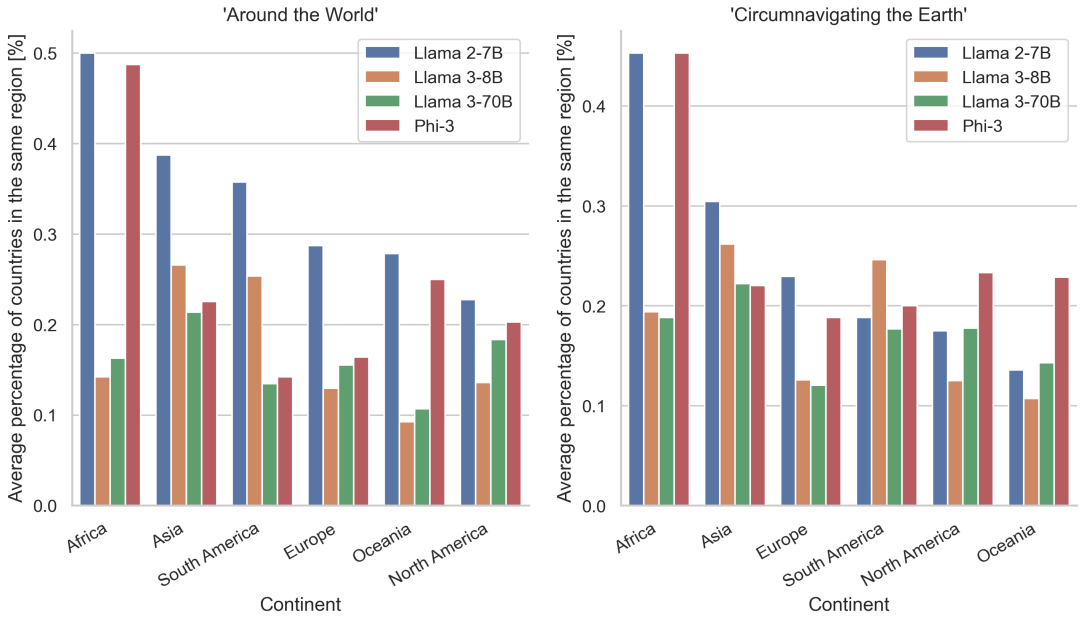


Figure 5: Average percentage of stops within the same continent using countries in different continents as starting points, per continent for the four LLMs and two syntactic variations. Left: using the lexical choice "around the world". Right: using the lexical choice "circumnavigating the Earth".

5. Discussion

Some of the inaccuracies generated by LLMs are expected, due to the training data of LLMs, which includes geographic data (e.g., from Wikipedia, social media, etc.) that are potentially skewed towards western and more affluent parts of society. Although LLMs might not possess, or be able to interpret, polygon information of countries to perform reverse geocoding accurately, the models did attempt to guess the country in most cases, even when given an option to answer with 'not available'. This effect could partially be due to the choice of models, prompts, and model parameters, which could be further optimized in the future. Surprisingly, the Llama 2 model has a small number of selected countries, which could be the result of the difference in the size of the training data compared to the other models. The Phi-3-mini-4k, a much smaller model, has competitive performance in terms of accuracy and bias. Surprisingly, the least biased model was Llama 3-8B, and performed even better than the Llama 3-70B model in the first experiment. This indicates that perhaps a larger model is not necessarily superior to smaller models in the geographic information extraction task. However, in the second experiment the larger model proved to be more robust to lexical choice and provided more circular routes than the other models, suggesting improved geospatial reasoning capabilities. The Phi-3 model generated longer trips too; however, it was less robust to formulation and formed trip shapes that are more chaotic. Some potential explanations for less accurate routes could include travel restrictions to certain countries and not enough training data from people traveling around the world from these regions. Another interesting result is that while toponym resolution remains a challenge in GeoAI, specifically when probing LLMs for geographic knowledge, we found that unifying country names using country codes worked well. Nevertheless, the geocoordinates provided by the LLMs had different formats and required manual work and resolution. In some cases, they were beyond the geographic limits, or misrepresented the complementary country. The optimization, verification and standardization of the output of LLMs in the geographic context could be a future research direction.

6. Conclusion

Large Language Models exhibit biases, including geographic biases. In this paper, we demonstrate that the geographic knowledge of LLMs is partially inaccurate and biased against less economically strong regions of the world. We conducted experiments with four LLMs to assess their ability to perform two geospatial tasks. We find that the representation of objective geographic knowledge is unequal between regions and that some countries are overrepresented. Future work includes testing more LLMs, using multilingual prompts and chain-of-thought prompts with provided polygon information, a larger amount of generated geocoordinates, and prompts using different geographic granularity. In addition, spatial fairness is an important aspect in avoiding common pitfalls [43]. Since random geocoordinates are used in the first experiment, it is crucial to conduct a thorough analysis of these locations. For instance, in densely populated areas, random geocoordinates may provide more information compared to those in rural or sparsely populated regions. Furthermore, conducting correlation analyzes with data sources such as worldwide travel data and other sources of LLMs' training data could be beneficial. Finally, bias in LLMs affects a variety of emerging socially critical applications, e.g., in human resources, journalism, and education [44]. Furthermore, integrating geolocation smoothly with LLMs for remote sensing and mobility applications would require high accuracy and trustworthiness. Uncovering such biases and knowledge gaps is a critical first step towards improving the explainability of these models and for the development of future solutions.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Improve writing style of some of the sentences. Further, the author(s) used ChatGPT to generate code for data analysis functions, which were used as a template and modified to fit the relevant data and tasks. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] S. Lobry, D. Marcos, J. Murray, D. Tuia, Rsvqa: Visual question answering for remote sensing data, IEEE Transactions on Geoscience and Remote Sensing 58 (2020) 8555–8566.

[2] J. Feng, Y. Du, T. Liu, S. Guo, Y. Lin, Y. Li, CityGPT: Empowering urban spatial cognition of large language models (2024). URL: http://arxiv.org/abs/2406.13948. doi:10.48550/arXiv.2406.13948, arXiv:2406.13948 [cs].

[3] B. Zhou, L. Zou, A. Mostafavi, B. Lin, M. Yang, N. Gharaibeh, H. Cai, J. Abedin, D. Mandal, Victimfinder: Harvesting rescue requests in disaster response from social media with bert, Computers, Environment and Urban Systems 95 (2022) 101824.

[4] X. X. Zhu, Y. Wang, M. Kochupillai, M. Werner, M. Häberle, E. J. Hoffmann, H. Taubenböck, D. Tuia, A. Levering, N. Jacobs, A. Kruspe, K. Abdulahhad, Geoinformation harvesting from social media data: A community remote sensing approach, IEEE Geoscience and Remote Sensing Magazine 10 (2022) 150–180.

[5] K. Janowicz, S. Gao, G. McKenzie, Y. Hu, B. Bhaduri, GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond, 2020.

[6] G. Mai, C. Cundy, K. Choi, Y. Hu, N. Lao, S. Ermon, Towards a foundation model for geospatial artificial intelligence (vision paper), in: Proceedings of the 30th International Conference on Advances in Geographic Information Systems, 2022, pp. 1–4.

[7] S. Gao, Y. Hu, W. Li, Handbook of geospatial artificial intelligence, CRC Press, Boca Raton, 2023.

[8] M. M. Louwerse, R. A. Zwaan, Language encodes geographical information, Cognitive Science 33 (2009) 51–73.

[9] W. Gurnee, M. Tegmark, Language models represent space and time, arXiv preprint arXiv:2310.02207 (2023).
[10] R. Manvi, S. Khanna, G. Mai, M. Burke, D. Lobell, S. Ermon, Geollm: Extracting geospatial knowledge from large language models, arXiv preprint arXiv:2310.06213 (2023).
[11] J. Roberts, T. Lüddecke, S. Das, K. Han, S. Albanie, Gpt4geo: How a language model sees the world's geography, arXiv preprint arXiv:2306.00020 (2023).
[12] K. Salmas, D.-A. Pantazi, M. Koubarakis, Extracting Geographic Knowledge from Large Language Models: An Experiment, in: KBC-LM'23: Knowledge Base Construction from Pre-trained Language Models workshop at ISWC 2023, 2023.
[13] R. Navigli, S. Conia, B. Ross, Biases in large language models: origins, inventory, and discussion, ACM Journal of Data and Information Quality 15 (2023) 1–21.
[14] J. Dunn, B. Adams, H. T. Madabushi, Pre-trained language models represent some geographic populations better than others, arXiv preprint arXiv:2403.11025 (2024).
[15] A. Kruspe, Towards detecting unanticipated bias in large language models, arXiv preprint arXiv:2404.02650 (2024).
[16] R. Manvi, S. Khanna, M. Burke, D. Lobell, S. Ermon, Large language models are geographically biased, arXiv preprint arXiv:2402.02680 (2024).
[17] A. Kruspe, M. Stillman, Saxony-Anhalt is the worst: Bias towards german federal states in large language models, in: German Conference on Artificial Intelligence (Künstliche Intelligenz), Springer, 2024, pp. 160–174.
[18] S. Mirza, B. Coelho, Y. Cui, C. Pöpper, D. McCoy, Global-Liar: Factuality of LLMs over time and geographic regions, arXiv preprint arXiv:2401.17839 (2024).
[19] F. Faisal, A. Anastasopoulos, Geographic and geopolitical biases of language models, arXiv preprint arXiv:2212.10408 (2022).
[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[21] E. Cambria, L. Malandri, F. Mercorio, N. Nobani, A. Seveso, Xai meets llms: A survey of the relation between explainable ai and large language models, arXiv preprint arXiv:2407.15248 (2024).
[22] J. Thebault-Spieker, B. Hecht, L. Terveen, Geographic Biases are 'Born, not Made' Exploring Contributors' Spatiotemporal Behavior in OpenStreetMap, in: Proceedings of the 2018 ACM International Conference on Supporting Group Work, 2018, pp. 71–82.
[23] C. Hube, Bias in wikipedia, in: Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 717–721.
[24] M. Graham, B. Hogan, R. K. Straumann, A. Medhat, Uneven geographies of user-generated information: Patterns of increasing informational poverty, Annals of the Association of American Geographers 104 (2014) 746–764.
[25] B. Hecht, M. Stephens, A tale of cities: Urban biases in volunteered geographic information, in: Proceedings of the international AAAI conference on Web and Social Media, volume 8, 2014, pp. 197–205.
[26] L. Li, M. F. Goodchild, B. Xu, Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr, Cartography and geographic information science 40 (2013) 61–77.
[27] I. Ilyankou, M. Wang, S. Cavazzi, J. Haworth, Quantifying geospatial in the common crawl corpus, in: Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems, 2024, pp. 585–588.
[28] P. Bhandari, A. Anastasopoulos, D. Pfoser, Are large language models geospatially knowledgeable?, in: Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, 2023, pp. 1–4.
[29] Z. Li, H. Ning, Autonomous GIS: the next-generation AI-powered GIS, Int. J. Digit. Earth 16 (2023) 4668–4686.
[30] R. Decoupes, R. Interdonato, M. Roche, M. Teisseire, S. Valentin, Evaluation of Geographical Distortions in Language Models: A Crucial Step Towards Equitable Representations, arXiv preprint arXiv:2404.17401 (2024).

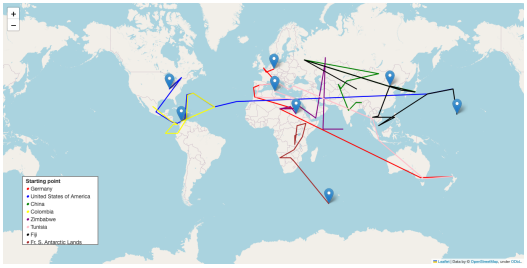
[31] P. Schwöbel, J. Golebiowski, M. Donini, C. Archambeau, D. Pruthi, Geographical erasure in language generation, arXiv preprint arXiv:2310.14777 (2023).
[32] S. Das, Evaluating the Capabilities of Large Language Models for Spatial and Situational Understanding, Ph.D. thesis, Thesis (MA). University of Cambridge, 2023.
[33] P. Mooney, W. Cui, B. Guan, L. Juhász, Towards understanding the geospatial skills of chatgpt: Taking a geographic information systems (gis) exam, in: Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, 2023, pp. 85–94.
[34] I. Momennejad, H. Hasanbeig, F. Vieira Frujeri, H. Sharma, N. Jojic, H. Palangi, R. Ness, J. Larson, Evaluating cognitive maps and planning in large language models with cogeval, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 69736–69751. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/dc9d5dcf3e86b83e137bad367227c8ca-Paper-Conference.pdf.
[35] M. Moayeri, E. Tabassi, S. Feizi, Worldbench: Quantifying geographic disparities in llm factual recall, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 1211–1228.
[36] S. M. Jain, Hugging face, in: Introduction to transformers for NLP: With the hugging face library and models to solve problems, Apress Berkeley, CA, 2022, pp. 51–67.
[37] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).
[38] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).
[39] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl, et al., Phi-3 technical report: A highly capable language model locally on your phone, arXiv preprint arXiv:2404.14219 (2024).
[40] T. Wolf, Transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2020).
[41] K. Jordahl, J. V. den Bossche, M. Fleischmann, J. Wasserman, J. McBride, J. Gerard, J. Tratner, M. Perry, A. G. Badaracco, C. Farmer, G. A. Hjelle, A. D. Snow, M. Cochran, S. Gillies, L. Culbertson, M. Bartos, N. Eubank, maxalbert, A. Bilogur, S. Rey, C. Ren, D. Arribas-Bel, L. Wasser, L. J. Wolf, M. Journois, J. Wilson, A. Greenhall, C. Holdgraf, Filipe, F. Leblanc, geopandas/geopandas: v0.8.1, 2020. URL: https://doi.org/10.5281/zenodo.3946761. doi:10.5281/zenodo.3946761.
[42] C. C. Robusto, The cosine-haversine formula, The American Mathematical Monthly 64 (1957) 38–40.
[43] S. Shaham, G. Ghinita, C. Shahabi, Models and mechanisms for spatial data fairness, in: Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases, volume 16, NIH Public Access, 2022, p. 167.
[44] C. Filippo, G. Vito, S. Irene, B. Simone, F. Gualtiero, Future applications of generative large language models: A data-driven case study on chatgpt, Technovation 133 (2024) 103002.

A. Frequency of use of countries by the LLMs

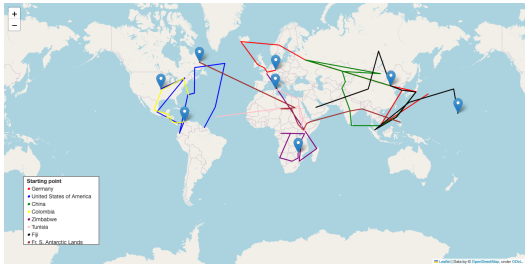
Llama 2-7B		Llama 3-8B		Llama 3-70B		Phi-3-mini-4k	
Country	Freq.	Country	Freq.	Country	Freq		
Indonesia	384	South Africa	109	China	266	South Africa	140
South Africa	327	Mozambique	108	Italy	228	Turkey	129
Turkey	260	Italy	81	Indonesia	147	Ghana	116
Germany	219	Turkey	79	India	139	Kenya	107
Morocco	196	Liberia	75	Germany	125	Egypt	100
Ethiopia	190	United States	70	Chile	119	Israel	91
India	178	Greece	67	Madagascar	95	Argentina	77
Argentina	176	Egypt	66	Norway	94	Senegal	76
Poland	174	Mexico	66	France	91	Colombia	73
Egypt	173	Argentina	63	Croatia	89	Thailand	67
France	171	Romania	63	Japan	88	Philippines	66
United States	148	Venezuela	61	United States	88	Tanzania	56
Australia	100	Israel	59	Russian Federation	86	Canada	55
Colombia	90	Spain	58	Morocco	79	New Zealand	54
Brazil	63	France	55	Argentina	73	United Kingdom	54
Belgium	55	Senegal	54	South Africa	70	Albania	51
Nepal	51	Morocco	51	Canada	67	Estonia	50
Peru	45	Mali	51	Vietnam	66	United States	50
Uruguay	36	Indonesia	47	Nicaragua	63	Dominican Republic	49
Canada	35	Albania	45	Peru	61	Belgium	48

Table 3
20 most frequently given answers by each LLM and the frequency in which they were mentioned by the LLMs in the first experiment.

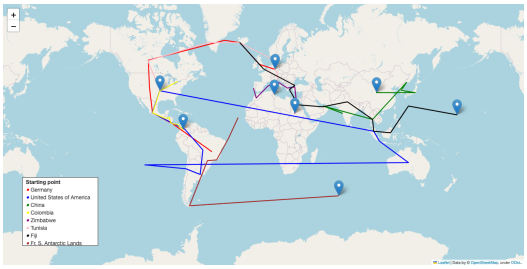
B. Round-the-World Trip examples



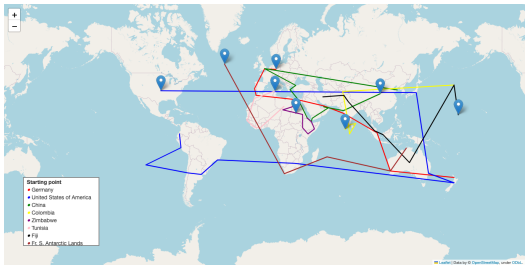
(a) Llama 2-7B, ”trip around the world”



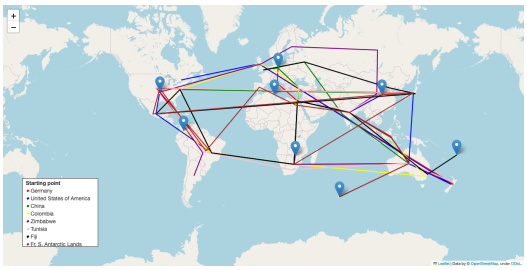
(b) Llama 2-7B, ”trip circumnavigating the Earth”



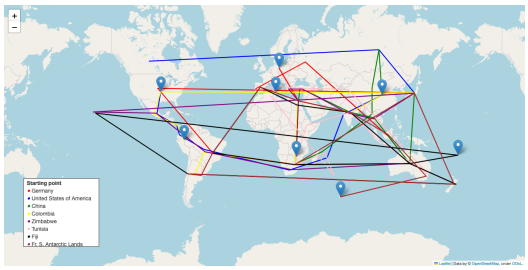
(c) Llama 3-8B, ”trip around the world”



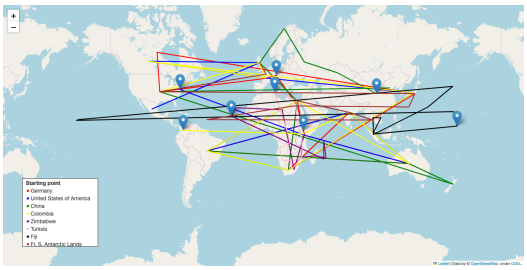
(d) Llama 3-8B, ”trip circumnavigating the Earth”



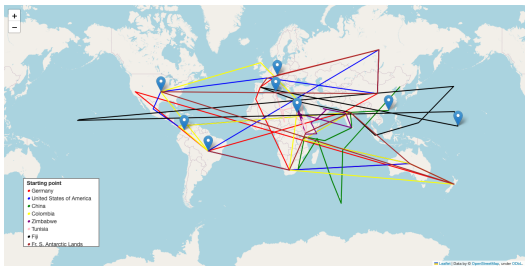
(e) Llama 3-70B, ”trip around the world”



(f) Llama 3-70B, ”trip circumnavigating the Earth”



(g) Phi-3-mini-4k, ”trip around the world”



(h) Phi-3-mini-4k, ”trip circumnavigating the Earth”

Figure 6: Round-the-world trip examples that the four models generated, demonstrating trips from different countries as starting points, namely: Germany, United States of America, China, Colombia, Zimbabwe, Tunisia, Fiji and Fr. S. Antarctic Lands using the two lexical choices: ”trip around the world” and ”trip circumnavigating the Earth”.