BIASED GEOLOCATION IN LLMS: EXPERIMENTS ON PROBING LLMS FOR GEOGRAPHIC KNOWLEDGE

Mila Stillman
Department of Computer Science and Mathematics
Munich University of Applied Sciences
Munich, Bavaria
mila.stillman@hm.edu

Anna Kruspe
Department of Computer Science and Mathematics
Munich University of Applied Sciences
anna.kruspe@hm.edu

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. Moreover, LLMs are prone to hallucinations, which may lead to the generation of incorrect or fabricated information. In this paper, we present an analysis of the geographic knowledge and geospatial reasoning of LLMs through two distinct experiments conducted on a global scale. Specifically, we evaluate the geospatial capabilities of three open-source LLMs, namely: Llama-2, Llama-3 and Phi-3, and demonstrate that geographic knowledge within LLMs is unevenly distributed across different regions of the world. This imbalance could lead to an unfair treatment of certain areas and impact various applications that use geographic knowledge, including remote sensing applications, that rely on LLMs for data analysis and decision-making.

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 Lobry et al. (2020). Furthermore, a geographic "cognitive map" could be useful for orientation and path finding in mobility applications and user interfaces Feng et al. (2024). During disaster events, an effective integration of textual data, such as social media data into GeoAI frameworks is invaluable for disaster management and response Zhou et al. (2022)Zhu et al. (2022). Finally, the automation of many geospatial tasks could become possible Janowicz et al. (2020); Mai et al. (2022); Gao et al. (2023).

Language models (LMs) have demonstrated an ability to encode geographic and spatial information, as is evident in earlier Language Models (LMs) and in recently released LLMs Louwerse & Zwaan (2009); Gurnee & Tegmark (2023); Manvi et al. (2023); Roberts et al. (2023); Salmas et al. (2023). However, LLMs also exhibit geospatial biases, representing certain populations, languages and countries better than others Navigli et al. (2023); Dunn et al. (2024). Biases were found globally in objective as well as subjective subjects Manvi et al. (2024), factual accuracy Mirza et al. (2024), and inaccuracy due to geopolitical favoritism Faisal & Anastasopoulos (2022). In this study, we test LLMs' geographic knowledge through two separate experiments. First, we probe random uniformly distributed geocoordinates from a large number of countries, and ask the LLMs to provide the country name to which these geocoordinates belong. Knowing that LLMs do not posses specific 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. Second, we ask the LLMs to provide a trip-itinerary for a round-the-world trip, using the countries we test in the first experiment as the 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 geography, and planning abilities of such a trip using a fixed number of stops in different countries. Here, we test the LLMs' ability to combine those reasoning skills in different regions.

2 Related Work

Current LLMs use Transformers Vaswani et al. (2017) as their backbone architecture. They are black boxes and are not interpretable without additional components Cambria et al. (2024). These models are trained on large datasets that often contain inherent biases, which could skew the data representation Navigli et al. (2023). For instance, crowd-sourced geographic information such as geotagged social media posts typically favor urban and wealthy areas over rural and economically disadvantaged ones Hecht & Stephens (2014); Li et al. (2013); Zhu et al. (2022). This bias also extends to contributors of crowd-sourced geographic data in platforms like OpenStreetMap Thebault-Spieker et al. (2018) and to Wikipedia, which shows uneven geographic information distribution and systemic biases Hube; Graham et al. (2014).

Geographic knowledge in LLMs is an increasingly studied field. Research reveals the potential to use LLMs in applications, such as extracting coordinates of cities and locations Bhandari et al. (2023); Roberts et al. (2023), trip itineraries Roberts et al. (2023), and the use of GIS agents to automate geospatial tasks Li & Ning (2023). However, investigations into these models' geographic knowledge disparities show significant limitations and regional inequalities Decoupes et al. (2024); Dunn et al. (2024); Schwöbel et al. (2023); Bhandari et al. (2023). Mooney et al. (2023) research the geographic knowledge of ChatGPT by measuring its results while taking a Geographic Information Systems (GIS) exam. The research shows that the models GPT-3.5 and GPT-4 achieve 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, which raises questions about the role of memorization in task performance Das (2023); Roberts et al. (2023). Momennejad et al. (2023) 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 the cognitive maps, route-planning strategies, or inference capabilities. Furthermore, the authors indicate that LLMs tend to fail due to hallucinations, construction of overly long routes, or getting trapped in loops. Future mobility applications, such as autonomous driving, and 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 thus requires a careful examination.

3 Methodology

Our experimental setup includes probing and comparing the outputs of three LLMs for geographic knowledge from geocoordinates. Instead of asking LLMs to indicate locations of known geocoordinates of cities and points of interest, as has been done in previous research, we select coordinates in a randomized manner. The models are downloaded from Huggingface[1] in a quantized GGUF format, namely Llama-2, Llama-3 and Phi-3, and probed using llama.cpp and langchain[2]. We choose quantized open-source models due to the ability to run these locally without using excessive computational resources, or additional costs. Hence, all experiments are run using an Apple M2 processor with 24GB of memory. The exact models we used are: llama-2-7b-chat.Q4_K_M, Meta-Llama-3-8B-Instruct.Q4_K_M and Phi-3-mini-4k-instruct-q4. For Llama-2, the chat model was selected due to its better empirical performance in this task than the quantized instruct model. The prompts are adjusted to fit the geographic context via a system prompt. Programming is excluded in this task, to probe the pure geographic knowledge that the LLMs already possess. The answer 'not available' is acceptable when such data is not available to the LLM. The prompt example is adjusted to fit a template that is provided by each model, while the prompt text remained equal for all models. The value of `temperature` used is 0.85 and the `top_k` value is set to 1. The number of tokens is limited to 300 in the first experiment and to 1000 in the second experiment. The text of the prompt for the system is presented in the following textbox:

```
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'.
```

[1] https://huggingface.co/

[2] https://python.langchain.com/

In the first experiment, we select 177 countries. The list of countries is available in the `naturalearth_lowres` dataset from the Geopandas[3] library in Python. For each country, we generate 20 random uniformly distributed geocoordinates from within the country's borders. The geometries in the form of polygons or multipolygons are retrieved from the geometry column of the `'naturalearth_lowres'` dataset. For each pair of the generated geocoordinates, we ask the LLMs to indicate the country to which they belong. The text of the user prompt is presented below:

```
You will be given a set of geocoordinates in the form (lat,long). What country do these
     geocoordinates belong to? Provide only the country name (e.g., Germany).

Here are the geocoordinates:
```

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, to circumnavigate the Earth's surface using any means of travel. We ran this experiment using a different starting point each time, namely starting from each of the 177 countries used previously. 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 circumnavigation of the Earth, the experiment is repeated with revised wording, changing the phrase "a trip around the world" to "a trip circumnavigating the Earth's surface". In this experiment, the user prompt is as follows:

```
Your task is to plan a trip around the world. Write each country you will pass on that trip
     and provide the geocoordinates in that country  in the format (latitude,longitude) in
     decimal degrees only. Provide only the country name and geocoordinates. Use a maximum of
      12 stops. You can use any means of travel.

You start at:
```

The prompt is designed to make zero-shot geospatial predictions. For both tasks, multiple prompt variations are explored to achieve results that are concise and convey the necessary information. Further improvements to the prompts are left for future work. The code is available under https://github.com/LINK_WILL_BE_ADDED_IN_FINAL_VERSION.

4 RESULTS

4.1 FIRST EXPERIMENT

In the first experiment, we analyse how many of the countries are correctly identified by the LLMs. The Phi-3 model has a better overall performance than the Llama-2 and Llama-3 models, with 156 out of 177 countries in the dataset being correctly identified at least once. Llama-2 and Llama-3 identify 67 and 89 countries correctly at least once, respectively. For the three models, most countries in Africa are never identified correctly, similar to some countries in South America, the Middle East, Asia and Eastern European countries. The three models perform better in this task in Western and European countries, as well as some countries in South America and countries with a large population in Asia, i.e., India and China. The results for the three models are presented in figure 1 and summarized in table 1.

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, especially in the Phi-3 model. The Llama-3 model struggles the most in selecting countries in Africa. Both Antarctica and Seven Seas (open ocean) continents in the dataset used for the experiments have a single country each The results are presented in figure 2

The Llama-2 and Llama-3 models show a much lower diversity of countries than the Phi-3 model. For Llama-2, countries such as France, Germany and Indonesia are frequently used. In Llama-3, the most frequently used countries were Spain and Brazil. Interestingly, in the Phi-3 model there is a much higher diversity of countries, and the frequency of the countries used was more balanced. The top 20 countries and their frequency of occurrence for each model are presented in table 3 in Appendix A. While Llama-2 indicates that the country is "not available" 475 times, Llama-3

[3]https://geopandas.org/en/v0.10.0/docs/reference/api/geopandas.datasets.available.html

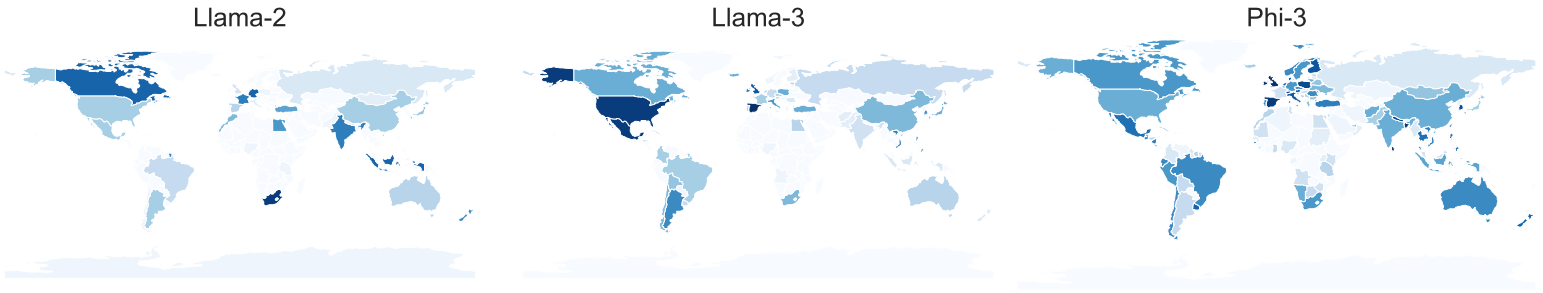

Figure 1: Number of correctly identified continents from a set of random geocoordinates within the country, for three large language models: Llama-2, Llama-3 and Phi-3. Values range between 0 (the lightest shade), to 20 (the darkest shade.)

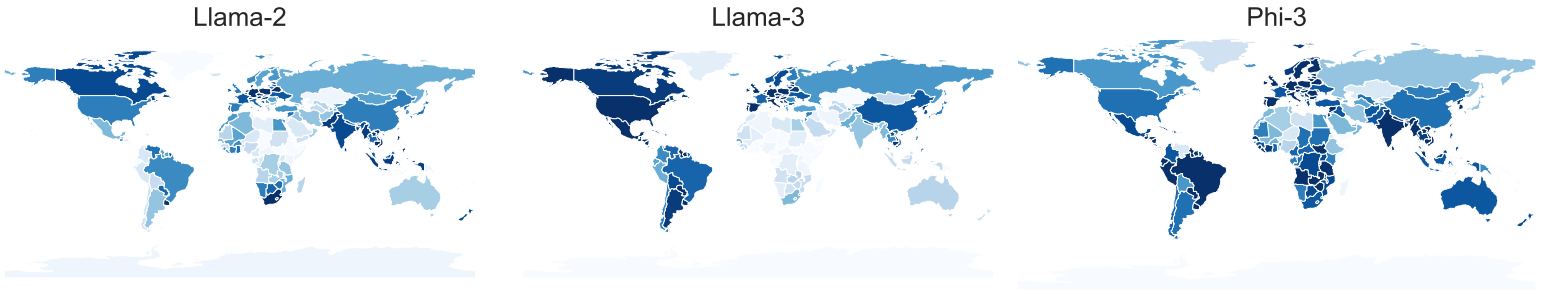

Figure 2: Number of correctly identified continents from a set of random geocoordinates within the country, for three large language models: Llama-2, Llama-3 and Phi-3. Values range between 0 (the lightest shade), to 20 (the darkest shade.)

Table 1: Results for the first experiment for the three LLMs

		Llama-2	Llama-3	Phi-3
# of unique countries correctly identified (at least once)		67	89	156
# of total correctly identified countries	Africa	0.041	0.017	0.099
	Antarctica	0.05	0.0	0.0
	Asia	0.066	0.083	0.301
	Europe	0.059	0.183	0.515
	North America	0.083	0.225	0.386
	Oceania	0.136	0.064	0.329
	Seven seas (open ocean)	0.0	0.0	0.0
	South America	0.05	0.192	0.346

Table 2: Point biserial correlation coefficient and p-values

	Llama-2	Llama-3	Phi-3
Point biserial correlation coefficient (GDP)	0.267	0.219	0.126
p-value (GDP)	0.0003	0.0034	0.095
Point biserial correlation coefficient (population estimate)	0.27	0.181	0.118
p-value (population estimate)	0.00028	0.0157	0.1173

indicates this only 30 times. The Phi-3 model indicates this option 61 times. Overall, all models have a strong tendency to provide an answer rather than acknowledging the unavailability of the requested information. We also analyze the most frequently used countries in terms of their Gross Domestic Product (GDP) and population estimate. The Phi-3 model is the least biased in that sense, while both Llama-2 and Llama-3 have their most frequently used countries in the upper percentile of both GDP and population estimation. We calculate the point biserial correlation coefficient Lev (1949) to assess the relationship between a binary variable, indicating whether a country was mentioned by the LLM, and the country's population estimate or GDP. We find significant results for both GDP and population estimation in the Llama-2 and Llama-3 models. For instance, a correlation with GDP results in p-values of 0.0003 and 0.0034 for Llama-2 and Llama-3, respectively. For Phi-3 the p-values correlating selection of countries with their GDP and population estimate are around 0.1, indicating a weak correlation. The calculated point biserial correlation coefficients and the p-values are summarized in table 2.

4.2 Second experiment

In the second experiment, the LLMs generate trip-itineraries around the world. We see that the average distance traveled was significantly longer in the Phi-3 model compared to the Llama-2 and Llama-3 models, where the distance resembled the Earth's circumference of around 40,000 km. Choice of wording does not have a large effect of the average traveled distance. For most combinations of model and wording, the average distance traveled from countries in Africa is the shortest, followed by Asia and Europe.

Additionally, the percentage of countries chosen for the trip around the world that belong to the same continent as that of the starting country are found to be the highest in Africa. The effects of choice of wording on the shape of the trips could be studied in the future. The average distance traveled and the percentage of stops within the same continent of staring point per continent for the three models are presented in figure 3 and figure 4, respectively.

5 Discussion

Some of the inaccuracies generated by the LLMs are expected, due to the training data of LLMs, which includes some geographic data (e.g., from Wikipedia), however, it typically describes named locations, such as cities and points of interest in the form of single points. LLMs might not possess or be able to interpret polygon information of countries to perform such calculation accurately.

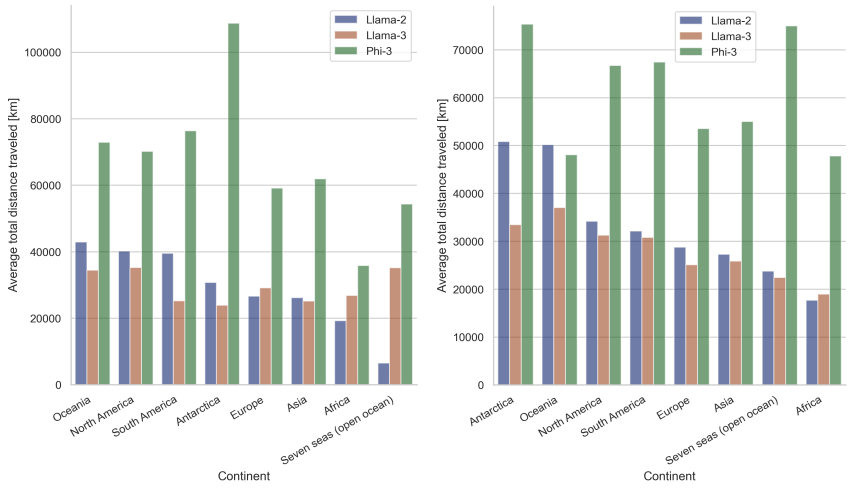

Figure 3: Average distance traveled per continent for the three LLMs. Left: using the wording ”around the world trip”. Right: using the wording ”circumnavigating the Earth’s surface”.

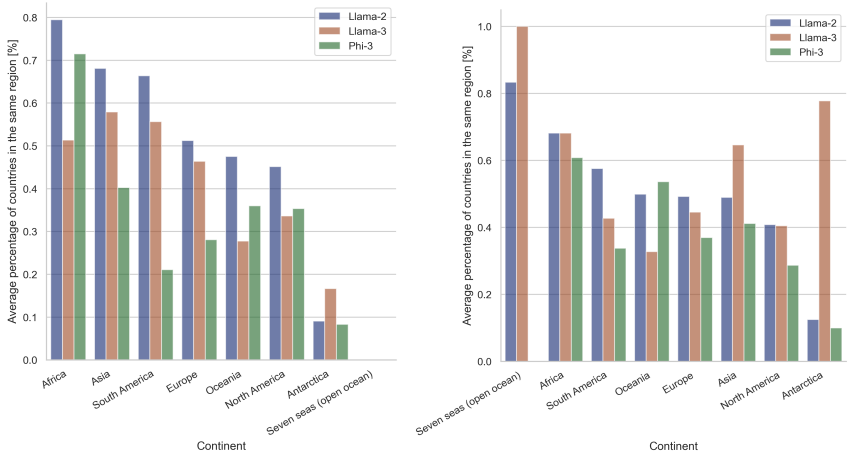

Figure 4: Average percentage of stops within the same continent for the three LLMs. Left: using the wording ”around the world trip”. Right: using the wording ”circumnavigating the Earth’s surface”.

The interesting result is, however, that the models did attempt to guess the country of origin in most cases, even when given an option to answer with 'not available'. This effect could partially be due to the choice of temperature, which could be further optimized in the future. Surprisingly, the Llama-2 and Llama-3 models select a relatively small number of countries, which could be the result of the difference in the size of training data of both models. The Phi-3-mini model, a much smaller model, has a significantly better performance in terms of both accuracy and bias. However, in the second experiment it provides a much longer distance than the Llama models. This could be explained by the model's tendency to be optimized towards a larger coverage of the Earth, a lack of reasoning ability to comprehend the meaning of a round-the-world trip, or a tendency to hallucinate. The average distance traveled for trips starting in the African continent, however, is significantly shorter than that of the other continents for all three models. For the Llama models, the average distance resembles the distance of the Earth's circumference of around 40,000km for both choices of wording. It is shorter (below 30,000km on average) for Europe, Asia and Africa. Some potential explanations include travel restrictions to certain countries, and not enough training data from people travelling around the world from certain regions. In this experiment, spatial fairness is an important aspect in order to avoid common pitfalls Shaham et al. (2022). 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. Another interesting result we discover is that toponym resolution remains a challenge in GeoAI, specifically when probing LLMs for geographic knowledge. We found that unifying coordinate formats, country name variations and undefined territories given by the LLMs required a large amount of manual work and resolution. The optimization and standradization of the output of LLMs in the geographic context could be a future research direction.

6 Conclusion and Future Work

Large Language Models exhibit biases, including geographic biases. In this paper, we demonstrate that the geographic knowledge of LLMs is biased. We experiment with three open-source LLMs to determine their geographic accuracy on two geospatial tasks, finding . Future work could include testing larger LLMs, using a larger amount of generated geocoordinates and conducting correlation analyses using data sources such as global statistical travel data or other relevant datasets on which the LLMs may have been trained, as these could potentially skew the results.

Finally, lack of knowledge, bias and prejudice in LLMs affects a variety of emerging socially critical applications, e.g., human resources management, journalism, and education Filippo et al. (2024). Integrating LLMs in remote sensing and mobility applications would require high accuracy and trustworthiness. Uncovering such biases and gaps in knowledge is a critical first step toward improving the explainability of these models and lays the groundwork for developing future solutions.

References

Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. Are large language models geospatially knowledgeable? In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pp. 1–4, 2023.

Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Navid Nobani, and Andrea Seveso. Xai meets llms: A survey of the relation between explainable ai and large language models. *arXiv preprint arXiv:2407.15248*, 2024.

Sowmen Das. *Evaluating the Capabilities of Large Language Models for Spatial and Situational Understanding*. PhD thesis, Thesis (MA). University of Cambridge, 2023.

Rémy Decoupes, Roberto Interdonato, Mathieu Roche, Maguelonne Teisseire, and Sarah Valentin. Evaluation of Geographical Distortions in Language Models: A Crucial Step Towards Equitable Representations. *arXiv preprint arXiv:2404.17401*, 2024.

Jonathan Dunn, Benjamin Adams, and Harish Tayyar Madabushi. Pre-trained language models represent some geographic populations better than others. *arXiv preprint arXiv:2403.11025*, 2024.

Fahim Faisal and Antonios Anastasopoulos. Geographic and geopolitical biases of language models. *arXiv preprint arXiv:2212.10408*, 2022.

Jie Feng, Yuwei Du, Tianhui Liu, Siqi Guo, Yuming Lin, and Yong Li. Citygpt: Empowering urban spatial cognition of large language models. *arXiv preprint arXiv:2406.13948*, 2024.

Chiarello Filippo, Giordano Vito, Spada Irene, Barandoni Simone, and Fantoni Gualtiero. Future applications of generative large language models: A data-driven case study on chatgpt. *Technovation*, 133:103002, 2024.

Song Gao, Yingjie Hu, and Wenwen Li. *Handbook of geospatial artificial intelligence*. CRC Press, Boca Raton, December 2023.

Mark Graham, Bernie Hogan, Ralph K Straumann, and Ahmed Medhat. Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers*, 104(4):746–764, 2014.

Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

Brent Hecht and Monica 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, pp. 197–205, 2014.

Christoph Hube. Bias in wikipedia. In *Proceedings of the 26th International Conference on World Wide Web Companion*.

Krzysztof Janowicz, Song Gao, Grant McKenzie, Yingjie Hu, and Budhendra Bhaduri. GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond, 2020.

Joseph Lev. The point biserial coefficient of correlation. *The Annals of Mathematical Statistics*, 20 (1):125–126, 1949.

Linna Li, Michael F Goodchild, and Bo Xu. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and geographic information science*, 40(2):61–77, 2013.

Zhenlong Li and Huan Ning. Autonomous GIS: the next-generation AI-powered GIS. *Int. J. Digit. Earth*, 16(2):4668–4686, December 2023.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020.

Max M Louwerse and Rolf A Zwaan. Language encodes geographical information. *Cognitive Science*, 33(1):51–73, 2009.

Gengchen Mai, Chris Cundy, Kristy Choi, Yingjie Hu, Ni Lao, and Stefano Ermon. Towards a foundation model for geospatial artificial intelligence (vision paper). In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pp. 1–4, 2022.

Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213*, 2023.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*, 2024.

Shujaat Mirza, Bruno Coelho, Yuyuan Cui, Christina Pöpper, and Damon McCoy. Global-Liar: Factuality of LLMs over time and geographic regions. *arXiv preprint arXiv:2401.17839*, 2024.

Ida Momennejad, Hosein Hasanbeig, Felipe Vieira Frujeri, Hiteshi Sharma, Nebojsa Jojic, Hamid Palangi, Robert Ness, and Jonathan Larson. Evaluating cognitive maps and planning in large language models with cogeval. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 69736–69751. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/dc9d5dcf3e86b83e137bad367227c8ca-Paper-Conference.pdf.

Peter Mooney, Wencong Cui, Boyuan Guan, and Levente 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*, pp. 85–94, 2023.

Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21, 2023.

Jonathan Roberts, Timo Lüddecke, Sowmen Das, Kai Han, and Samuel Albanie. Gpt4geo: How a language model sees the world's geography. *arXiv preprint arXiv:2306.00020*, 2023.

Konstantinos Salmas, Despina-Athanasia Pantazi, and Manolis 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.

Pola Schwöbel, Jacek Golebiowski, Michele Donini, Cédric Archambeau, and Danish Pruthi. Geographical erasure in language generation. *arXiv preprint arXiv:2310.14777*, 2023.

Sina Shaham, Gabriel Ghinita, and Cyrus Shahabi. Models and mechanisms for spatial data fairness. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 16, pp. 167. NIH Public Access, 2022.

Jacob Thebault-Spieker, Brent Hecht, and Loren 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*, pp. 71–82, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Bing Zhou, Lei Zou, Ali Mostafavi, Binbin Lin, Mingzheng Yang, Nasir Gharaibeh, Heng Cai, Joynal Abedin, and Debayan Mandal. Victimfinder: Harvesting rescue requests in disaster response from social media with bert. *Computers, Environment and Urban Systems*, 95:101824, 2022.

Xiao Xiang Zhu, Yuanyuan Wang, Mrinalini Kochupillai, Martin Werner, Matthias Häberle, Eike Jens Hoffmann, Hannes Taubenböck, Devis Tuia, Alex Levering, Nathan Jacobs, Anna Kruspe, and Karam Abdulahhad. Geoinformation harvesting from social media data: A community remote sensing approach. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):150–180, 12 2022.

Llama-2		**Llama-3**		**Phi-3**	
Country	Freq.	Country	Freq.	Country	Freq
France	494	Spain	522	Turkey	85
Not Available	475	Brazil	317	Switzerland	81
Germany	454	Germany	212	Romania	80
Indonesia	382	Poland	162	Puerto Rico	79
India	247	Italy	153	Tanzania	75
South Africa	208	Mexico	148	Italy	74
Turkey	194	Turkey	141	Guinea-Bissau	72
Egypt	100	Czechia	133	Belgium	69
Brazil	82	United States of America	110	Namibia	69
Argentina	78	Portugal	107	Spain	67
China	70	South Africa	103	Nigeria	66
Morocco	60	Chile	97	Canada	64
New Zealand	55	Argentina	79	Not Available	61
United States of America	53	Russia	79	Brazil	58
Canada	51	Uruguay	66	Jordan	57
Ethiopia	50	France	59	Poland	56
Poland	44	Japan	58	Kazakhstan	54
Spain	39	China	58	Kenya	53
Nepal	35	Vietnam	57	Israel	52
Antarctica	24	Dominican Republic	56	India	52

Table 3: 20 most frequently given answers by each LLM and their frequency in the experiment

A Appendix / supplemental material

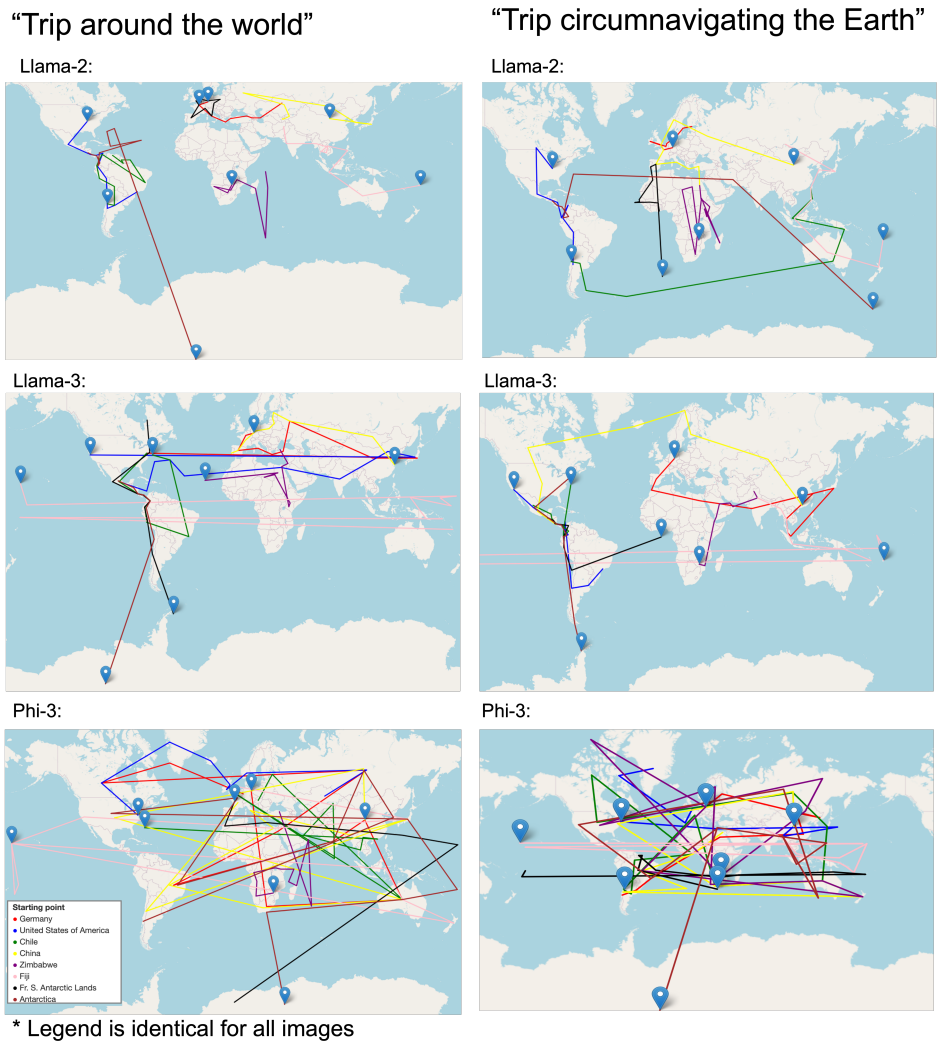

Figure 5: Example of round-the-world trip routes starting from a country from each of the continents for the three LLMs.

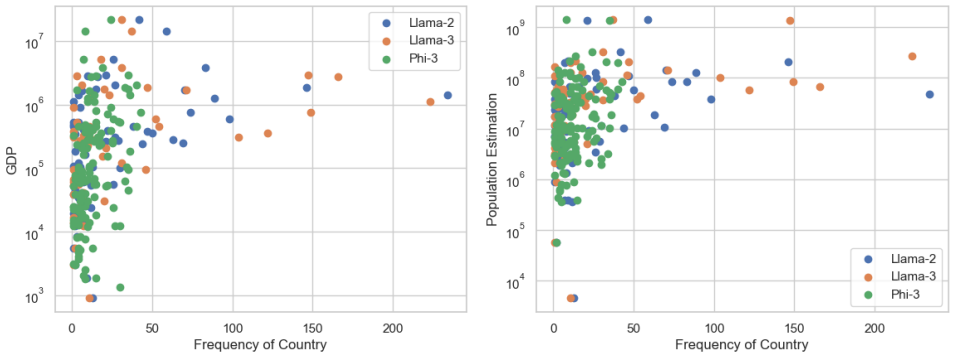

Figure 6: GDP and Population estimation of the countries compared to how frequently they were selected by the LLMs in the first experiment.