Evidence for Systematic Bias in the Spatial Memory of Large Language Models

Nir Fulman[1,*,†], Abdulkadir Memduhoğlu[1,2,†] and Alexander Zipf[1,3]

[1]GIScience Chair, Institute of Geography, Heidelberg University, Heidelberg, Germany
[2]Department of Geomatic Engineering, Faculty of Engineering, Harran University, Sanliurfa, Türkiye
[3]HeiGIT at Heidelberg University, Heidelberg, Germany

Abstract

We report our initial findings from an examination of potential systematic bias in Large Language Models' (LLM) spatial reasoning capabilities. We devised a series of questions to probe the spatial reasoning abilities of four LLMs: GPT-3.5, GPT-4, Gemini, and Llama-2, targeting four specific biases rooted in human spatial perception: hierarchical, proximity and directional biases. The questions encompassed scenarios challenging the models' spatial reasoning, and each question was posed 10 times independently to gauge the consistency of the LLMs' responses. The models demonstrated a strong understanding of straightforward geographical relationships, achieving 87% accuracy in questions that did not challenge biases in spatial reasoning. However, when faced with questions highlighting these biases, the models' accuracy dropped to 24%. We discuss the design of a large-scale experiment aimed at examining spatial cognition biases in large language models and identifying potential mitigation strategies.

Keywords

geographic reasoning, large language models, cognitive maps, systematic bias

1. Introduction

Recent studies primarily view Large Language Models (LLMs) in geography as tools linking natural language to geographic information systems [1]. However, Roberts et al. [2] showcased GPT-4's [3] inherent ability to perform spatial reasoning tasks. They highlighted tasks that extend beyond mere recall of factual information, namely GPT-4's proficiency in calculating the final destinations of routes based on initial locations, modes of transport, directions, and travel durations, without reliance on external processing engines. These capabilities open up practical applications such as creating personalized travel itineraries. Identifying the weaknesses of LLMs in spatial tasks may assist in guiding their development in this direction.

We investigate the possibility that biases in human spatial reasoning may manifest in LLMs, focusing on four well-studied ones: (a) Hierarchical bias refers to the cognitive tendency to infer the direction between two points based on the dominant geographical orientation of their

GeoExT 2024: Second International Workshop on Geographic Information Extraction from Texts at ECIR 2024, March 24, Glasgow, Scotland
*Corresponding author.
†These authors contributed equally.
nir.fulman@uni-heidelberg.de (N. Fulman); memduhoglu@uni-heidelberg.de (A. Memduhoğlu); zipf@uni-heidelberg.de (A. Zipf)
0000-0002-2629-2641 (N. Fulman); 0000-0002-9072-869X (A. Memduhoğlu); 0000-0003-4916-9838 (A. Zipf)

larger categorical groups (such as states or regions), leading to inaccuracies when exceptions to these general orientations exist [4]. (b) People often underestimate distances within the same categorical group, perceiving them as shorter than distances between different groups, even when the distances across groups are actually shorter [5]. (c) Rotation bias refers to the tendency to adjust the mental representation of geographical elements, aligning them more closely with conventional cardinal directions than their actual orientations [6]. This simplification leads to misconceptions about the true positions of locations, as individuals mentally 'rotate' geographical layouts to fit a more straightforward, north-south/east-west alignment, irrespective of their true, more complex orientations. (d) Alignment bias refers to the cognitive inclination to overestimate the alignment of geographically grouped locations, leading to skewed perceptions of their actual latitudinal or longitudinal relationships [7].

LLMs rely on associative learning and contextual data processing to understand and generate human-like text [8]. These models may manifest systematic biases in spatial reasoning due to their training data and learning mechanisms. While human biases in spatial reasoning are rooted in mental mapping [7], which LLMs do not possess, we hypothesize that they may exhibit similar biases, based on three considerations. First, LLMs learn from textual data, and biases in human spatial reasoning can be present in textual descriptions of geography. Second, as humans generalize and simplify in their cognitive maps, leading to biases, they also do so in their textual descriptions of locations. LLMs, learning from such descriptions, could inherit and perpetuate these biases in their spatial reasoning abilities. For example, the US is often described as south of Canada, despite some areas within the US being located to the north. Third, LLMs might prioritize conceptual associations, such as assuming a 'west coast' city to be the westernmost point without accounting for the curvature of the coastline.

To investigate hierarchical bias in large language models (LLMs), we initially conducted a study with ten questions, five challenging questions where bias is likely to be exhibited based on scenarios where humans typically struggle, and five control questions to serve as a baseline for comparison. This study included four models: GPT-3.5 [3], GPT-4, LLaMA 2 [9], and Gemini 1.0 Pro [10], among which GPT-4 demonstrated superior performance. Based on this outcome, we narrowed our focus to GPT-4 for an analysis of the other types of bias. For each type, we formulated four questions, maintaining a balance between challenging and control scenarios. Each question in our study was posed ten times, employing a 'zero-shot' mode to reset the model after every question, ensuring that responses remained uninfluenced by previous interactions. The questions were directly drawn from or inspired by well-known experiments in cognitive psychology literature, as referenced below. This paper extends the work of Fulman et al. [11], who provided evidence of hierarchical bias in LLMs.

2. Results

The outcomes for hierarchical bias (a) are illustrated in Table 1. The models were instructed to determine intercardinal directions between cities, using the prompt: 'What is the intercardinal direction from [City A] to [City B]?' For example, all models consistently (0/10) provided inaccurate directions between Portland and Toronto. We attribute the error to the general northward alignment of Canada relative to the United States (Figure 1a). This observation is

in line with the findings of Stevens and Coupe [4], who observed a similar misperception in humans, presumably influenced by the overarching southward position of the United States relative to Canada, leading most to incorrectly assume Toronto is north of Portland. Conversely, when assessing the relationship between Dallas and San Antonio, both in Texas, the models consistently provided the correct answer (10/10).

GPT-4 demonstrates the highest accuracy in this assessment, achieving a 75% success rate, followed by Gemini with 55%, GPT-3.5 with 53%, and LLaMA-2 at 47%. In scenarios designed to highlight hierarchical bias, GPT-4 distinctly outperforms its counterparts, registering a 50% accuracy rate. In comparison, Gemini scores 34%, GPT-3.5 26%, and LLaMA-2 only 10%. However, when evaluating tasks absent of suspected hierarchical bias, all models exhibit improved performance, with accuracy rates exceeding 75%. The remainder of this study will focus on GPT-4 to explore further biases.

Table 1
Overview of Hierarchical Bias and Model Performance Evaluation

Bias Type	Cities	GPT4	GPT3.5	Gemini	Llama	Correct Answer	Susp. Bias	Bias Ratio
Hierarchical	Portland OR to Toronto CAN	0	0	0	0	Southeast		
	Tijuana MEX to San Antonio TX	3	5	0	0	Southeast		
	Wilmington NC to Philadelphia PA	10	0	10	5	Northeast	Yes	30%
	San Diego CA to Reno NV	2	8	0	0	Northwest		
	Memphis TN to Milwaukee WI	10	0	7	0	Northeast		
	Santo Domingo DOM to Miami FL	10	10	0	10	Northwest		
	Minneapolis MN to Chicago IL	10	10	9	10	Southeast		
	Dallas TX to San Antonio TX	10	10	10	10	Southwest	No	85%
	Havana CUB to Philadelphia PA	10	0	10	8	Northeast		
	San Antonio TX to Houston TX	10	10	9	4	Northeast		
	Model Performance	**75%**	**53%**	**55%**	**47%**			

The outcomes for biases (b) through (d) are presented in Table 2. To demonstrate proximity bias (b), GPT-4 was tasked with evaluating the relative distances between cities, employing the query: 'Which is closer to [City X]: [City A] or [City B]?' For instance, despite New Haven, Connecticut being closer to Philadelphia, Pennsylvania by both road distance (~250km) and great circle measurements, the model consistently determined that Pittsburgh, Pennsylvania is the closer city (~450km) (0/10) (Figure 1b). However, when New Haven is replaced with Johnstown, which is ~390km from Philadelphia in Pennsylvania, the model consistently gives the correct answer (10/10). This possibly reflects a bias of perceiving distances within the state as shorter than across states.

In examining the rotation bias (c), the model was asked: 'Which city is further west, [City A] or [City B]?' For instance, when inquiring which city is further west between Wilmington, North Carolina and Jacksonville, North Carolina, the model mistakenly (2/10) pointed to the latter, possibly reflecting a simplification of the US east coast curvature (Figure 1c). However, it correctly identified the relative westward position when comparing Wilmington to Morehead City, North Carolina, both being coastal cities, possibly suggesting that the presence of a common geographical feature, forces more precise comparisons (10/10).

Table 2
Overview of question types and model performance evaluation

Bias Type	Cities	GPT4	Correct Answer	Susp. Bias	Bias Ratio
Proximity	Philadelphia PA: Pittsburgh PA or New Haven CN	0	New Haven	Yes	0
	Dallas TX: Houston TX or Oklahoma City OK	0	Oklahoma City		
	Dallas TX: Houston TX or Austin TX	10	Austin	No	100%
	Philadelphia PA: Johnstown PA or Pittsburgh PA	10	Johnstown		
Rotation	San Diego CA or Fresno CA	0	Fresno	Yes	10%
	Wilmington NC or Jacksonville NC	2	Wilmington		
	Wilmington NC or Morehead City NC	10	Wilmington	No	75%
	Los Angeles CA or San Francisco CA	5	San Francisco		
Alignment	Monaco MCO to Chicago IL	0	Southwest	Yes	0
	Rome ITA to Philadelphia PA	0	Southwest		
	Lisbon PRT to New York City NY	10	Northwest	No	100%
	Madrid ESP to Boston MA	10	Northwest		

Alignment bias (d) was examined through the query: 'What are the intercardinal directions from [City A] to [City B]?' For example, the model inaccurately determined the direction between Monaco, situated in the southern part of Europe, and Chicago, located in the northern United States (0/10). This error may reflect the common misconception that North America and Europe align on the same east-west axis, while in reality, Europe is predominantly north of the United States [7] (Figure 1d). However, it correctly ascertained the direction from Lisbon to New York City, which is straightforward because Lisbon is indeed to the south of New York City (10/10).

3. Discussion

We report our initial findings from an examination of potential systematic bias in the spatial reasoning capabilities of GPT-3.5, GPT-4, Gemini, and Llama-2. The models show distinct patterns in their performance: They achieve 87% accuracy in questions that do not challenge biases in spatial reasoning, indicating a strong understanding of straightforward geographical relationships. On the other hand, they only achieved a 24% accuracy rate in the questions which highlight these biases.

Our current study draws from the psychology experiments that served as our inspiration; however, it does not offer a statistically valid assessment of how biases in spatial perception impact LLMs. To address this limitation, our future exploration will focus on the proximity bias, denoted as (b), which relates to the tendency to underestimate distances within categories while overestimating distances between them. This analysis will involve querying the models with hundreds of relevant cities and examining their interrelationships.

While our approach may allow us to verify the existence of biases in LLM's spatial reasoning skills, we may not be able to pinpoint the source of these biases – whether they originate from

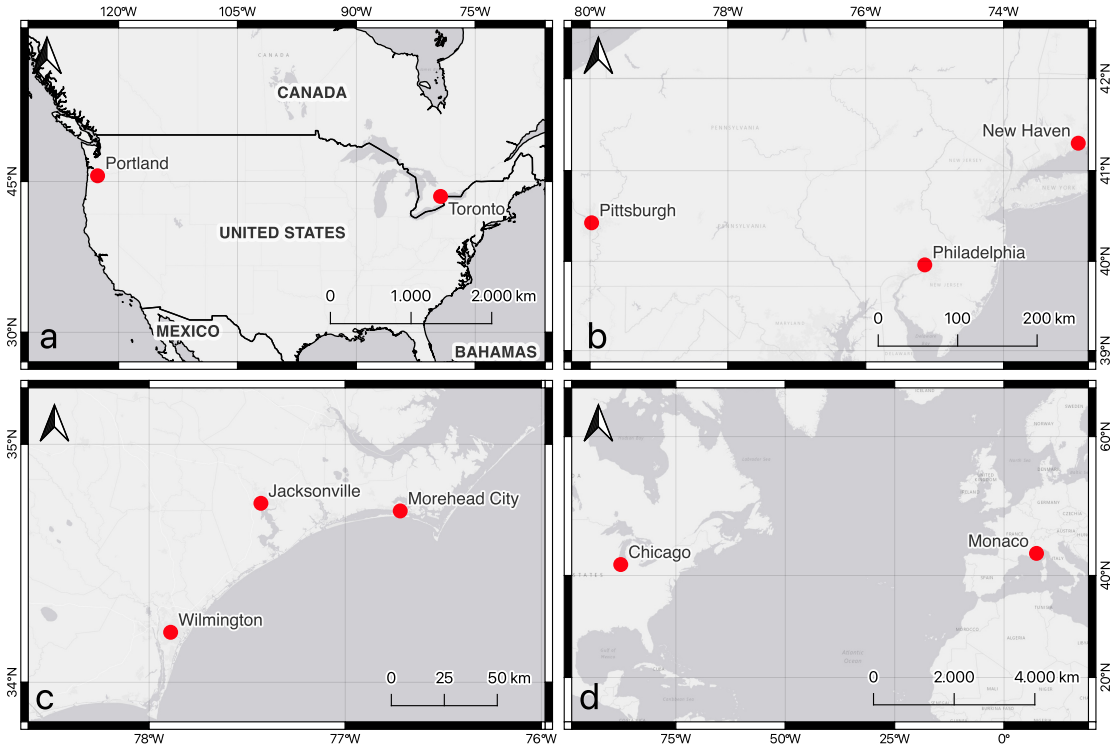

Figure 1: Illustration of cities demonstrating the four types of bias: (a) hierarchical, (b) proximity, (c) rotation, and (d) alignment

learned human errors, generalized geographical input data, or the models' inherent tendencies towards conceptual associations. Nevertheless, we may be able to mitigate these issues. One potential strategy involves training LLMs with datasets explicitly detailing spatial relationships between various locations. Utilizing Natural Language Geographic Data for this purpose could usher in a deliberate development of spatial reasoning skills within these models, enabling them to more accurately comprehend and process geographic relationships. In the next phase of our research, we plan to explore methods for fine-tuning an open-source LLM using such spatially explicit datasets, evaluating its ability to discern intercardinal directions and ultimately enhance its spatial reasoning capabilities.

Acknowledgments

N. Fulman was supported by the Health + Life Science Alliance Heidelberg Mannheim and received state funds approved by the State Parliament of Baden-Württemberg.
A. Memduhoğlu was supported by the Scientific and Technological Research Council of Türkiye (TUBITAK) under the program 2219 (1059B192202917).

References

[1] Y. Zhang, C. Wei, S. Wu, Z. He, W. Yu, Geogpt: Understanding and processing geospatial tasks through an autonomous gpt, arXiv preprint arXiv:2307.07930 (2023).

[2] 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).

[3] OpenAI, Models: Gpt-4 turbo (128k context window) and gpt-3.5 turbo (16k context window), 2023. URL: https://platform.openai.com/docs/models, version November 21, 2023.

[4] A. Stevens, P. Coupe, Distortions in judged spatial relations, Cognitive Psychology 10 (1978) 422–437.

[5] L. P. Acredolo, L. T. Boulter, Effects of hierarchical organization on children's judgments of distance and direction, Journal of Experimental Child Psychology 37 (1984) 409–425.

[6] L. G. Braine, A new slant on orientation perception, American Psychologist 33 (1978) 10.

[7] B. Tversky, Distortions in cognitive maps, Geoforum 23 (1992) 131–138.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need 30 (2017).

[9] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, T. ... Scialom, Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[10] G. Team, R. Anil, S. Borgeaud, Y. Wu, J. B. Alayrac, J. Yu, J. ... Ahn, Gemini: a family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).

[11] N. Fulman, A. Memduhoğlu, A. Zipf, Distortions in judged spatial relations in large language models: The dawn of natural language geographic data?, arXiv preprint arXiv:2401.04218 (2024).