International Cartographic Association
Association Cartographique Internationale

Exploring challenges of Large Language Models in estimating the distance

Mina Karimi [a,*], Krzysztof Janowicz [a]

[a] *Department of Geography and Regional Research, University of Vienna, Vienna, Austria – mina.karimi@univie.ac.at , krzysztof.janowicz@univie.ac.at*

* Corresponding author

Keywords: Large Language Models (LLMs), ChatGPT, distance, spatial reasoning

Abstract:

Generative Artificial Intelligence (GenAI), particularly Large Language Models (LLMs), has experienced significant growth and attention in recent years. These advanced models, exemplified by ChatGPT, have demonstrated remarkable capabilities in generating text and image results using prompts from natural language. The ability to generate coherent and contextually relevant responses has led to widespread applications across diverse domains such as education, transportation, healthcare, law, finance, scientific research, and geography (Dash et al., 2023; Zhao et., 2023).

LLMs, despite only being trained to predict the next token, have shown significant abilities (Bubeck et al., 2023). This has led to questions about what these models have truly learned. One idea is that LLMs gather many correlations between words but don't truly understand coherent model or the data they are trained on. Another idea is that during training, LLMs develop clearer and more understandable models of the data generating process, known as a "world model" (Gurnee & Tagmark, 2023). In essence, ChatGPT goes beyond the traditional GPT model and can respond to a wide range of human queries and incorporate facts from external sources to enhance accuracy and reliability with the use of techniques like retrieval-augmented generation (RAG). However, as ChatGPT becomes more prevalent, there is a growing attention to its geographical perceptions and the accuracy of its outputs in this domain. Questions arise about how well LLMs, trained on vast textual datasets, truly understand geographical information and whether they can provide trustworthy responses of geographical concepts. As these models continue to evolve and find applications in various domains, there is a heightened focus on examining their effectiveness and limitations in handling geographical queries. (Ji & Gao, 2023).

In this paper, we investigate the potential of LLMs, for instance ChatGPT, in understanding geographical concepts such as estimating the distances between cities. We examined its performance for well-known and major cities as well as smaller or lesser-known cities. When asked about the distance between major cities like *Vienna* and *Salzburg* in *Austria* or *Tehran* and *Isfahan* in *Iran*, ChatGPT can provide a reasonable estimation based on its training data. These cities are commonly featured in datasets used for training such models, allowing ChatGPT to generate accurate distances. However, the story changes when it comes to smaller or less prominent cities. For cities that are not as widely known, such as *Zwettl* and *Bad Goisern* in *Austria* or *Kashmar* and *Kesheh* in *Iran*, ChatGPT struggles to provide accurate distances. Due to the limited representation of such cities in its training data, ChatGPT may resort to generating random or incorrect values. These inaccurate distances persist even when we specify the province of these small cities to the language model (e.g., *Kashmar* in *Khorasan Razavi*, and *Kesheh* in *Isfahan)*. The examples are shown in Table 1.

This limitation is especially noticeable when asking for distances between cities with similar names, such as the various *Springfields* in the *United States* and the *Feldkirchs* in *Alsace*, *France*, and *Austria*. Despite being distinct cities, ChatGPT may produce erroneous distances or fail to differentiate between them, leading to ambiguous or incorrect responses. This underscores the need for caution when relying on LLMs for geographical information, especially for lesser-known or similarly named places.

Furthermore, although promising in urban science, the utilization of generative AI also encounters ethical issues like misinformation and bias, sometimes lacking accuracy in portraying compositions and locations under specific circumstances (Jang et al., 2023; Kang et al., 2023). For example, there exists a significant bias in the language used to query distances. When asked in English, ChatGPT may provide more accurate results for well-known cities due to the prevalence of English-language datasets containing information on these cities. However, when asked in Persian, the results vary, as the model's training data may not be as rich or diverse in Persian-language geographical information.

Abstracts of the International Cartographic Association, 7, 68, 2024.
European Cartographic Conference – EuroCarto 2024, 9–11 September 2024, TU Wien, Vienna, Austria.
https://doi.org/10.5194/ica-abs-7-68-2024

In summary, while ChatGPT excels at estimating distances between major cities, its performance diminishes for smaller or less common locations. The language used to query distances also has a noticeable bias, and difficulties arise with locations that have similar names. These limitations highlight the challenges of relying solely on LLMs for accurate and reliable geographical information, especially in diverse linguistic contexts and for lesser-known cities.

Origin	Destination	Google Map Routing (km)	ChatGPT (km)
Vienna	Salzburg	295.4	295
Zwettl	Bad Goisern	211	220
New York City	Springfield	Massachusetts, USA: 230.1 Virginia, USA: 397.5	Massachusetts, USA: 227 Virginia, USA: 365
Vienna	Feldkirch	Vorarlberg, Austria: 633 Alsace, France: 871	Austria: 560 France: 825
Tehran	Isfahan	440	Prompt in English: 445 Prompt in Persian: 455
Kashmar	Kesheh	895	Prompt in English: 110 Prompt in Persian: 57 Prompt in English (Providing States): 465

Table 1. The examples of distances (km) between two cities

References

Dash, D., Thapa, R., Banda, J., Swaminathan, A., Cheatham, M., Kashyap, M., Kotecha, N., Chen, J. H., Gombar, S., Downing, L., Pedreira, R. A., Goh, E., Arnaout, A., Morris, G. K., Magon, H., Lungren, M. P., Horvitz, E., and Shah, N. H. Evaluation of gpt- 3.5 and gpt-4 for supporting real-world information needs in healthcare delivery. *ArXiv*, abs/2304.13714, 2023. URL https://api.semanticscholar. org/CorpusID:258331653.

Gurnee, W., & Tegmark, M. (2023). Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Jang, K. M., Chen, J., Kang, Y., Kim, J., Lee, J., & Duarte, F. (2023). Understanding Place Identity with Generative AI (Short Paper). In 12th International Conference on Geographic Information Science (GIScience 2023). Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Ji, Y., & Gao, S. (2023). Evaluating the effectiveness of large language models in representing textual descriptions of geometry and spatial relations. *arXiv preprint arXiv:2307.03678*.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Kang, Y., Zhang, Q., & Roth, R. (2023). The ethics of ai-generated maps: A study of dalle 2 and implications for cartography. *arXiv preprint arXiv:2304.10743*.

Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min et al. "A survey of large language models." *arXiv preprint arXiv:2303.18223* (2023).

https://doi.org/10.5194/ica-abs-7-68-2024