Are Large Language Models Geospatially Knowledgeable?

Prabin Bhandari
pbhanda2@gmu.edu
Department of Computer Science
George Mason University
Fairfax, Virginia, USA

Antonios Anastasopoulos
antonis@gmu.edu
Department of Computer Science
George Mason University
Fairfax, Virginia, USA

Dieter Pfoser
dpfoser@gmu.edu
Department of Geography and
Geoinformation Science
George Mason University
Fairfax, Virginia, USA

ABSTRACT

Despite the impressive performance of Large Language Models (LLM) for various natural language processing tasks, little is known about their comprehension of geographic data and related ability to facilitate informed geospatial decision-making. This paper investigates the extent of geospatial knowledge, awareness, and reasoning abilities encoded within such pretrained LLMs. With a focus on autoregressive language models, we devise experimental approaches related to (i) probing LLMs for geo-coordinates to assess geospatial knowledge, (ii) using geospatial and non-geospatial prepositions to gauge their geospatial awareness, and (iii) utilizing a multidimensional scaling (MDS) experiment to assess the models' geospatial reasoning capabilities and to determine locations of cities based on prompting. Our results confirm that it does not only take larger but also more sophisticated LLMs to synthesize geospatial knowledge from textual information. As such, this research contributes to understanding the potential and limitations of LLMs in dealing with geospatial information.

CCS CONCEPTS

• **Computing methodologies** → *Natural language generation.*

KEYWORDS

Large Language Models, Geospatial knowledge, Geospatial awareness, Geospatial Reasoning

ACM Reference Format:
Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. 2023. Are Large Language Models Geospatially Knowledgeable?. In *The 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23), November 13–16, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589132.3625625

1 INTRODUCTION

The recent proliferation of pretrained large language models (LLMs) like GPT-3 [2] and their impressive performance on several downstream tasks has led the natural language processing (NLP) community to consider the implicit knowledge these models may contain in their parameters. Authors have shown that LLMs can function, to an extent, as knowledge bases [8], since they store various types of knowledge, such as common sense, relational, and linguistic aspects in their parameters [3, 9]. This paper explores whether and to what extent **geospatial knowledge** is encoded in LLMs and whether such models have **geospatial awareness**. Finally, we examine the models' **geospatial reasoning** potential. Geospatial knowledge includes the factual understanding of geographic data such as location, distance, and area. Geospatial awareness is concerned with the ability to perceive and comprehend geographical information. Finally, geospatial reasoning is the use of geospatial knowledge and awareness for informed decision making.

To evaluate LLMs with respect to their geospatial knowledge, awareness, and reasoning capabilities, we conducted the following experiments. First, we probe the LLMs for actual geo-coordinates of cities. This should provide us with an idea about their concrete geospatial knowledge. To assess their geospatial awareness, we evaluate whether geospatial prepositions such as "near" translate into smaller distances when used in sentences to generate nearby cities as opposed to a control scenario which simply uses the conjunction "and". Last, to gauge the geospatial reasoning potential of LLMs, we perform a multidimensional scaling(MDS) [1] experiment, in which we compare the predicted layout of cities using real distances to a distance measure derived from LLMs.

Our findings reveal that LLMs are becoming more adept at handling and comprehending geospatial data, as evidenced by their encoded geospatial knowledge and subsequent geospatial awareness while generating texts. Our results also show the possibility of using LLMs in geospatial reasoning tasks.

2 METHODOLOGY

Our methodology involves three different tasks to assess different aspects of the geospatial capabilities of LLMs.

For the first task, evaluating the geospatial knowledge encoded within LLMs, the objective is to correctly predict the locations and coordinates of cities. The second task, assessing geospatial awareness, analyzes the expressions generated by LLMs when leveraging geospatial prepositions vs. generic expressions, e.g., "near vs. "and" by comparing their resulting respective distances, i.e., are cities that use "near" actually closer than when using "and"? Lastly, to assess the LLMs' usefulness for geospatial reasoning, we devise a problem where the goal is to predict the locations of cities based on the relative distances between cities. We generate two "constellations", one which uses the actual distances, compared to another one that uses LLM-derived distances.

The LLMs that we use are OPT (6.7B and 13B), LLaMA model (7B and 13B), and Alpaca model (instruction-tuned from 7B LLaMA model) for your geospatial knowledge task. Based on the results from this task we only use the 13B LLaMA for our other tasks.

SIGSPATIAL '23, November 13–16, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0168-9/23/11.
https://doi.org/10.1145/3589132.3625625

Table 1: Mean Error Distances (km) for Coordinate Prediction

Model	Template	Prompt	Error (km)	P-Rate (%)
Word2Vec	-	-	**2612**	-
BERT-L	-	-	3077	-
GPT-2	-	-	4498	-
LLaMA (7B)	2	0-shot	521	10
LLaMA (7B)	2	3-shot	1469	99
LLaMA (13B)	1	0-shot	864	89
LLaMA (13B)	1	3-shot	1069	99
LLaMA (13B)	2	0-shot	**386**	31
LLaMA (13B)	2	3-shot	1634	99
Alpaca (7B)	1	-	1799	76
Alpaca (7B)	2	-	2158	99

For the geospatial awareness task, we employ tok-k sampling with $k = 100$ and a temperature of 0.9, while using beam search with five beams for other tasks.

3 MEASURING GEOSPATIAL KNOWLEDGE

This first experiment simply probes LLMs to determine the coordinates (latitude and longitude) of cities. This task serves as an indicator of the extent to which LLMs encode geospatial knowledge.

3.1 Experimental Setup

Prompting, introduced by Brown et al. [2], refers to appending a few sample input-output along with a textual prompt to a pretrained LLM, which is then expected to provide a relevant completion of this input based on the sample inputs and outputs. This approach is sometimes referred to as in-context learning. In our experiments, we use prompts as : *The geo-coordinates of Kathmandu are ...*

We use both 3-shot and zero-shot inference for the location prediction of cities. The cities that we use as examples while prompting are selected randomly from a list of 3,527 cities having a population greater than 100k in MaxMind database[1]. We experiment with an additional prompt template, where *"geo-coordinates"* in our prompt is replaced by *"latitude and longitude"*. For the case of the instruction-following Alpaca model, we provide instructions to provide the geo-coordinates of cities with two templates similar to those above using the instruction template format of the model.

3.2 Results and Discussion

The results of our coordinate prediction task in Table 1 consist of the first three rows from the results reported in Liétard et al. [5] and subsequent rows from our work. The "P-rate" column refers to the prediction rate which indicates the LLM's success rate in correctly generating another city.

Previous results by Liétard et al. [5] imply limited encoded geospatial knowledge in LLMs, with comparisons to Word2Vec [7] favoring Word2Vec's performance. Liétard et al. [5] also suggests that larger LLMs might perform better for tasks related to geographic information. This idea is well supported by our results, with the caveat that not all LLMs perform equally well. Despite having the same number of parameters, the OPT and LLaMA models produced vastly different results in this coordinate prediction task. The LLaMA model was much better than OPT, that is why we have not included it in Table 1. This demonstrates the effect of the model architecture, design decisions, and the pre-training dataset on the model's performance. Particularly in the zero-shot setting, *the 13-billion variant of LLaMA showed a remarkable 85% reduction in prediction error for coordinate prediction compared to the best baseline(Word2Vec).* Overall, larger variants of LLMs generally do perform better. Our research also shows that different prompts produce different outcomes. This leaves room for improvement by employing continuous prompts [4, 6].

Finally, our results show that prompting can improve prediction rate, approaching Alpaca model performance. However, their performance is slightly lower than the zero-shot setting, which is somewhat counterintuitive. This difference mat arise from LLMs potentially lacking relevant geo-coordinate examples during pre-training, leading to lower prediction rates in zero-shot scenarios, but often yielding higher accuracy. In the 3-shot setting or instruction-following setup, LLMs are compelled to provide predictions, sometimes inaccurately. This shows that we can extract the geospatial knowledge encoded in LLMs more efficiently with proper prompt engineering and exposure to diverse geospatial datasets during LLM pre-training

In conclusion, our results show that LLMs are becoming more adept at encoding geospatial knowledge. Furthermore, we see that the zero-shot setting outperforms the few-shot setting in terms of accuracy, partly because the few-shot setting leads to higher prediction rates.

4 MEASURING GEOSPATIAL AWARENESS

Geospatial awareness refers to the perception of space and the use of spatial information during everyday activities. This idea also applies to generative language models, i.e., the degree to which LLMs capture geospatial information and how this is evident when generating text. To assess the geospatial awareness of LLMs, we utilize geospatial prepositions, i.e., prepositions that describe spatial relationships between objects or places in a geographical setting.

4.1 Experimental Setup

We want the LLM to generate sentences such as `"<City-A> is near <City-B>"`, where `"<City-A> is near"` is passed as context. Assuming that the model has geospatial awareness, `<City-B>` should be geographically close to `<City-A>`.

In our experiments, we contextualize the LLM input with a geospatial preposition and evaluate the output the LLM generates and prompt the model as: *Albany is near ...*

We analyze whether the generation of <City-B> given the context of "<City A> is near" is affected by the presence of the preposition "near" or not. In addition to "near", we also use the prepositional phrases "close to" and "far from". We contrast the results of the above experiments with a control experiment where the geospatial preposition is replaced with the conjunction "and". We prompt the

[1] https://www.kaggle.com/datasets/max-mind/world-cities-database

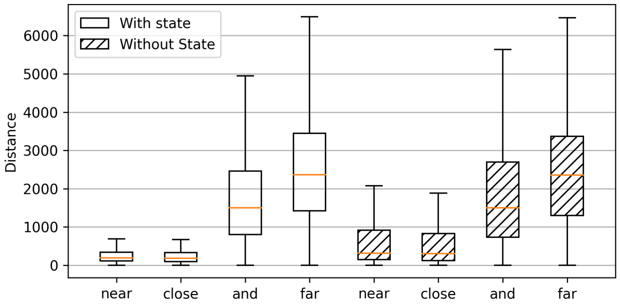
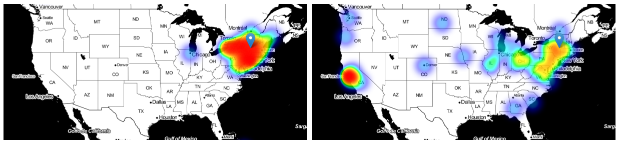


Figure 1: Predicted city distances with different prepositions.



(a) Albany, New York is near ... **(b) Albany is near ...**

Figure 2: Heatmaps of the places generated with “near". State information enhances city disambiguation.

model in both zero-shot and three-shot settings. Additionally, we also append the state of the city in all our inputs.

We curate a list of 93 cities in the contiguous United States to perform an in-depth analysis of the geospatial awareness of LLMs. The list balances city size with coverage of most of the contiguous states. We use ten different prompts per city, where each prompt is created by randomly selecting a city and its closest city in our list. We generate fifty samples for each prompt.

4.2 Results and Discussion

Figure 1 displays a box plot of the statistics of the actual distances between the generated places and the original city in our experiment. The visualization makes it evident that the use of geospatial prepositions in the sentences has an impact on the generated cities. The sentences contextualized with geospatial prepositions that indicate close proximity, such as “near" and “close to", yielded cities that are physically closer to the original city. Conversely, when the context was “far”, a geospatial preposition indicating distant location, the generated cities tended to be farther away from the original city. For our control experiment (with the non-geospatial word “and”), the observed differences in the distances of the predicted cities provide compelling evidence of the geospatial awareness of LLMs. The varying magnitude of differences in the distances of predicted cities for the different geospatial prepositions further reinforces the notion of geospatial awareness in LLMs.

Figure 2 provides a specific example showing that the inclusion or exclusion of the state name in the city names influences the generated cities. The generated cities are occasionally further away from the source cities when state names are not included in the prompt. We believe that this discrepancy is due to the limitation of LLMs in resolving the exact location of a city when the state information is missing: the lack of state information may lead LLMs to confuse cities with the same name (disambiguation).

In conclusion, our results provide compelling evidence that LLMs are indeed geospatially aware.

5 LLMS AND GEOSPATIAL REASONING

Geospatial reasoning refers to understanding and analyzing geospatial information to draw conclusions and make decisions. In order to assess the usefulness of LLMs for this task, we devise an experiment to predict the locations of cities using dissimilarity measures, such as for example distances between the cities.

5.1 Experimental Setup

We use dissimilarity measures to establish a 2-dimensional geometric representation of cities. We accomplish this through the application of multi-dimensional scaling (MDS) [1]. Specifically, we begin with a list of cities with known locations and with a test city whose location and coordinates we want to predict. Knowing the distance between all cities (including the test city), we then use a least-squares estimation of transformation parameters between two point patterns [10] to get the transformation matrix that maps the 2-dimensional geometric space coordinates generated by MDS to actual geo-coordinates using the cities for which the geo-coordinates are known. Finally, we use this transformation matrix to determine the geo-coordinates for the test city.

We use actual distances as a benchmark for dissimilarity measures and the co-occurrence counts between each city pair as our baseline measure to establish a comparative reference point. For any value that is a measure of similarity like Co-occurrence, we consider its reciprocal value to convert it into a dissimilarity measure. By utilizing a dissimilarity measure between cities to predict their geo-location, our designed task illustrates a practical application of geospatial reasoning. We extract diverse measures of dissimilarity from the LLM and conduct a comparative analysis against our predefined benchmark and baseline. The *dissimilarity measures* include the following: (i) **Predicted Distance**: We predict the distances between each city pair in a zero-shot setting from LLM , and (ii) **Generation Frequency**: We count the generation frequency of each city in relation to the remaining cities from our measuring geospatial awareness task.

We use the same list of 93 cities in the contiguous United States presented in Section 4. Each city in our dataset is considered a test city for which we want to predict its coordinates and we use the remaining cities to sample cities with known locations. Based on the results of Sec:4, we include the state names in the prompts.

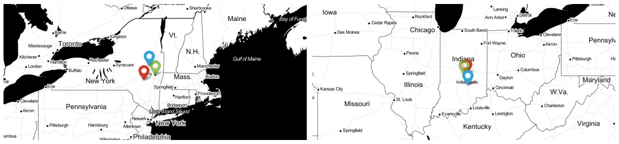
5.2 Results and Discussion

We present our location prediction task results in Table 2, including two mean error distances. The first estimates geo-coordinates by considering all other cities in contiguous US, while the second considers cities within one of the nine different US Census Bureau-designated divisions. To establish a benchmark, we calculate the minimum attainable error using actual distances. We also employ a “random” baseline, averaging errors from ten random predictions for test city locations.

Our results indicate improved accuracy when focusing on smaller geographical regions (using divisions) rather than considering the entire contiguous US. This can be attributed to the inherent ease

Table 2: Mean error distances for geo-coordinate predictions of cities from dissimilarity measure using MDS

		Mean error distance (km)	
	Measure	Contiguous	Divisions
	Actual distance	**190.41**	**56.78**
	Random distance	1440.01	483.15
i	Co-occurrence count	1237.01	425.64
ii	"and" generations count	1359.51	453.73
iii	"near" generations count	750.66	328.91
iv	"close to" generations count	782.22	324.06
v	"far from" generations count	1383.23	455.76
vi	Predicted distance	**346.65**	**177.41**



(a) Albany, New York **(b) Indianapolis, Indiana**

Figure 3: Real and predicted locations: Green (Real), Blue (predicted from actual), Red (predicted from predicted).

in predicting closer cities. Further, the errors associated with co-occurrence counts (row i) and "and" generation counts (row ii) are similar to random distance, as they don't reflect proximity-based similarity. This trend is more prominent for the "far from" generation counts (row v). Conversely, "near" and "close to" counts (rows iii and iv) exhibit much lower errors, supporting LLM's geospatial awareness and reasoning ability. However, due to generation count sparsity, they don't closely match predicted distances. Predicted distances (row vi) yield results much closer to actual distances and far better than random guesses.

It is important to note that our task's goal was not to assess LLMs' geo-coordinate prediction accuracy but to evaluate their geospatial reasoning capabilities, potentially useful for predicting relative city orientation rather than exact locations. Figure 3 shows the actual (Blue) and predicted locations using both actual distances and predicted distances (Green and Red) for two cities. As shown in the figure, predicted coordinates closely align with actual values in some locations and exhibit slight deviations in others. These disparities would only be marginally noticeable at city scales when focusing on city orientation rather than its precise location. While the predicted locations based on predicted distances differ from actual-distance-based predictions, they remain reasonably close.

In conclusion, our results demonstrate the potential use of LLMs for geospatial reasoning tasks. While it is important to note that the values produced by LLMs may not precisely match the actual values, they still show a remarkable level of similarity. Thus, LLMs have great potential for supporting humans in geospatial reasoning and analysis tasks with targeted fine-tuning tailored to a certain use case.

6 CONCLUSIONS AND FUTURE WORK

This work demonstrates notable improvements in LLMs' ability to handle geospatial data, due not only to the increasing size of models, but also facilitated by novel techniques such as instruction tuning. We show that LLMs encode geospatial knowledge, which can be leveraged for tasks that are simple, such as obtaining coordinates and locations for cities by probing those models, or more complex, such as a quantitative understanding of spatial prepositions. All this information can be extracted and utilized using the proper "querying" techniques such as prompting. We demonstrate that LLMs show potential for geospatial reasoning tasks, but further enhancements are needed to meet the desired accuracy and performance levels. Overall, LLMs have come a long way and now exhibit geospatial awareness when generating text.

Future research will focus on examining the practical applicability of LLMs in real-world applications involving geospatial data, as well as utilizing even larger models, and doing so for languages other than English.

Experiment Code: https://github.com/prabin525/spatial-llm.

ACKNOWLEDGMENTS

This work has been supported by the National Science Foundation Grant No. IIS-2127901. Additionally, this work was supported by resources provided by the Office of Research Computing, George Mason University and by the National Science Foundation (Awards Number 1625039 and 2018631).

REFERENCES

[1] Ingwer Borg and Patrick JF Groenen. 2005. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Jeff Da, Ronan Le Bras, Ximing Lu, Yejin Choi, and Antoine Bosselut. 2021. Analyzing commonsense emergence in few-shot knowledge models. arXiv:2101.00297.

[4] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 4582–4597. https://doi.org/10.18653/v1/2021.acl-long.353

[5] Bastien Liétard, Mostafa Abdou, and Anders Søgaard. 2021. Do Language Models Know the Way to Rome?. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 510–517. https://doi.org/10.18653/v1/2021.blackboxnlp-1.40

[6] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

[8] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2463–2473. https://doi.org/10.18653/v1/D19-1250

[9] Tara Safavi and Danai Koutra. 2021. Relational World Knowledge Representation in Contextual Language Models: A Review. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1053–1067. https://doi.org/10.18653/v1/2021.emnlp-main.81

[10] Shinji Umeyama. 1991. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 13, 04 (1991), 376–380.