MAKING GEOGRAPHIC SPACE EXPLICIT IN PROBING MULTIMODAL LARGE LANGUAGE MODELS FOR CULTURAL SUBJECTS

Zilong Liu
Department of Geography and Regional Research
University of Vienna
1010 Vienna, Austria
zilong.liu@univie.ac.at

Kitty Currier
Department of Geography
University of California, Santa Barbara
Santa Barbara, California 93106, USA
kcurrier@ucsb.edu

Krzysztof Janowicz
Department of Geography and Regional Research
University of Vienna
1010 Vienna, Austria
krzysztof.janowicz@univie.ac.at

ABSTRACT

The ability of large language models (LLMs) to recall factual knowledge has driven recent work to probe them for geo-diverse common sense. However, currently most benchmarks fail to consider geographic context in their probing. Due to LLMs' increasing multimodality, this problem is particularly prominent because we are now exposed to the multi-perspective representations for a single cultural subject. In this provocation, we draw insights from geography and acknowledge their value in AI evaluation. More concretely, we describe potential risks, such as the modifiable areal unit problem, in current benchmarks. In addition, we envision knowledge graph (KG) construction as a solution to facilitate geospatially large-scale and granular probing, and advocate a reorientation of the focus of geospatial KGs to cultural places and a need of geospatial scoping of statements in KG completion. To sum up, we propose to make geographic space explicit in probing multimodal LLMs for cultural subjects.

Large language models (LLMs) are able to encode a large amount of relational knowledge from their training data, making themselves competent when compared with existing knowledge bases (Petroni et al., 2019). Recent work has tried to develop a technique to probe the degree to which a LLM exhibits *geo-diverse common sense*—presumed shared knowledge that can vary across regions due to cultural differences (Yin et al., 2022). This work examined how a prompt's language–Chinese, English, Hindi, etc.–influenced pre-trained language models' responses to topics like "the color of a wedding dress", and provided linguistic insights such as how much common sense expressed in native and non-native languages they may have encoded. Meanwhile, LLMs are becoming multimodal in addition to being multilingual, making it possible to probe factual knowledge by requesting output in formats besides text, such as image. Along with many other works, Qadri et al. (2023) have attempted to discover cultural limitations (e.g., *regional cultural defaults*) in text-to-image generation models with a community-based study on South Asia. There are also more generative AIs that support visual question answering. Therefore, it is necessary to evaluate in a *geospatially* scalable manner their ability to understand images besides their generative capacity. This will provide us with more holistic representations that involve multiple perspectives about a cultural subject.

Though the idea of geo-diverse common sense could be an oxymoron, it raises questions regarding how geographic space is represented in probing, particularly on how benchmarks should be constructed for probing cultural subjects. It is undoubtedly difficult to make benchmarks that are both applicable worldwide yet granular enough to account for the modifiable areal unit problem (Openshaw, 1984), also known as aggregation bias (Suresh & Guttag, 2019) in machine learning. In short, a probing task involving five countries is inadequate, and one that considers only the country as the geographic unit for analysis is as deceiving as gerrymandering. Without considering geographic space, a prompting template would also suffer from place name ambiguity (Leidner, 2007), the problem that a single place name may refer to many locations.

From the perspective of cultural geography, cultures are "locatable, specific phenomena" (Crang, 2013). This means a cultural subject can be grounded with a spatial reference, but it does not necessarily need to have a geometry as a latitude-longitude pair. A cultural subject can be a *place of interest*, such as World Heritage sites that "belong to all the peoples of the world, irrespective of the territory on which they are located"[1], or a *concept* described with a collection of statements, each of which has a respective geographic scope. The former sees space and place (Tuan, 1979) from a humanistic perspective, while the latter aims to map the spatial heterogeneity of the validity of an assertion. In addition, as most large-scale benchmarks were created from symbolic knowledge that can be transformed to prompts in natural language, knowledge graphs (KGs), as the data providers behind symbolic AI, are playing an important role in probing subsymbolic systems. Regarding KG construction, we should reorient the focus of geospatial KGs (Mai et al., 2022) to cultural places among natural and artificial features, and complement KG-completion studies that by now have only focused on temporal scoping (Rula et al., 2019) with geospatial-scoping research that could mitigate inferential bias (Janowicz et al., 2018). The explicitness of geographic space will provide us with a *vocabulary* when speaking with LLMs, leading us to *think geographically* (Jackson, 2006) by considering scale effects, proximity, and outsider gazes forming beyond certain constructions.

To sum up, we propose to **make geographic space explicit in probing multimodal LLMs for cultural subjects**, with an exploration of potential risks and opportunities. We think the value of making geographic space explicit is not limited to conventional geography-adjacent disciplines such as urban planning or environmental science. Meanwhile, we consider our ideas a complement to a research agenda on *representation in AI evaluations* (Bergman et al., 2023). An implicit geographic scope is often associated with impacted communities, be they *the Global South* or *Native Americans*. In addition, when it comes to evaluating AI systems, *representative of where?* is as important as *representative of when?*, since a globally inclusive AI should value all cultures equally.

Acknowledgement

The authors would like to thank Jiaqi Liang and Meilin Shi for sharing valuable feedback when forming the initial idea of this work.

References

A Stevie Bergman, Lisa Anne Hendricks, Maribeth Rauh, Boxi Wu, William Agnew, Markus Kunesch, Isabella Duan, Iason Gabriel, and William Isaac. Representation in ai evaluations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 519–533, 2023.

Mike Crang. *Cultural geography*. Routledge, 2013.

Peter Jackson. Thinking geographically. *Geography*, 91(3):199–204, 2006.

Krzysztof Janowicz, Bo Yan, Blake Regalia, Rui Zhu, and Gengchen Mai. Debiasing knowledge graphs: Why female presidents are not like female popes. In *ISWC (P&D/Industry/BlueSky)*, 2018.

Jochen L Leidner. Toponym resolution in text: annotation, evaluation and applications of spatial grounding. In *ACM SIGIR Forum*, volume 41, pp. 124–126. ACM New York, NY, USA, 2007.

Gengchen Mai, Yingjie Hu, Song Gao, Ling Cai, Bruno Martins, Johannes Scholz, Jing Gao, and Krzysztof Janowicz. Symbolic and subsymbolic geoai: Geospatial knowledge graphs and spatially explicit machine learning. *Trans GIS*, 26(8):3118–3124, 2022.

Stan Openshaw. The modifiable areal unit problem. *Concepts and techniques in modern geography*, 1984.

[1] https://whc.unesco.org/en/about

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 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, 2019.

Rida Qadri, Renee Shelby, Cynthia L Bennett, and Emily Denton. Ai's regimes of representation: A community-centered study of text-to-image models in south asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 506–517, 2023.

Anisa Rula, Matteo Palmonari, Simone Rubinacci, Axel-Cyrille Ngonga Ngomo, Jens Lehmann, Andrea Maurino, and Diego Esteves. Tisco: Temporal scoping of facts. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 959–960, 2019.

Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2(8), 2019.

Yi-Fu Tuan. Space and place: humanistic perspective. In *Philosophy in geography*, pp. 387–427. Springer, 1979.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. Geomlama: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2039–2055, 2022.