Measuring Geographic Diversity of Foundation Models with a Natural Language–based Geo-guessing Experiment on GPT-4

Zilong Liu [1], Krzysztof Janowicz[1,2], Kitty Currier [2], and Meilin Shi [1]

[1]Department of Geography and Regional Research, University of Vienna, Austria
[2]Department of Geography, University of California, Santa Barbara, USA

Correspondence: Zilong Liu (zilong.liu@univie.ac.at)

Abstract. Generative AI based on foundation models provides a first glimpse into the world represented by machines trained on vast amounts of multimodal data ingested by these models during training. If we consider the resulting models as knowledge bases in their own right, this may open up new avenues for understanding places through the lens of machines. In this work, we adopt this thinking and select GPT-4, a state-of-the-art representative in the family of multimodal large language models, to study its geographic diversity regarding how well geographic features are represented. Using DBpedia abstracts as a ground-truth corpus for probing, our natural language–based geo-guessing experiment shows that GPT-4 may currently encode insufficient knowledge about several geographic feature types on a global level. On a local level, we observe not only this insufficiency but also inter-regional disparities in GPT-4's geo-guessing performance on UNESCO World Heritage Sites that carry significance to both local and global populations, and the inter-regional disparities may become smaller as the geographic scale increases. Morever, whether assessing the geo-guessing performance on a global or local level, we find inter-model disparities in GPT-4's geo-guessing performance when comparing its unimodal and multimodal variants. We hope this work can initiate a discussion on geographic diversity as an ethical principle within the GIScience community in the face of global socio-technical challenges.

Keywords. GPT-4, foundation models, geographic features, geographic diversity, knowledge probing

1 Introduction

Like humans, machines are capable of learning from observations to draw inferences. However, if we do not fully understand the components and nature of the geo-data landscape, naively feeding these data to machines for training, validation, and testing purposes could yield unexpected and undesired results. In a pioneering work in image classification, Shankar et al. (2017) conducted a stress test on the generalizability of two classifiers pre-trained on two of the most commonly used image benchmark datasets. For images crowdsourced from Hyderabad, India, neither classifier could recognize well categories like *groom* and *bridegroom*. Also, the classifier trained on one dataset showed poorer performance on web images from the Global South, e.g., Ethiopia. Such failures could be attributed to a more *Western* representation bias exhibited by both benchmark datasets. Situating GIScience in the current AI4Science[1] trend, we must ask ourselves: Are these models being developed and used for knowledge discovery for the benefit of all, irrespective of where we are or where we come from (Janowicz, 2023)?

The issues of geographic diversity exist not only in computer vision, but also in natural language processing tasks such as geoparsing (Liu et al., 2022). Just as we have realized this fact, the geo-data landscape is facing a disruption brought by the release of ChatGPT as a recent breakthrough in foundation models (Bommasani et al., 2021). More recent large language models (LLMs) also support modalities such as images, greatly improving text-to-image generation and visual question answering. This success in multimodality is significant for the next generation of GeoAI models that could also be pre-trained with geo-data ranging from location descriptions to remote sensing and street-level images, and from vector data to cartographic maps. However, such models would still suffer from a lack of geographic diversity when learning latent spatial representations in a task-agnostic manner (Mai et al., 2022). Additionally, more and more geo-data could become generated by machines at scale. On HuggingFace, there are 43,616, 14,864, and 354 models for text,

---

[1]https://www.microsoft.com/en-us/research/lab/microsoft-research-ai4science

text-to-image, and image-to-text generation, respectively[2], which can be further deployed and fine-tuned for various purposes. Currently, it costs only $0.00025/1k characters for inputs and 20 times the price for outputs when using Gemini Pro, one of the state-of-the-art multimodal closed-source models[3]. The increasing accessibility of generative AI may foster a feedback loop, where content created by these models is used to train subsequent generations. This raises concerns about the potential to perpetuate and amplify biases present in current and future models.

In this short paper, we examine the geographic diversity—or lack thereof—of GPT-4[4], the state-of-the-art multimodal LLM in OpenAI's GPT series. Janowicz (2023) suggested that what an LLM reveals is a mirror of the world through multiple distortions, e.g., one from our observed world to the digital world and another from the sampled world to the learned (and possibly debiased) world, embedded in high-dimensional vector space. Our work uses this analogy to guide the investigation into the geographic diversity of GPT-4, in the process examining what it means for a foundation model to be called geographically diverse. The main subject of our investigation is the collection of **geographic features**[5] that constitute gazetteers referred to as the *vocabulary* of geography (Jackson, 2006). This subject is different from previous studies that may fall into an environmental-determinism trap, as they tend to attribute local machine-learning failures simply to data bias against a studied area. Also, previous work ignores the modifiable areal unit problem (Openshaw, 1984), most often using country-level differences in data distribution and model performance as the sole indicator of geographic diversity. Stemming from the *platial* root of GIScience, we consider that the notion of geographic diversity has another facet, i.e., how well geographic features are represented. These features could be areas where a concept holds true but shifts, physical features that extend across the landscape, or human-made sites that carry historical and cultural meaning. In addition to countries, other kinds of relevant geographical units could be used when assessing geographic diversity.

We approach this notion of geographic diversity centered around the *extension* (i.e., the instances to which a category applies) of geographic feature types, and we believe it is necessary not only to record where models would fail but also to develop innovative ways of assessing geographic diversity. Therefore, we design a natural language–based geo-guessing experiment, and suggest using its performance as an indicator. During the experiment, we mask the geographic feature mentioned in a piece of text and ask GPT-4 to supply its actual name.

---

[2]Retrieved from https://huggingface.co/models on January 10, 2024

[3]https://blog.google/technology/ai/gemini-api-developers-cloud

[4]https://openai.com/research/gpt-4

[5]https://wiki.gis.com/wiki/index.php/Geographic_feature

2 Related work

Zhao et al. (2021) were among the first to try to theorize about the intersection of generative AI, GIScience and the broader discipline of geography. They raised the problem of *deep fake geography*, which situates fake geography (e.g., location spoofing or the fact that maps could tell lies) in the deep-learning era, and conducted an empirical study by using generative adversarial networks to inject landscape features from two other cities into satellite images of Tacoma in Washington, United States. As the resulting images appear to be authentic, the authors later developed detection models using visual and frequency-domain features. In the same work, it was also predicted that deep fakes would become an inevitable part of our society, and therefore, how to understand the fast emergence and negative impacts of associated techniques remains a key question.

Interestingly, the rapid progress in LLMs makes it important to look at generative AI as not merely a data generator but as a knowledge base. Petroni et al. (2019) conducted a fill-in-the-blank cloze test on a wide range of pre-trained language models including BERT[6], an early language model using the Transformer architecture which forms the fundamental building block of today's LLMs. They found that BERT can store relational knowledge in its training data and recall factual, commonsense knowledge without fine-tuning.

More recent work that involves knowledge extraction indicates that geographic knowledge, as a kind of specialized knowledge, is encoded in these models, as well. Liétard et al. (2021) designed three probing tasks about coordinates, population sizes, and neighboring countries, respectively. As the model size increased, more geographic knowledge was found to be learned. Similarly, Bhandari et al. (2023) focused directly on LLMs and probed for coordinates of cities. They found that LLaMA[7] in zero-shot settings can outperform LLaMA in few-shot settings. In addition, they discovered that LLMs have the ability to predict a place based on contextual information (containing an input place and a spatial preposition) and to achieve distance-based spatial reasoning about cities. Jang et al. (2023) retrieved textual responses (structured as bullet points) from ChatGPT and street-level images from DALL·E 2[8] to study the place identity of 31 cities. Then, they examined the semantic similarity between the place identity from the perspective of the models and the place identity embedded in two ground-truth text and image datasets. The results showed that ChatGPT and DALL·E 2 can represent salient features of cities.

These post-BERT works suggest that the usage of generative AI in the form of LLMs should not be limited to

---

[6]http://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

[7]https://ai.meta.com/blog/large-language-model-llama-meta-ai

[8]https://openai.com/dall-e-2

content generation. Using GPT-4 as an example, we focus on its learned representation (rather than reasoning) about geographic features beyond administrative features, e.g., cities or countries. We probe it for factual knowledge in the form of unstructured texts rather than triples. In addition, our experiment differs from mainstream probing techniques that query about feature attributes. Instead, we query GPT-4 about a feature, itself, based on the assumption that contextual words are geo-indicative.

3 Ground-Truth Data Acquisition

Our ground-truth corpus is retrieved via SPARQL queries from DBpedia[9]. DBpedia is currently one of the largest open knowledge bases that uses Semantic Web and Linked Data technologies to extract structured data from Wikipedia (Lehmann et al., 2015). We select geographic features that belong to subclasses of the `dbo:Place` category and subsequent subclasses, as well. This selection includes a subset of all geographic features that exist in DBpedia, in which other classes, such as `dbo:ArchitecturalStructure`, also contain relevant features.

As our work does not explicitly involve the probing of multilingual knowledge of GPT-4, we retrieve only English abstracts which, however, may contain non-English feature names. Features that lack an English abstract and that lack mentions of their names in the abstract are omitted from our study. These additional classes are not considered in this work. Figure 1 shows the retrieval workflow, in which the first step is to retrieve `dbo:Place` subclasses and subsequent subclasses, and the second step is to retrieve the name and the abstract of an instance.

4 Geo-guessing Experiment

DBpedia abstracts allow us to conduct a geo-guessing experiment on GPT-4. We use two GPT-4 variants, `gpt-4-1106-preview` and `gpt-4-vision-preview`. Both models were trained with data up to April 2023. Compared with the `gpt-4-1106-preview` (that was the GPT-4 Turbo model before the more recent release of `gpt-4-0125-preview`), `gpt-4-vision-preview` has the additional ability to understand images, and therefore, `gpt-4-vision-preview` is multimodal. We probe both models in zero-shot settings and set the temperature (i.e., the randomness in the output) to 0. No candidate answer is provided for the model in the experiment, meaning that it is an open-ended question-answering task. Figure 2 shows an example of how the experiment can be achieved in the OpenAI Playground[10]. The system

[9]https://www.dbpedia.org
[10]https://platform.openai.com/playground

prompt is `Return only the name of XX in the given paragraph`. The user prompt is an abstract that masks the target feature as `XX`. In this example, `gpt-4-1106-preview` outputs `Gulf of Thailand` as the correct answer. It is also worth noting that as GPT-4 uses *both publicly available data (such as Internet data) and data licensed from third-party providers* (Achiam et al., 2023), its training data may include DBpedia as an open knowledge source. Therefore, we assume that GPT-4 should output the precisely correct answer if it memorizes the corresponding parts of its training data.

5 Evaluation Results

Our current experiment involves four subclass types, `dbo:Valley`, `dbo:Bay`, `dbo:Sea`, and `dbo:WorldHeritageSite`. The feature types `dbo:Bay` and `dbo:Sea` are subclasses of `dbo:BodyOfWater`, and `dbo:BodyOfWater` and `dbo:Valley` are subclasses of `dbo:NaturalPlace`. Both `dbo:NaturalPlace` and `dbo:WorldHeritageSite` are subclasses of `dbo:Place`. Figure 3 illustrates a DBpedia geographic feature-type hierarchy, in which the grey circle represents `dbo:Place` subclasses excluded from our current work. In total, there are 15 `dbo:Valley`, 40 `dbo:Bay`, 152 `dbo:Sea`, and 981 `dbo:WorldHeritageSite` instances, respectively, used in our experiment.

5.1 Analysis Results on a Global Level

First, we measured the geo-guessing performance as the percentage of features correctly named by GPT-4. Table 1 shows the evaluation results by model and feature type. For each feature type, `gpt-4-vision-preview` correctly predicted fewer than half of the total features. The model `gpt-4-1106-preview` correctly predicted slightly more than half of the features belonging to `dbo:Bay` and `dbo:Sea`, with values of 0.55 and 0.51, respectively. Both `gpt-4-1106-preview` and `gpt-4-vision-preview` performed lowest on `dbo:Valley` and highest on `dbo:Bay`. Except for `dbo:Valley` (0.2 versus 0.27), surprisingly, `gpt-4-1106-preview` outperformed `gpt-4-vision-preview` on three other feature types. This may indicate that a `gpt-4-vision-preview` trained on additional image data (e.g., image–text pairs) does not necessarily encode more geographic knowledge than the pure language model `gpt-4-1106-preview`.

5.2 Local-Analysis Results about UNESCO World Heritage Sites

From the four selected feature types, we focus on `dbo:WorldHeritageSite` features next. According

Figure 1. The retrieval process of a `dbo:Sea` feature `dbr:Mediterranean_Sea` and its abstract from DBpedia
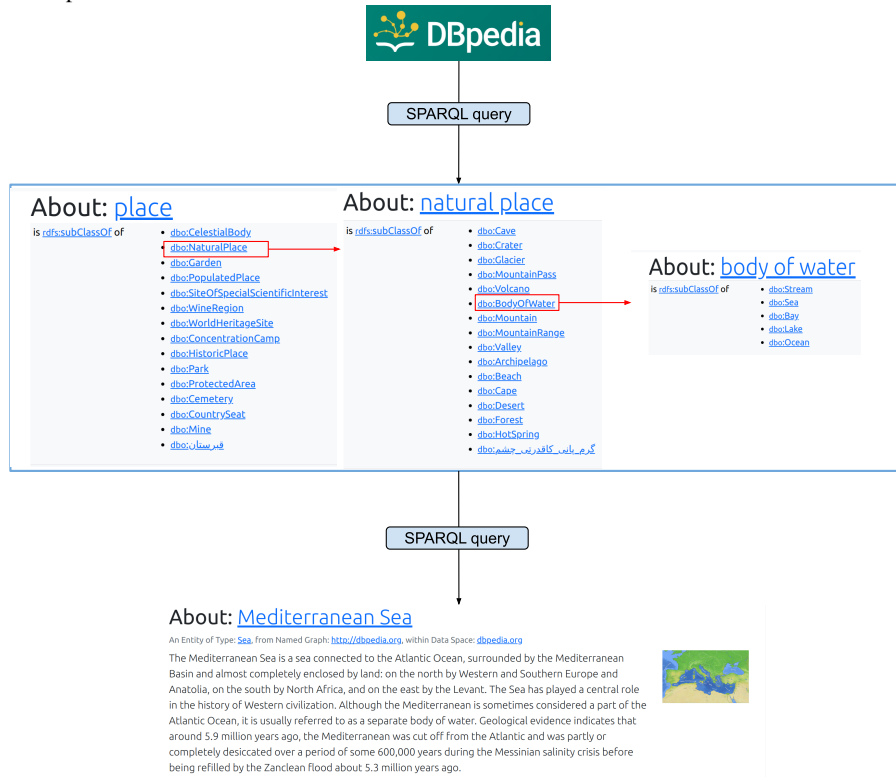

Figure 2. An example geo-guessing experiment about a `dbo:Bay` feature `dbr:Gulf_of_Thailand`, implemented with the `Chat` mode in OpenAI Playground

Figure 3. The hierarchy of DBpedia's `dbo:Place` subclasses used in our work-in-progress
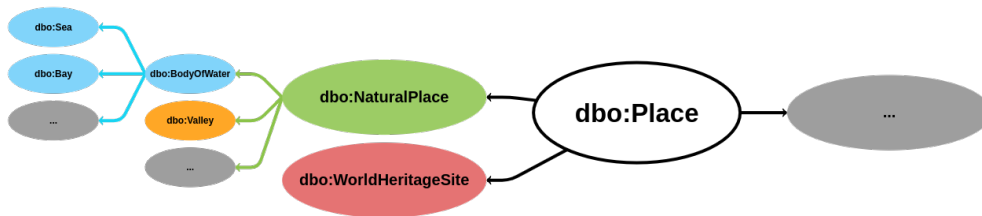
to the United Nations Educational, Scientific and Cultural Organization (UNESCO), "World Heritage sites belong to all the peoples of the world, irrespective of the territory on which they are located"[11]. Therefore, these sites are geographic features that carry both interpretations by local populations and universal values for

---

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

Table 1. The percentage of correct predictions by `gpt-4-1106-preview` and `gpt-4-vision-preview` on features belonging to `dbo:WorldHeritageSite`, `dbo:Valley`, `dbo:Bay`, and `dbo:Sea`

Feature Types	gpt-4-1106-preview	gpt-4-vision-preview
dbo:WorldHeritageSite	0.38	0.31
dbo:Valley	0.2	0.27
dbo:Bay	0.55	0.475
dbo:Sea	0.51	0.46

all of humanity. Compared with the previous analysis on `dbo:WorldHeritageSite` features from a *global* perspective, here we define localness with two kinds of geographical units to examine GPT-4's performance regarding this unique feature type. One kind of unit is countries, and the other one is regions defined by UNESCO for its activities[12]. We then measured GPT-4's performance as the percentage of correct predictions aggregated by these two units. When assessing by countries, we only include countries with more than ten sites in our ground-truth corpus.

Table 2 shows the top ten countries ordered by the percentage of correct predictions by `gpt-4-1106-preview` and `gpt-4-vision-preview`, respectively, on `dbo:WorldHeritageSite` features. For both models, there were inter-country disparities in their geo-guessing performance. In addition, their performance was less than or equal to 0.5, indicating a severe lack of encoded knowledge in GPT-4 about `dbo:WorldHeritageSite` on a country level. Both models had the same accuracy for four countries, including Spain (0.31), Germany (0.23), Switzerland (0.17), and Chile (0.07). However, `gpt-4-1106-preview` performed better in all the rest six countries, including France (0.5 versus 0.2), India (0.47 versus 0.41), China (0.39 versus 0.33), Italy (0.38 versus 0.29), Belgium (0.33 versus 0.25), and Japan (0.19 versus 0.13). The inter-model disparities indicate that `gpt-4-1106-preview` generally had a better country-level performance than `gpt-4-vision-preview` when geo-guessing `dbo:WorldHeritageSite` features.

Table 3 shows the UNESCO-regions ordered by the percentage of correct predictions by `gpt-4-1106-preview` and `gpt-4-vision-preview`, respectively, on `dbo:WorldHeritageSite` features. In addition to inter-UNESCO-regional disparities in the performance of both models, we again observe that their performance was less than 0.5, which indicates a similar lack of UNESCO-region-level knowledge about `dbo:WorldHeritageSite` encoded in GPT-4. Except for Arab States (0.28), `gpt-4-1106-preview` had a better performance than `gpt-4-vision-preview` in all the rest four UNESCO regions, including Latin America and the Caribbean (0.413 versus 0.26), Asia and the Pacific (0.407 versus 0.36), Africa (0.4 versus 0.37), and Europe and North America (0.36

versus 0.27). Again, this reveals inter-model disparities in GPT-4's geo-guessing performance on `dbo:WorldHeritageSite` features on a UNESCO-region level, and `gpt-4-1106-preview` generally performed better on this level as well.

When comparing Table 2 and Table 3, we notice greater disparities in the country-level performance than in the UNESCO-region-level performance. The `gpt-4-1106-preview` model had an accuracy with a range of 0.43 on a country level, compared with a range of 0.133 on a UNESCO-region level. Same for `gpt-4-vision-preview`, the accuracy had a range of 0.34 on a country level, which was larger than a range of 0.11 on a UNESCO-region level. This means that as the geographic scale increased from countries to UNESCO regions, inter-region disparities in the geo-guessing performance of both models on `dbo:WorldHeritageSite` features might become smaller.

6 Conclusions and Future Work

In this initial work, we explore the notion of geographic diversity through the lens of LLMs, aiming to better understand how well geographic features are represented. In contrast to the common perspective of seeing GPT-4 as a data generator, we also consider it a geographic knowledge base in its own right. We study geographic diversity with a geo-guessing experiment as an open-ended question-answering test, where GPT-4 is utilized to predict a geographic feature masked in a piece of text. Using English-language DBpedia abstracts, we find that GPT-4 may encode insufficient geographic knowledge about several feature types, including `dbo:WorldHeritageSite`, `dbo:Valley`, `dbo:Bay`, and `dbo:Sea`, on a global level. On a local level, we observe not only this insufficiency but also inter-regional disparities in GPT-4's geo-guessing performance for `dbo:WorldHeritageSite` features that carry both local and global significance. Interestingly, when assessing on a larger geographic scale, inter-regional disparities may become smaller. Moreover, the multimodal variant of GPT-4 may encode even less geographic knowledge than the unimodal version, whether on a global level for all selected feature types or on a local level for `dbo:WorldHeritageSite` alone. We speculate that GPT-4 does not perform well in our experiment due to reasons such as the loss in training data compression, the vulnerability to factual contradictions appear-

---

[12] https://whc.unesco.org/en/activities

Table 2. The top ten countries (with more than ten sites) ordered by the percentage of correct predictions by `gpt-4-1106-preview` and `gpt-4-vision-preview`, respectively, on `dbo:WorldHeritageSite` features

gpt-4-1106-preview	gpt-4-vision-preview
France (0.5)	India (0.41)
India (0.47)	China (0.33)
China (0.39)	Spain (0.31)
Italy (0.38)	Italy (0.29)
Belgium (0.33)	Belgium (0.25)
Spain (0.31)	Germany (0.23)
Germany (0.23)	France (0.2)
Japan (0.19)	Switzerland (0.17)
Switzerland (0.17)	Japan (0.13)
Chile (0.07)	Chile (0.07)

Table 3. The regions (defined by UNESCO for its activities) ordered by the percentage of correct predictions by `gpt-4-1106-preview` and `gpt-4-vision-preview`, respectively, on `dbo:WorldHeritageSite` features

gpt-4-1106-preview	gpt-4-vision-preview
Latin America and the Caribbean (0.413)	Africa (0.37)
Asia and the Pacific (0.407)	Asia and the Pacific (0.36)
Africa (0.4)	Arab States (0.28)
Europe and North America (0.36)	Europe and North America (0.27)
Arab States (0.28)	Latin America and the Caribbean (0.26)

ing in data conflation, the tendency for LLMs to repeat other named entities (in the prompt) as the correct answer, and so forth. Considering that the training data of GPT-4 is likely to have already included DBpedia, one promising way of enhancing its performance is to implement retrieval-augmented generation (Lewis et al., 2020), a general-purpose fine-tuning approach that could use DBpedia again as an external knowledge base.

Future work will require a larger-scale but granular analysis of geographic features, supported by various ground-truth knowledge corpora and comprehensive probing techniques. While our experiment provides linguistically and geographically contextual (unstructured) data about a target feature, it is neither a geoparsing task where the feature is unmasked nor a visual GeoGuessr[13] game where a player is asked to locate where a photo was taken. However, these two tasks could give us the inspiration to develop better probing techniques for geographic knowledge. For instance, one could ask LLMs to output a feature name along with geospatial information if representing it with different geometric primitives (e.g., points, lines, polygons), or to list features that are topologically connected if spatial predicates are given. Also, one could replace a masked abstract with their own dataset consisting of multi-perspective descriptions about a geographic feature. In fact, knowledge graphs, such as DBpedia, provide a rich body of structured knowledge, which could help achieve both mainstream probing and conduct our proposed geo-guessing experiment. As knowledge graphs also provide information ontologies, we could study both the *intension* (i.e., the properties of a category) and the

extension of a geographic feature type and their roles in foundation models.

7 Acknowledgement

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report, arXiv preprint arXiv:2303.08774, 2023.

Bhandari, P., Anastasopoulos, A., and Pfoser, D.: Are Large Language Models Geospatially Knowledgeable?, in: Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, pp. 1–4, 2023.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258, 2021.

Jackson, P.: Thinking geographically, Geography, 91, 199–204, 2006.

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

---

[13]https://www.geoguessr.com

Janowicz, K.: Philosophical foundations of geoai: Exploring sustainability, diversity, and bias in geoai and spatial data science, in: Handbook of Geospatial Artificial Intelligence, pp. 26–42, CRC Press, 2023.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., et al.: Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia, Semantic web, 6, 167–195, 2015.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems, 33, 9459–9474, 2020.

Liétard, B., Abdou, M., and Søgaard, A.: Do Language Models Know the Way to Rome?, in: Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pp. 510–517, 2021.

Liu, Z., Janowicz, K., Cai, L., Zhu, R., Mai, G., and Shi, M.: Geoparsing: Solved or biased? an evaluation of geographic biases in geoparsing, AGILE: GIScience Series, 3, 9, 2022.

Mai, G., Cundy, C., Choi, K., Hu, Y., Lao, N., and Ermon, S.: 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.

Openshaw, S.: The modifiable areal unit problem, Concepts and techniques in modern geography, 1984.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., and Riedel, S.: Language models as knowledge bases?, arXiv preprint arXiv:1909.01066, 2019.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., and Sculley, D.: No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World, in: NIPS 2017 workshop: Machine Learning for the Developing World, 2017.

Zhao, B., Zhang, S., Xu, C., Sun, Y., and Deng, C.: Deep fake geography? When geospatial data encounter Artificial Intelligence, Cartography and Geographic Information Science, 48, 338–352, 2021.