Does Cognitive Plausibility Matter for Agentic GeoAI Persona? Testing Source–Target Asymmetry in Distance Estimation

Krzysztof Janowicz[1], Alexandra Fortacz-Lazan[1], Seyedeh Negin Kosari[1], Linus Lulkiewicz[1], Carl Feldmann[1], Sebastian-Paul Lambrecht[1], Bastian Maximilian Arnold[1], Fabio Spitzenstätter[1], Alexander Filipsky[1], Brenda Tarn[1], Nelly Dorrmann[1], Martin Jakob Winder[1], Paul Vitovec[1] and Sebastian Kerstein[1]

[1]*Department of Geography and Regional Research, University of Vienna, Austria*

Abstract

AI agents are increasingly acting on our behalf, either directly as agentic systems or indirectly as synthetic participants or personas used to model and stand in for human behavior. For instance, recent work even explores whether such AI agents can be used for public participatory planning. Such work relies on the often implicit assumption that these agents can meaningfully mimic (groups of) humans. This, in part, explains the rapid rise in work that studies how generative AI performs tasks such as topological reasoning or even map generation. While the results are mixed, frontier reasoning models can now indeed perform many human-level tasks. *But what if we are asking the wrong questions?* If AI agents are supposed to meaningfully mimic human behavior, e.g., for public participatory planning, would they not also have to mimic human heuristics and cognitive patterns? Put differently, most current work focuses on what agents *can do*, not whether they do it *similarly to humans*. While this may not matter for arithmetic or note-taking, it certainly matters for evacuation simulations, mobility, privacy, behavioral geography, and spatial decision-making. Hence, in this paper, we are not going to focus on the task alone, here distance estimation, but also on whether we can observe *asymmetry* in these estimations under source→target reversal, as reported by classic geographic analysis literature from the past. Our results indicate that distance estimation is indeed subject to direction, but that this observation is more likely related to prompt sensitivity (prompt direction) than to human-like cognitive bias.

Keywords

AI persona, agentic AI, cognitive plausibility, behavioral geography, large language models

1. Introduction

AI agents are increasingly entering fields previously thought to require human-level intelligence. They do so either directly, for instance, as autonomous or semi-autonomous agentic systems, or more indirectly as synthetic participants, personas, or even as a simulated public used to approximate human behavior, preferences, or even decisions [1]. Assessing whether these agents can truly carry out such work is difficult. Nonetheless, for tasks such as summarization, calendar and task management, digital asset culling, and even programming (vibe coding), we can establish benchmarks to study the (increasing) accuracy of these systems over model generations. This also explains the rapid increase in research literature documenting the ability [2] of AI systems and their underlying foundation models, be they large language models (LLMs), text-to-speech (TTS), or text-to-image (T2I) models.

Unfortunately, however, for other tasks, focusing on accuracy alone may fall short. For example, public participatory planning [1], policy making, human mobility, tourism recommendations, or disaster mitigation [3] all involve perception, individual preferences, local knowledge, social context, and also understanding cognitive heuristics. This also applies to economics, where Tversky and Kahneman [4] famously argued that humans are not perfectly rational agents and that accounting for their heuristics and biases is key to understanding decision-making, e.g., under risk. One such bias documented in behavioral geography is an *asymmetry in distance estimations* of landmarks [5] that

First International Workshop on Spatial Intelligence and Reasoning Enabled by Knowledge Graphs and Foundation Models (SPARK 2026), in ISWC 2026 Workshops Joint Proceedings, October 25–26, 2026, Bari, Italy

krzysztof.janowicz@univie.ac.at (K. Janowicz); alexandra.fortacz-lazan@univie.ac.at (A. Fortacz-Lazan)

© 2026 Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-05

CEUR-WS.org/Vol-4313/SPARK-paper6.pdf

arises from the observation "that the cognitive distance between reference points and nonreference points is asymmetrical; nonreference points are judged nearer to reference points than are reference points to nonreference points." [6, p. 420]. In the 1970s, these arguments were used to question an often implicit assumption of research on cognitive mapping, namely that a *cartographic representation of such cognitive maps is meaningful* in the first place.

Reflecting on this research from today's perspective, we may ask a similar question. When using agentic AI systems as stand-ins for (groups of) humans: **is it sufficient to study only the ability of these systems and not also whether their behavior is cognitively plausible?** Concerns about the cognitive plausibility of computational approaches are not new to our domain. For example, in studying geographic feature types, researchers have examined the plausibility of semantic similarity measures using classical feature-based approaches and those based on conceptual spaces [7, 8]. Interestingly, this is not yet the case for studying the spatial abilities of generative AI where current research is often focused on accuracy alone. Nonetheless, while the results are still mixed, researchers have documented progress in spatial, topological and directional reasoning and planning abilities of (multi-modal) foundation models [2, 9, 10, 11]. A recent study by Fulman et al. [12], however, did consider the *bounded rationality* of AI agents in simulated parking search.[1]

Hence, in this short paper, we would like to address this *research gap* and perform a distance estimation experiment not merely to measure accuracy but also to study whether asymmetry will arise as well. **Consequently, our work will address the following questions.**

1. Can large language models (LLMs) accurately estimate the geographic distance between source→target pairs latently, i.e., without lookup or using tools (e.g., reasoners)?
2. Do ordered source→target pairs remain invariant under reversal, i.e., are distance judgments symmetric?
3. Is the lack of symmetry caused by prompt sensitivity and the probabilistic nature of generative AI, or are outputs systematically biased in one direction?
4. Finally, we are also interested in aspects of sustainability: can smaller, and, hence, cheaper and potentially less resource-intensive models compete with full SotA frontier models?

Note that these questions also hint at the relation between using (or combining) symbolic and neural systems. For a *metric* notion of distance, for instance, the question of whether symbolic systems would return the same results under reversal does not even arise by definition.

2. Experimental Setup

Before explaining our experimental setup, please note that we are not advocating that agentic AI systems should fall victim to the same cognitive heuristics and biases that humans do, but that tasks where agents are supposed to mimic human behaviors will also have to be judged by how cognitively plausible their behavior is. In many cases, for instance, it may be more important to understand *how* humans evacuate than *whether* they are taking the best evacuation routes.

As this work is a joint classroom experiment, we selected several models from the ChatGPT family as well as one open-weights model, namely the new Gemma 4 31B-it.[2] As these models have a substantial environmental footprint and potentially high costs per token, we also wanted to test substantially smaller models to understand whether they would be able to perform this (relatively simple) task of distance estimation. Hence, we selected the following models: gpt-5.4-2026-03-05, gpt-4.1-2025-04-14, gpt-4.1-**mini**-2025-04-14, gpt-4.1-**nano**-2025-04-14, gpt-5.4-**nano**-2026-03-17, and a locally hosted gemma-4-31B-it developed by Google.

For each of these models, we ran at least one full test consisting of 2x 15,000 source and target city pairs with the following user prompt and system instructions. Temperature for all models

[1] More generally, the large-scale study by Coda-Forno et al. [13] provides an early example of applying a cognitive evaluation approach to LLMs to study *how* a certain task is solved.

[2] In the future, we may also test further models. Note, however, that prior work has shown that the GPT-family substantially outperforms models such as DeepSeek, Llama, etc., on geographic and topological tasks.

was set to *0.3*, and reasoning effort was set to *none* for the GPT-5 family (a prerequisite to setting model temperature).[3] The local Gemma 4 installation was also queried using the same OpenAI API style.

User prompt: *What is the distance between {source city} and {target city}?*

System prompt: *Use your latent estimate. No lookup of the distances. Return only the distance in kilometers without units. Just return the plain number, e.g., 654.7, no text before or after.*

We did not explicitly specify using great-circle distance, nor provide further hints that may *indirectly imply a desire for metric properties.*[4] As earlier runs demonstrated that the smaller models would not always follow instructions, we limited token size to 20, thereby leaving enough room for brief error tracking, e.g., when the model answered "Approximately 850 km, instead of 850". Overall, this was not a major issue with even the smallest models producing fewer than 1% numerically invalid responses.

As we needed the actual ground truth distance and to make sure that the large language model did not instead pick any of the many possible places called, say, London or York, we used qualified place names, e.g., Pokrzywnica, Poland or Shanghai, China. While this does not exclude places such as Springfield appearing multiple times in a country (here, USA), it strikes a balance between not providing too many details to the model to make distance estimation trivial and limiting ambiguity. Put differently, even though some of these place names may appear multiple times within the same country, this is unlikely to have a noticeable statistical effect over the 15,000 records (per run) overall.

To establish source→target city pairs with a clear major city and a minor city, thereby taking inspiration from the distance between reference points and non-reference points in prior work [5], we first selected the top 500 cities from GeoNames by population. This resulted in a list whose smallest cities still had roughly 1 million inhabitants. For each of those top 500 cities, we picked 30 other cities (with a population over a thousand) at random. Finally, we reversed each pair so each data frame contains 2x 15,000 pairs. Consequently, each city appeared in both source and target slots. For internal use, we also stored the geographic coordinates, population, actual great-circle distance, and continents (for source and target).

3. Results and Discussion

Here we report the results of our experiment. First, we are interested in the magnitude and the direction of the percentage error, and less in the raw error. Simply put, we want to account for the fact that if two cities are nearby but the distance has been misestimated by 100 km, this is substantially more significant than if this error occurs for two cities thousands of km apart. Hence, we define the two errors simply as follows:

$$\text{Signed percentage error} = 100 \cdot \frac{\hat{d}_i - d_i}{d_i} \tag{1}$$

$$\text{Absolute percentage error} = 100 \cdot \frac{\left|\hat{d}_i - d_i\right|}{d_i} \tag{2}$$

As these percentage-based measures do not provide an intuition for the scale of errors, we note that the best-performing model (gpt-4.1) has a median/mean absolute error of **244/391 km**, while the weakest model (gpt-4.1-nano) has a median/mean absolute error of **1598/2428 km**. The newest model (gpt-5.4) scores second best at 256/401 km. These differences are very small; hence, we assume both models (4.1 and 5.4) to perform on par.

[3] We selected a temperature of 0.3 as (in real deployment cases) agents and generative AI more generally rarely run at a temperature of zero. For later comparison, we ran models such as gpt-5.4 multiple times. We will return to both issues later.

[4] Asking for a *latent* estimate does not guarantee by itself that a model will not use RAG or other tooling. In our case, we can confirm that the local Gemma model had no access to RAG or any tooling. For OpenAI's models, it is our understanding that

Table 1
Estimation error per model. Signed percentage error indicates systematic over/under-estimation, while absolute percentage error indicates the magnitude irrespective of direction.

Model	Median/mean signed error (%)	Median/mean absolute error (%)
gpt-4.1	-1.45 / -2.73	3.38 / 6.83
gpt-5.4	-1.06 / -2.08	3.44 / 7.59
gemma-4-31B-it	1.29 / 2.97	6.16 / 11.69
gpt-4.1-mini	-2.23 / -1.81	7.04 / 11.94
gpt-5.4-nano	-4.14 / -1.31	12.12 / 20.19
gpt-4.1-nano	**3.37** / **27.00**	**21.02** / **46.89**

As summarized in Table 1, we can make several observations about the percentage errors: **(1)** four of the six models tend to underestimate distances; **(2)** the smaller models perform substantially worse. While gpt-4.1-mini could still be considered usable, both *nano* models fall behind; **(3)** the open-weights model Gemma 4 performs (perhaps surprisingly) well given that it can run on a single GPU workstation; **(4)** the noticeable gaps between median and mean show that all models returned some large-error outliers; and, finally, **(5)** on a minor note, in contrast to all other models, Gemma 4 did not use decimal places and only returned multiples of 5, i.e., 115, 850, 2480, but never 2486.7, over 14,999 pairs.

Next, we have to look at the asymmetry, i.e., whether there is a difference in estimation from major-to-minor versus minor-to-major city pair direction. Here Mm is the **M**ajor-**m**inor direction, while mM will signify the reversal.

$$\text{Signed percentage asymmetry} = 100 \cdot \frac{\hat{d}_i^{Mm} - \hat{d}_i^{mM}}{d_i} \tag{3}$$

$$\text{Absolute percentage asymmetry} = 100 \cdot \frac{\left|\hat{d}_i^{Mm} - \hat{d}_i^{mM}\right|}{d_i} \tag{4}$$

Table 2
Relative directional asymmetry per model for source→target reversals.

Model	Med./mean signed asymmetry (%)	Med./mean absolute asymmetry (%)
gpt-4.1	0.00 / -0.10	1.81 / 3.93
gpt-5.4	0.12 / -0.31	2.06 / 4.84
gemma-4-31B-it	**0.80** / **1.79**	2.82 / 6.84
gpt-4.1-mini	0.20 / -0.48	4.69 / 9.02
gpt-5.4-nano	0.42 / 0.74	11.06 / 21.10
gpt-4.1-nano	0.16 / -0.87	**12.13** / **42.68**

As Table 2 shows, distance estimations are indeed asymmetric. However, the signed results show that this effect is unlikely to reflect a strong systematic directional bias. For the *mini* and *nano* models, the (absolute) asymmetry is substantially larger, while remaining very low for the full models. This could imply that the results are driven by **prompt sensitivity**, i.e., differences arising purely because of the syntactic prompt order, and/or because of outliers. To give a concrete example, for gpt-5.4, the typical pair has a small tendency for the major-to-minor direction to be larger, but across all pairs, the effect is slightly negative. This suggests that some of the larger cases fall along the minor-to-major order.

We also checked the fraction of entirely symmetric pairs; here, gpt-4.1 returns 17.5% of such pairs (partially also explaining the results in Table 2), while, on the other end, gpt-5.4-nano contains fewer than 1% symmetric pairs. An additional full run of gpt-5.4 at a **temperature of zero** shows 13% symmetric pairs. As we ran models such as gpt-5.4 multiple times, it is worth comparing across runs as

tool_choice defaults to 'none'; see https://developers.openai.com/api/reference/cli/resources/chat/subresources/completions.

well. While the median of the absolute reversal asymmetry (for a full run of 15k pairs) comes out at about 2.06% at a temperature of 0.3, it drops to 1.62% at temperature zero. *Across* our three independent gpt-5.4, temp=0.3 runs, the median variation between the pairwise comparisons for the same source→target pair, e.g., the 3 individual estimates for Shanghai→Breitenburg, is only 1.18%. Hence, this indicates that the results shown above are unlikely just a sampling artifact of the model temperature.

Among other issues, this raises the question of whether estimates become more difficult as city distance increases. **Interestingly, Figure 1 shows the opposite in relative terms.** In fact, as ground truth distance grows, the absolute percentage error shrinks (while the absolute error in km only increases modestly). Across all models, Spearman rank correlations ρ between ground truth distance and absolute percentage error ranged from **-0.28 to -0.42**, indicating that with larger distances we see lower absolute percentage error. Given the sample size, all results are statistically significant ($p < 0.001$).

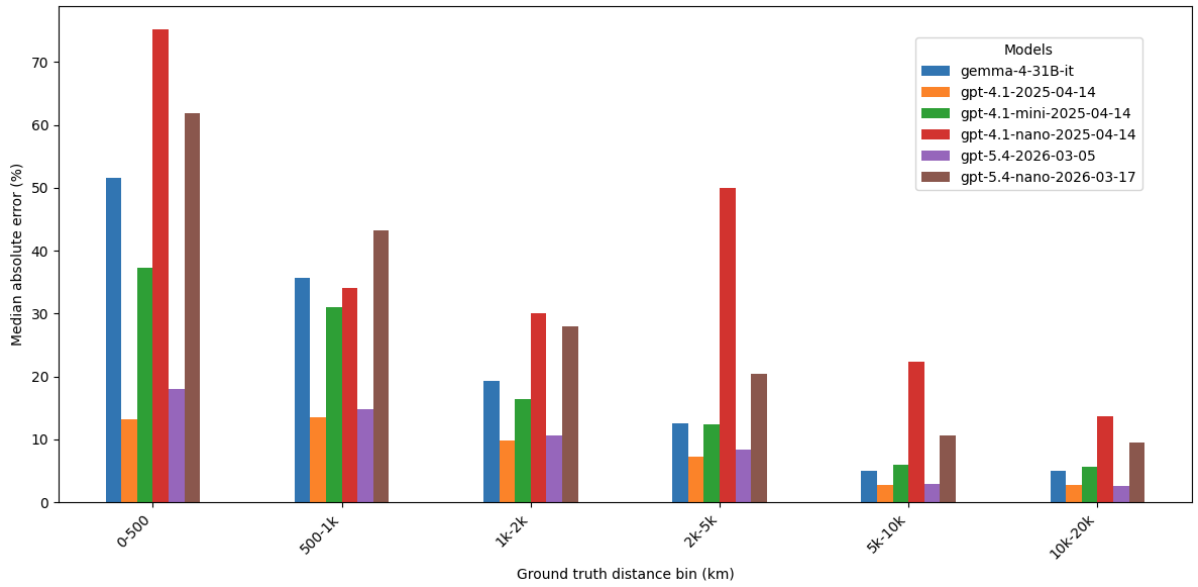


Figure 1: Distance estimation error by binned ground truth distance per model.

There are several potential explanations for this observation. Addressing each of them in a controlled experiment would fall outside the scope of this short paper and is left for future work. Hence, we focus on one potential explanation here: it seems likely that the models have no detailed latent representation of most of the (very) small minor cities. Yet, the training data contain orders of magnitude more information about countries and their relations and locations. Consequently, the models may well represent the approximate coordinates of, say, Shanghai, but for its paired minor city of Breitenburg, Germany, **models may attend more strongly to Germany than to Breitenburg**. Such an approach would still yield good accuracy for smaller nations common in Europe and elsewhere. Conversely, this may yield high percentage errors when comparing cities within Germany or across (small) nations such as Switzerland and Austria. Note that for very short city pairs, the fact that the ground truth dataset is based on point locations may also introduce larger errors at small distances. However, only a few such cases (54 city pairs less than 200 km apart) were present in our data. Hence, we do not consider this a likely explanation.

Finally, we wanted to compare the AI results to human-level performance. For a small set of 2x9 random pairs and n=11 participants, the median and mean absolute percentage errors were **13.7 and 19.8**, respectively. Without reading too much into this small sample, these results are roughly comparable to gpt-5.4-nano but fall substantially behind those of all larger models.

4. Conclusion

In this short paper, we studied the ability of large language models to estimate geographic distances of ordered source→target city pairs. Inspired by prior results from behavioral geography, our interest, however, was not so much in overall accuracy, but in whether we would observe asymmetric effects under source→target reversal. Our main motivation for doing so arose from the observation that current literature on the spatial reasoning abilities of generative AI often focuses on improvements in accuracy alone. This leaves the question of whether these systems produce outputs that are **cognitively plausible** largely unanswered [14]. While this may not matter for arithmetic or note-taking, it may matter greatly for applications such as public participatory planning or when using LLM agent-based personas for evacuation simulation, where human cognitive biases, heuristics, cultural sensitivities, and so on shape the output of such processes.

With regard to our initial research questions, **(RQ1)** we find that LLMs can provide overall accurate geographic distance estimates. However, the differences between smaller and larger models are very noticeable. From the viewpoint of API costs as well as sustainability [15], this, unfortunately, also implies that the nano models are not (yet) suitable for reliable estimation **(RQ4)**. Interestingly, however, the locally hosted Gemma 4 performed remarkably well. Further, our experiment also confirms that results differ under source→target pair reversal **(RQ2)**. However, this lack of symmetry is not systematic in the sense of producing distinct major-to-minor versus minor-to-major distance patterns **(RQ3)**. While we cannot entirely exclude that such effects do exist in LLMs, the results could also be explained by prompt sensitivity to the city-pair order and other aspects. **Consequently, we cannot claim that the distance estimations of generative AI in our experiment are cognitively plausible.** *Whether and when this is a problem is up for debate.* Similarly, there may be other cases where AI agents do behave in line with human heuristics [12]. All this also points to an interesting role for symbolic systems, e.g., based on (logical) deduction and knowledge graphs, which (in *theory*) should not show such syntactic sensitivity. Unfortunately, however, from an actual system *deployment* perspective, retrieval-augmented generation (RAG) and tooling may reintroduce sensitivity to order back into the AI agent's intermediate reasoning and thus its final results (even if the model's temperature would be set to zero). We believe that these issues require further attention.

Finally, it is interesting to note that when comparing the errors across **continent pairs**, e.g., European cities to North American cities versus Asian cities to African cities, we found substantial differences in estimation quality. These results could point to a potential representational bias [16] worth investigating in future work. We do not include this analysis here because the results also raise several methodological questions. For instance, the distribution of the top 500 cities is geographically uneven, as is the presence of small communities in Oceania and Africa in our GeoNames dataset. Similarly, despite the large overall sample, our dataset becomes sparse for pairs such as Oceania–Oceania. Along the same lines, our (largely) aggregated analysis could be complemented with further local investigation out of scope for this short paper.

Declaration on Generative AI

Following the CEUR GenAI Usage Taxonomy, *Grammarly* (in non-generative mode) was used for "grammar and spelling check". *ChatGPT* was used for "peer review simulation" and coding support. The authors reviewed AI outputs and take full responsibility for the publication.

References

[1] J. Du, X. Ye, P. Jankowski, T. W. Sanchez, G. Mai, Artificial intelligence enabled participatory planning: a review, International Journal of Urban Sciences 28 (2024) 183–210.

[2] A. G. Cohn, R. E. Blackwell, Evaluating the ability of large language models to reason about cardinal directions, arXiv preprint arXiv:2406.16528 (2024).

[3] Y. Hu, G. Mai, C. Cundy, K. Choi, N. Lao, W. Liu, G. Lakhanpal, R. Z. Zhou, K. Joseph, Geo-knowledge-guided gpt models improve the extraction of location descriptions from disaster-related social media messages, International Journal of Geographical Information Science 37 (2023) 2289–2318.

[4] A. Tversky, D. Kahneman, Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty., science 185 (1974) 1124–1131.

[5] E. K. Sadalla, W. J. Burroughs, L. J. Staplin, Reference points in spatial cognition., Journal of experimental psychology: human learning and memory 6 (1980) 516.

[6] W. J. Burroughs, E. K. Sadalla, Asymmetries in distance cognition, Geographical Analysis 11 (1979) 414–421.

[7] M. Raubal, B. Adams, The semantic web needs more cognition, Semantic Web 1 (2010) 69–74.

[8] C. Keßler, What is the difference? a cognitive dissimilarity measure for information retrieval result sets, Knowledge and Information Systems 30 (2012) 319–340.

[9] Y. Ji, S. Gao, Y. Nie, I. Majić, K. Janowicz, Foundation models for geospatial reasoning: assessing the capabilities of large language models in understanding geometries and topological spatial relations, International Journal of Geographical Information Science 39 (2025) 1866–1903.

[10] I. Majic, Z. Wang, K. Janowicz, M. Karimi, Spatial task-explicity matters in prompting large multimodal models for spatial planning, in: Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, 2024, pp. 99–105.

[11] M. Aghzal, E. Plaku, Z. Yao, Can large language models be good path planners? a benchmark and investigation on spatial-temporal reasoning, arXiv preprint arXiv:2310.03249 (2023).

[12] N. Fulman, A. Memduhoğlu, A. Zipf, Utilizing large language models to simulate parking search, Transportation Research Part A: Policy and Practice 199 (2025) 104542.

[13] J. Coda-Forno, M. Binz, J. X. Wang, E. Schulz, Cogbench: a large language model walks into a psychology lab, in: R. Salakhutdinov, Z. Kolter, K. A. Heller, A. Weller, N. Oliver, J. Scarlett, F. Berkenkamp (Eds.), Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024, volume 235 of *Proceedings of Machine Learning Research*, PMLR / OpenReview.net, 2024, pp. 9076–9108. URL: https://proceedings.mlr.press/v235/coda-forno24a.html.

[14] K. Janowicz, G. Mai, R. Zhu, S. Gao, Z. Wang, Y. Hu, L. Bennett, Geography according to chatgpt-how generative ai represents and reasons about geography, in: Geography According to Foundation Models, SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2026, pp. 2–11.

[15] M. Shi, K. Janowicz, J. Verstegen, K. Currier, N. Wiedemann, G. Mai, I. Majic, Z. Liu, R. Zhu, Geography for ai sustainability and sustainability for geoai, Cartography and Geographic Information Science 52 (2025) 331–349.

[16] R. Manvi, S. Khanna, M. Burke, D. Lobell, S. Ermon, Large language models are geographically biased, arXiv preprint arXiv:2402.02680 (2024).