LARGE LANGUAGE MODELS FOR UNDERSTANDING SPATIAL EXPERIENCE? BIAS, VARIATION, AND PERSONA EFFECTS

Yue Lin[a]*

[a] Department of Geography and Geographic Information Science, University of Illinois Urbana-Champaign, Urbana, USA
* linyue@illinois.edu

Keywords: artificial intelligence, cognitive mapping, GeoAI, vague cognitive regions

Introduction

Studies of spatial cognition and experience, which examine how humans perceive, interpret, and interact with places and space, have traditionally relied on methodologies such as observational studies, surveys, and mixed-method analyses. Many of these approaches rely on direct human participation, often through instruments such as questionnaires. However, data collection from human participants has become increasingly difficult in recent years due to rising costs and declining response rates (Williams & Brick, 2018). This has prompted growing interest in looking for cost-efficient alternatives.

Advances in artificial intelligence (AI), particularly large language models (LLMs), are starting to reshape how data collection for studying spatial cognition and experience can be approached (Grossmann et al., 2023). These transformer-based models, pretrained on vast amounts of text data, are increasingly capable of simulating human-like responses, which might offer new opportunities to test theories and hypotheses about human behavior at scale and speed. Notably, LLMs can also be assigned "personas" through system instructions, allowing them to role-play particular identities and potentially simulate the behavior of specific populations. In fact, social and behavioral scientists have already begun using LLMs as cost-effective substitutes for human participants, for example, to generate responses to polls (Bisbee et al., 2024), make moral judgements (Dillion et al., 2023), and support social simulations (Qu & Wang, 2024). Buts while LLMs present promising opportunities as low-cost alternatives to humans, their value should not be assumed. It remains critical to evaluate how closely their outputs align with human responses and to assess whether they introduce biases that could compromise reliability or cause harm.

This paper describes efforts to examine whether, and to what extent, LLMs produce synthetic responses that approximate human responses in understanding spatial cognition and experience through the lens of *vague cognitive regions*, which are geographic areas that do not have formally defined boundaries in people's mind. The geographic literature has long explored behavioral approaches to studying vague regions, such as downtown Santa Barbara (Montello et al., 2003), Northern and Southern California (Montello et al., 2014), Los Angelas Koreatown (Bae & Montello, 2018), and Central Ohio (Xiao, 2025), where participants with lived experience in these areas were asked questions designed to elicit their beliefs about where a region lies.

Building on existing work, this paper presents two case studies, one completed and one ongoing, comparing LLM responses with those of human participants in identifying and mapping vague cognitive regions: (1) Chicago neighborhoods, which are smaller, subordinate units embedded in the city's grid system, and (2) Central Ohio, which spans a larger area covering multiple counties. By examining these two geographically distinct types of vague regions, the paper seeks to contribute to critical literacy around emerging AI models in spatial research and mapping, with attention to how they might be responsibly leveraged for geographic research and practice.

Method

Human understandings of vague cognitive regions have traditionally been investigated through two main approaches. The first is the boundary-based approach, which assumes these regions have inherent boundaries in people's minds, though the cognitive boundaries can vary among individuals (Lee, 2017). But some scholars argue against this assumption, citing contradictory evidence, and instead advocate for an anchor point approach, which identifies prototypical or typical locations and determines whether they are in or outside the region (Xiao, 2025). In this paper, the two case studies employ these conceptualizations separately: the Chicago neighborhoods study uses the boundary-based approach, and the Central Ohio study uses the anchor point approach.

The study on Chicago neighborhoods began by crowdsourcing neighborhood names and boundary drawings from Chicago residents (Talen et al., 2026). This yielded over 5,500 responses from participants whose demographics (e.g., tenure, gender identity, age, and racial/ethnic identity) roughly reflected Chicago's overall population. After filtering out neighborhoods with low response frequencies, 4,529 responses were retained, representing 127 unique neighborhood names. Next, GPT-4o (a widely used LLM and the latest model available at the time) was applied using structured prompts that incorporated the solicited neighborhood names and requested corner point addresses for each neighborhood's boundaries based on the model's judgment (Lin et al., 2025). These corner points were then geocoded using Google's geocoding service and mapped to visualize the LLM-generated boundaries. Figure 1 shows the drawn Chicago neighborhood boundaries by (a) human participants and (b) GPT-4o.

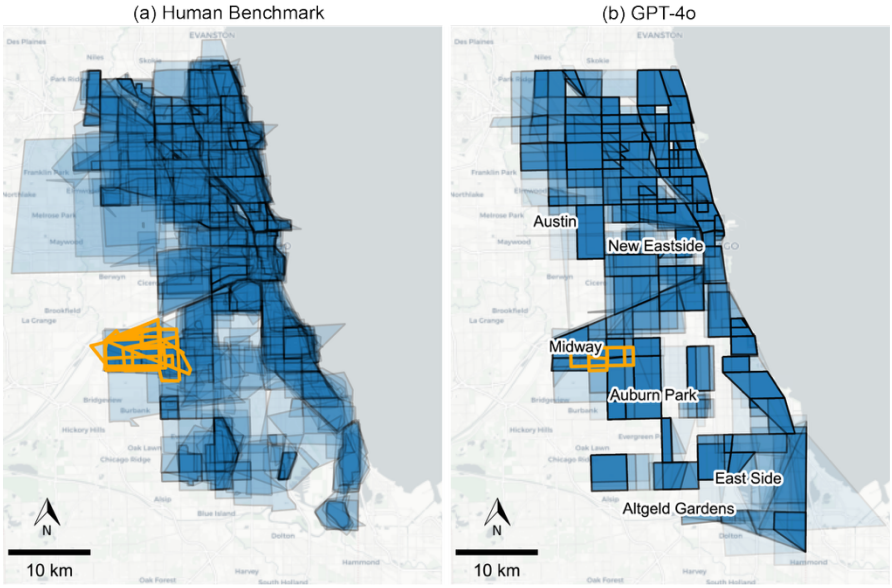

Figure 1. Drawn Chicago neighborhood boundaries by (a) human participants and (b) GPT-4o (Lin et al., 2025).

The Central Ohio study began with a prior survey in which 232 individuals were asked whether each of 67 Ohio cities (selected to offer a representative sample of Ohio, serving as anchor points) are in or not in Central Ohio (Xiao, 2025). The responses were compared with outputs from different LLMs, with each model prompted the same number of times as there were human participants to ensure comparable sample sizes. Specifically, the study tested state-of-the-art versions at the time of writing (GPT-4.1 and Gemini 2.5 Pro) as well as their previous best-performing versions (GPT-3.5 Turbo and Gemini 1.5 Pro). Additionally, the LLMs were prompted with assigned personas (e.g., a newcomer or long-term resident) through system instructions that prompted the models to simulate the experiences, knowledge, and traits associated with a given identity. Since these factors are known to affect human spatial cognition and experience, this approach tested whether personas similarly impact LLM responses. Figure 2 shows aggregated results from multiple responses classifying each Ohio city as in, not in, or indeterminate for Central Ohio according to (a) human participants and (b) GPT-4.1 (no specific persona assigned).

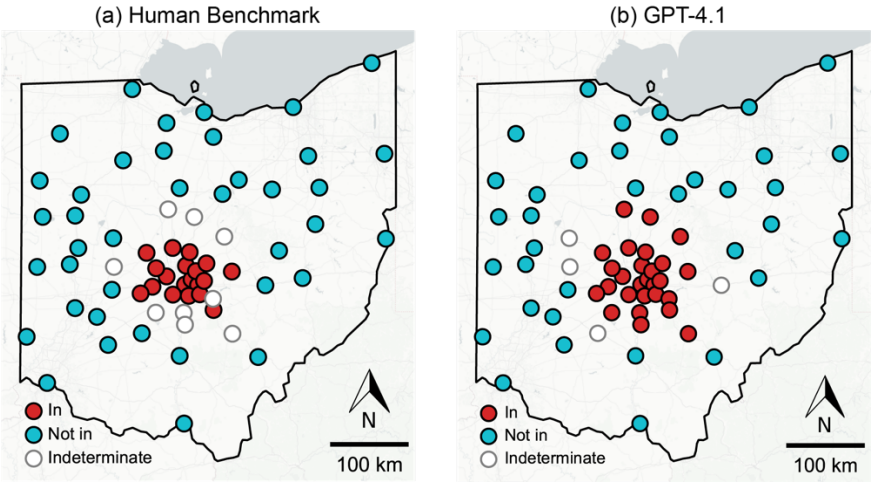

Figure 2. Mapped Ohio cities in terms of whether they are in, not in, or indeterminant for Central Ohio according to (a) human participants and (b) GPT-4.1 (no specific persona assigned).

Results

The results first reveal that LLM-generated responses about vague regions generally align with human responses. For Chicago neighborhoods, the agreement rate between LLM and human-drawn boundaries is mostly above 50%. A comparison of the aggregated classification of Ohio cities (as in, not in, or indeterminate for Central Ohio) between humans and LLMs across models and personas yields precision, recall, and F1 scores above 0.8, suggesting generally well-aligned classifications.

However, the limitations of LLMs for this application are also noticeable and warrant attention. First, geographic bias was identified in the LLM responses. In the Chicago neighborhood boundaries study, LLM-generated boundaries were less likely than

human-drawn boundaries to cover areas with lower population density and higher percentages of non-White and Hispanic populations, reflecting potential biases in the models' training data or geographic knowledge. Second, LLM responses showed a potential loss of variability and diversity commonly observed in human responses. In both studies, LLM responses about either a neighborhood's boundary or whether an Ohio city is in or not in Central Ohio remained highly consistent across multiple repetitions, whereas humans showed much greater variability, even for well-defined neighborhoods or cities that seemingly should be clearly in or out of Central Ohio (e.g., Bexley). Third, while prior research suggests that personas can influence LLM outputs, the results found no noticeable differences between responses generated under different personas, even when strongly framed. This suggests current limitations in how personas can meaningfully affect LLM responses in spatial cognition tasks.

Discussion and Conclusion

As AI increasingly makes its way into everyday use across many aspects of geographic and spatial research, questions arise about whether such technologies can effectively substitute for human participants. This paper attempts to addresses this question by presenting two case studies that compare LLM outputs with human responses in understanding spatial cognition and behavior. The findings reveal promise but also significant limitations of LLMs. While LLMs can approximate human responses in some contexts, they exhibit notable shortcomings including geographic bias, reduced variability compared to human responses, and limited responsiveness to persona-based prompting. These limitations suggest that LLMs cannot yet serve as reliable substitutes for human participants in spatial research. Given the growing interest in leveraging AI for spatial research and the potential implications of these technologies for how we understand human spatial cognition and experience, this paper highlights the critical need for domain-specific validation and careful scrutiny as new technologies are integrated into spatial research methodologies.

References

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. Political Analysis, 32(4), 401–416.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? In Trends in Cognitive Sciences (Vol. 27, Issue 7, pp. 597–600). Elsevier Ltd.

Grossmann, I., Feinberg, M., Parker, D. C., Christakis, N. A., Tetlock, P. E., & Cunningham, W. A. (2023). AI and the transformation of social science research. Science, 380(6650), 1108–1109.

Lee, S. A. (2017). The boundary-based view of spatial cognition: a synthesis. Current Opinion in Behavioral Sciences, 16, 58-65.

Lin, Y., Bae, C. J., & Talen, E. (2025). How would AI define neighborhood boundaries? A comparison with human-crowdsourced data. Environment and Planning B: Urban Analytics and City Science, 23998083251369570.

Qu, Y., & Wang, J. (2024). Performance and biases of Large Language Models in public opinion simulation. Humanities and Social Sciences Communications, 11(1), 1095.

Talen, E., Wileden, L., & Bae, C. (2026). Normative neighborhoods: does theory match perception?. Landscape and Urban Planning, 265, 105516.

Williams, D., & Brick, J. M. (2018). Trends in U.S. Face-To-Face Household Survey Nonresponse and Level of Effort. Journal of Survey Statistics and Methodology, 6(2), 186–211.

Xiao, N. (2025). Where Is Central Ohio? Many-Valued Logic Approaches to Understanding and Measuring Vague Geographic Regions. Annals of the American Association of Geographers, 1-24.