International Cartographic Association
Association Cartographique Internationale

From Maps to Meaning: An Exploratory Study of Generative AI Models' Use of Statistical Maps

Marta Solarz [1]*, Izabela Gołębiowska [1]

Department of Geoinformatics, Cartography and Remote Sensing, Faculty of Geography and Regional Studies, University of Warsaw, Warsaw, Poland – m.solarz2@uw.edu.pl, i.golebiowska@uw.edu.pl

* Corresponding author

Keywords: generative artificial intelligence, thematic maps, multimodal models, map use, map interpretation, geoAI

Abstract:

In recent years, artificial intelligence (AI), particularly its generative forms (GenAI), has rapidly advanced, gaining popularity across fields such as data analysis and visualization (Yilin et al., 2024). Currently, researchers are exploring how AI and Computer Vision can interpret visual data representations and extract information from graphical content (Kommisetty et al., 2024). In the context of cartography, extraction of information for map content has been formulated in three map use types: **map reading**, which involves identifying elements and decoding data (Buckley & Kimerling, 2021); **map analysis**, which is recognizing relationships between map elements and their real-life counterparts (Kimerling et. al., 2016); and **map interpretation**, which is a process requiring advanced reasoning and knowledge of the context (Kang et. al., 2024). In geoscience, AI has been applied to, among others, satellite image analysis, map description generation, and map-type classification (Logar et al., 2020; Robinson & Griffin, 2024; Wen et al., 2024), as well as to autonomous map creation via frameworks like MapGPT (Zhang et al., 2024). Nevertheless, many possibilities of using AI in geoscience remain unexplored, this also applies to statistical maps, which are the main subject of this study.

The study reported here aims to evaluate the ability of selected LLM-based generative AI models to extract information from statistical maps at three levels of a map use complexity: map reading, map analysis, and map interpretation (Fig. 1). The selection of models was guided by key criteria, including multimodality (ability to process both visual and textual input), demonstrated response quality in publicly available benchmarks, accessibility via chat interfaces, and representation of distinct vendors offering current, production-ready versions. For the pilot study, three GenAI models were selected: GPT-4o by OpenAI, Gemini 1.5 Pro by Google, and Claude Sonnet 3.5 v2 by Anthropic. The statistical maps applied for the models' evaluation were selected using the following criteria: uniform area presented (Europe), two levels of enumeration units (NUTS-0, NUTS-2), two map types (choropleth, and proportional symbols), and consistency in terms of timeliness, file type (PNG), map language (English). The maps came from reliable sources: statistical offices (e.g., Eurostat) and statistical atlases. Four statistical maps were selected: two choropleth maps and two proportional symbols maps (with either NUTS-0 or NUTS-2 per each map type).

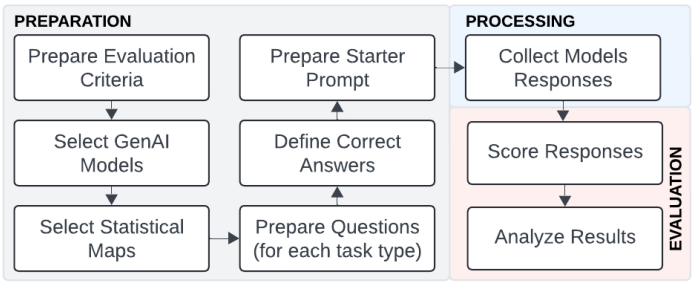

Figure 1. Workflow of the study.

Each model was asked six questions for each of the four statistical maps (24 questions in total), using the same starter prompt defining the rules for answering. The set of questions covered described above map use tasks, with two questions dedicated to each task type: map reading (e.g., *What symbols or additional elements are shown on the graphic outside the main statistical map?*), analysis (e.g., *What spatial patterns can be discerned in the distribution of regions with high and low Human Resources in Science and Technology (HRST)?*), and interpretation (e.g., *How can regional differences in HRST participation affect the economic development of the country's regions? Give specific three examples of countries where such regional differences are noticeable.*). The responses to all six questions for each map were scored according to a developed 6-point scale covering the following weighted evaluation criteria: correctness (30%), completeness and clarity (20%), support of answers with specific data (25%), justification and inference (25%). Consequently, each model could achieve a maximum score of 24 points for answers across all maps.

The collected results demonstrated fairly good and similar performance of GenAI models in retrieving information from statistical maps, with a slight advantage of Claude Sonnet 3.5 v2 (19.36 points out of 24) over GPT-4o (19.13 points) and Gemini 1.5 Pro (17.57 points). Claude Sonnet led in map reading and interpretation tasks, while GPT scored highest in map analysis questions, suggesting differentiated strengths that warrant deeper qualitative investigation in the

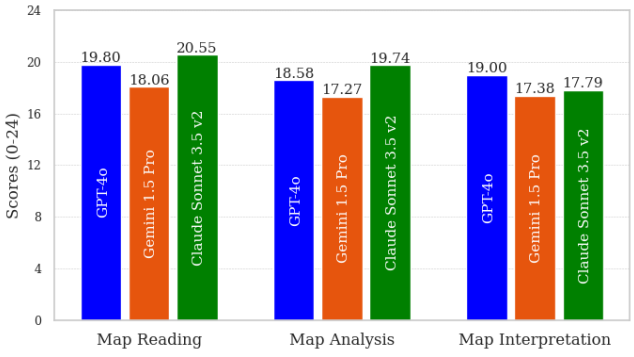

Figure 2. Performance of models across map use task types.

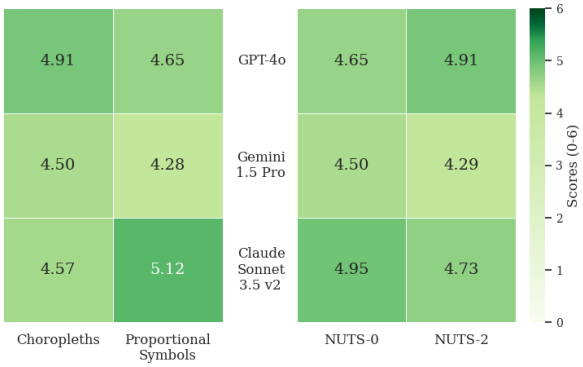

Figure 3. Performance of models across map types and enumeration unit levels.

future study. Gemini consistently performed the worst across all kinds of map usage (Fig. 2). However, the differences between the collected scores in each map use type did not exceed 2.5 points (out of 24).

For choropleth maps, models performed comparably (with scores ranging from 4.50 to 4.91), with GPT showing a modest edge (4.91 points out of 6). For proportional symbol maps, the scores diverged more considerably (from 4.28 to 5.12) than for choropleth maps, with Claude Sonnet emerging as the clear leader (5.12 points out of 6 points) and Gemini receiving the lowest score (4.28). At the level of countries (NUTS-0), the highest scores went to Claude Sonnet (4.95 points out of 6 points), while at the more detailed level of NUTS-2 units, GPT (4.91 points) was marginally better, ahead of Claude Sonnet (4.73 points) (Fig. 3). These collected results suggest that the effectiveness of the models may be affected by a type of map, an enumeration unit level, and complexity of a task. However, due to the limited number of maps and questions used in this exploratory stage, the generalizability of the findings remains constrained and will be addressed in the subsequent study.

The exploratory study reported here was a foundation for designing a full-scale study scheduled for spring 2025. It enabled the development and validation of the evaluation criteria, including scoring dimensions and the starter prompt structure. The main study will cover 12 generative AI models and 16 statistically diverse maps, selected to ensure proportional variation in cartographic methods, sources, and enumeration units. Each map will include 9 questions representing three levels of map use (reading, analysis, interpretation) and distinct cognitive operations (e.g., identifying, locating, associating, predicting). It will also feature expert evaluation of model responses, enabling a more comprehensive analysis of LLM-based generative AI models. This assessment will comprise both quantitative scoring and qualitative interpretation, supported by inferential statistics, to get insight into the capabilities of genAI models in extracting information from different kinds of thematic maps.

References

Buckley, A. & Kimerling, J. (2021). Map Reading. *Geographic Information Science & Technology Body of Knowledge, 2021*(Q1). https://doi.org/10.22224/gistbok/2021.1.8.

Kang, Y., Gao, S. & Roth, R. E. (2024). Artificial intelligence studies in cartography: A review and synthesis of methods, applications, and ethics. *Cartography and Geographic Information Science, 0*(0), 1–32. https://doi.org/10.1080/15230406.2023.2295943.

Kimerling, A. J., Muehrcke, P. C., Muehrcke, J. O., & Muehrcke, P. M. (2016). *Map use: Reading, analysis, interpretation*. ESRI Press Academic.

Kommisetty, P. D. N. K.,Vijay, A. & Rao, M. (2024). From Big Data to Actionable Insights: The Role of AI in Data Interpretation. *International Advanced Research Journal in Science, Engineering and Technology, 11*(8), https://doi.org/10.17148/IARJSET.2024.11831.

Logar, T., Aylett-Bullock, J., Nemni, E., Bromley, L., Quinn, J. & Luengo-Oroz, M. (2020). PulseSatellite: A Tool Using Human-AI Feedback Loops for Satellite Image Analysis in Humanitarian Contexts. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(9), https://doi.org/10.1609/aaai.v34i09.7101.

Robinson, A. C. & Griffin, A. L. (2024). Using AI to Generate Accessibility Descriptions for Maps. *Abstracts of the International Cartographic Association, 7*(139), https://doi.org/10.5194/ica-abs-7-139-2024.

Yilin, Y., Jianing, H., Yihan, H., Zhan, W., Shishi, X., Yuyu, L. & Wei, Z. (2024). Generative AI for visualization: State of the art and future directions. *Visual Informatics, 8*(2), 43–66, https://doi.org/10.1016/j.visinf.2024.04.003.

Wen Y., Zhou X., Li K., Li H. & Yan Z. (2024). Multi-task deep learning strategy for map-type classification, *Cartography and Geographic Information Science*, 51(6), 782–796, https://doi.org/10.1080/15230406.2024.2368574.

Zhang, Y., He, Z., Li, J., Lin, J., Guan, Q., & Yu, W. (2024). MapGPT: an autonomous framework for mapping by integrating large language model and cartographic tools. *Cartography and Geographic Information Science*, 51(6), 717–743. https://doi.org/10.1080/15230406.2024.2404868.