




# Large Language Models: Testing Their Capabilities to Understand and Explain Spatial Concepts

Majid Hojati   

Postdoctoral Fellow, School of Planning, University of Waterloo, Waterloo, Ontario, Canada

Rob Feick   

Associate Professor, School of Planning, University of Waterloo, Waterloo, Ontario, Canada

---

## Abstract

Interest in applying Large Language Models (LLMs), which use natural language processing (NLP) to provide human-like responses to text-based questions, to geospatial tasks has grown rapidly. Research shows that LLMs can help generate software code and answer some types of geographic questions to varying degrees even without fine-tuning. However, further research is required to explore the types of spatial questions they answer correctly, their abilities to apply spatial reasoning, and the variability between models. In this paper we examine the ability of four LLM models (GPT3.5 and 4, LLama2.0, Falcon40B) to answer spatial questions that range from basic calculations to more advanced geographic concepts. The intent of this comparison is twofold. First, we demonstrate an extensible method for evaluating LLM's limitations to supporting spatial data science through correct calculations and code generation. Relatedly, we also consider how these models can aid geospatial learning by providing text-based explanations of spatial concepts and operations. Our research shows common strengths in more basic types of questions, and mixed results for questions relating to more advanced spatial concepts. These results provide insights that may be used to inform strategies for testing and fine-tuning these models to increase their understanding of key spatial concepts.

**2012 ACM Subject Classification** Information systems; Information systems → Geographic information systems

**Keywords and phrases** Geospatial concepts, Large Language Models, LLM, GPT, Llama, Falcon

**Digital Object Identifier** 10.4230/LIPIcs.COSIT.2024.31

**Category** Short Paper

## 1 Introduction

In the rapidly advancing field of artificial intelligence (AI), Large Language Models (LLMs) are making significant progress, with applications extending across various fields. LLMs such as ChatGPT are forms of generative AI that can create human-like language responses [18]. LLMs are trained on large amounts of text, including books, news articles, and websites. They have demonstrated a strong understanding of human language, which allows them to be used for tasks such as reasoning, creative writing, code generation, translation, and information retrieval. Training LLMs can be multi-staged and engage varying degrees of human input. Through training, LLMs learn how words are combined in language, and they use these combinations to complete the language processing tasks. With more substantial training datasets, LLMs can recognize, interpret, and generate text with minimal or no specific fine-tuning. However, LLMs are vulnerable to a range of errors, including various types of factual inconsistency, misrepresentation errors, and geographic biases [16, 23, 6] [21].

Since LLM models contain embedded geographic knowledge and have shown abilities to apply it to geographic queries and reasoning tasks, interest has grown in using them for tasks as wide-ranging as interactive answering of geospatial questions, aiding learning of spatial concepts and software use, and translating natural language to spatial queries



© Majid Hojati and Rob Feick;

licensed under Creative Commons License CC-BY 4.0

16th International Conference on Spatial Information Theory (COSIT 2024).

Editors: Benjamin Adams, Amy Griffin, Simon Scheider, and Grant McKenzie; Article No. 31; pp. 31:1–31:9

Leibniz International Proceedings in Informatics



LIPIC Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 31:2 Large Language Models' Understanding of Spatial Concepts

(e.g., see [14, 13]). All of these use cases rely on LLM models being able to produce correct answers. Considering the limitations of current LLMs and their rapid evolution, an extensible framework is required to test different models' capabilities to answer spatial questions, perform geographic calculations, and aid users with more complex spatial reasoning problems. Since many geographic questions are context-dependent and can be answered in several feasible ways through data transformations [20], ongoing evaluation of this type will be needed to assess progress and understand evolving LLM geospatial capabilities. The main goals of this study are as follows:

- Introduce a methodology to automate testing of LLMs abilities to apply their geospatial knowledge to a range of spatial questions and reasoning tasks. A reproducible method allows us to test and compare various LLMs and different types of the tests over the time.
- Compare different LLMs and identify the spatial concepts that they need to be fine-tuned with.

### 2 Related research

With the increasing use of LLM models and the potential errors they carry, model evaluation has gained considerable attention recently (see [7, 24]). For example, Chang and Kidman [4] review a series of LLM evaluation methods, while [9] outline task-based evaluation for reasoning, medical usage, ethics, natural and social language fields. From the GIScience perspective, [25] used natural-language navigation tasks to evaluate LLMs' abilities to apply reasoning to spatial structures such as spatial and temporal distances, and shapes. Bhadari et al., [3] show that LLMs can complete spatial calculations correctly and, with limited accuracy, can apply common spatial prepositions (e.g., near, far) in queries. Aghzal et al., [2] examine LLMs' spatial reasoning abilities using end-to-end path planning tests in a grid environment and show that although fine-tuned LLMs can achieve impressive results in distributed reasoning tasks, they often struggle with long-term temporal reasoning and generalizing to more complex environments [2]. While Mai et al. showcase the potential of LLMs for GeoAI on various spatial semantic tasks [15], Li et al. [11] reported that ChatGPT had difficulty with questions that required more spatial reasoning abilities than more basic information retrieval.

### 3 Methodology

At a high level, our methodology consisted of two main stages: a) preparing questions as methodological ("how-to") and as spatial SQL problems, and b) evaluating the models' responses for both types of problems (see Figure 1). To recognize some of the range of LLM capabilities and the requirements to use their APIs, we compared two open source LLMs (Falcon-40B [19], Llama-2-7B [22]) and two of OpenAI's enterprise level models (ChatGPT 3.5 and 4). A set of 96 questions spanning a range of spatial concepts and operations were developed (Table 1). See the question list and model answers at:

[https://docs.google.com/spreadsheets/d/1hnRcIFB7-p6e5nE\\_ou\\_evxEC14S1DrxjEF0qcCFIWI/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1hnRcIFB7-p6e5nE_ou_evxEC14S1DrxjEF0qcCFIWI/edit?usp=sharing).

The following criteria were used to develop questions that: 1) targeted fundamental GIS concepts, 2) can be answered with SQL queries and not require complex modeling or coding, 3) have a quantitative answer that allows objective and automated or semi-automated evaluation. Questions were developed in part by drawing from the UCGIA's GIS&T Book of Knowledge <sup>1</sup> including sections such as "foundational concepts" (e.g., shape, direction,

---

<sup>1</sup> <https://gistbok.ucgis.org/knowledge-area/foundational-concepts>

distance), “data management” (e.g., coordinate systems), and analytics and modeling (e.g., buffers, spatial queries). While these sections do not encompass the breadth of the GIS&T BoK or GIScience more generally, they do span from the fundamentals of GIS through to some more advanced spatial concepts and operations. Finally, we endeavoured to include both “how to” questions that a non-expert may pose while learning GIS and “SQL” questions that users may deploy in an automated script that uses prompt engineering to perform tasks. To do this, each question was rephrased in two formats. First, to evaluate models’ abilities to aid geospatial learning, each question was formatted as a methodological or “how to” question to probe the models’ capacity to explain how a geographic calculation or problem could be addressed (see leftmost blue box in Figure 1 above). As noted by [4, 17, 14], LLMs are susceptible to producing overly generic or at times incorrect output. Second, to examine models’ potential to aid analytical tasks, each question was reformatted slightly to ask the LLMs to write an SQL query that would solve a problem (see leftmost green box in Figure 1). To simplify testing of the generated queries, we specifically asked for queries to be written for the PostGIS extension to PostgreSQL.

The Langchain Python framework was used to automate the process of LLM evaluation. Functions were created that drew “how to” and SQL questions from csv files, supplied these questions as prompts to the models’ APIs, and output their answers to another set of csv files. Since LLMs are evolving rapidly and replicability is important: 1) the entire pipeline is automated and can run multiple times, 2) the question set can be changed or improved over time, and 3) the SQL variants of the questions allow unambiguous response evaluations that minimize human biases.

Specific prompts were used to require models to evaluate each question independently and to differentiate between question types. To preclude models from drawing upon a history of preceding questions, each call to an LLM was a “cold start” [8, 17]. This was done for methodology questions by prefacing each question with the following prompt: *Without including any of the previous conversations, provide a methodology to answer the following question:*

For the SQL questions, the prompt was modified to focus specifically on PostGIS functions that align with the environment we used for output evaluation: *Without including any of the previous conversations, write a SQL query to answer the following question. Make sure to only use available SQL functions and PostGIS spatial functions:*

Finally, if a question included a specific data model or any assumption we used the following prompt: *Without including any of the previous conversations, write a SQL query to answer the following question. Make sure to only use available SQL functions and PostGIS spatial functions. Assume that our database includes the following tables:*

*Table name: resorts, Columns: id: main identifier column, name: name of the ski resort column, geom: geometry column, point geometry, with CRS 4326*

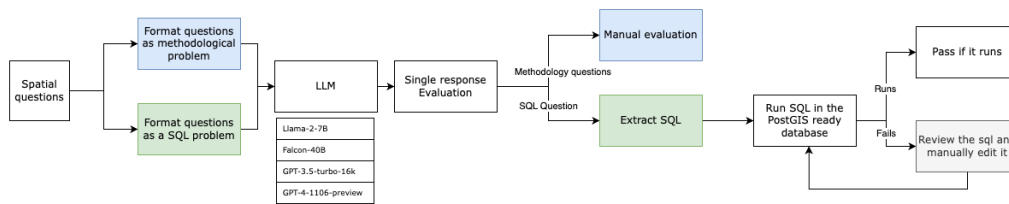
A few sample questions are shown below.

**Direction:** *Assume that the following coordinates are longitude and latitude, is the point(-150,30) located north of the point(-130,25)?*

**Area:** *What is area of this shape: polyline[[-150,30],[-155,35],[-155,25]]? Assume that the coordinates are longitude and latitude*

**Projection:** *Convert these coordinates point([-71,41]) to web mercator projection. Assume that the coordinates are longitude and latitude*

**Geohash:** *Calculate bounding box of the following geohash 9qqj7nm.xncgyy4d0dbxqz0*



■ **Figure 1** The overall workflow of the study.

■ **Table 1** The spatial concepts included in the question set.

Spatial concept	#	Spatial concept	#	Spatial concept	#
Area	7	Proximity	2	Distance	9
Angle	5	Projection	2	Geometry/Format	8
Boundary Box	2	Graph/Network	3	Center	6
Matrix Translation	3	Error/Accuracy	9	Geometry Validation	3
Interpolation, Variogram	2	3d Coordinates	5	Geohash	2
Direction	7	Spatial topologies	13	Midpoint, Dimension	3

The responses provided by the LLMs were evaluated through a series of six types of tests. This approach was developed following initial pilot-testing where model responses included providing no answer, partial answers, as well as complete answers that may be correct or incorrect. Test-1 focuses on if a model provided a final answer. Test-2 determined if a model’s response was correct by comparing it to the value returned when the authors ran the generated query. For Test-3, answers to the ‘How to’ methodological questions were evaluated manually by the first author. Test-4 assesses if a model can generate useful SQL queries. For this test, a small script that used common markdown code separators was run to parse SQL statements from the larger text responses they were embedded within (see Figure 2 below). Since the questions to evaluate the spatial functions were mainly developed to be commutable via SQL queries, Test-5 entailed executing the SQL query extracted in Test 4 in a PostGIS-enabled database. Test-6 evaluated the level of the edits that a SQL query needed to be able to return the correct answer. For example, a generated SQL query that included a misspelled function or a missing brackets in the SQL would not pass this test.

Since all of the questions are designed to return a numeric answer, the evaluation and scoring of answers was straightforward. To validate the final answers, the authors determined the correct answers from the relevant SQL queries and compared them with the model responses. If the parsing script mentioned above for Test-4 was not able to extract the SQL query from a model’s output, the test was marked as failed and the authors extracted the SQL text manually before executing the query in Test-5. Each test was assigned a maximum score of one, with values of .5 awarded where the answer was almost, but not completely, correct. In future studies with more specific testing needs, it may be appropriate to apply different weights to each test.

## 4 Results

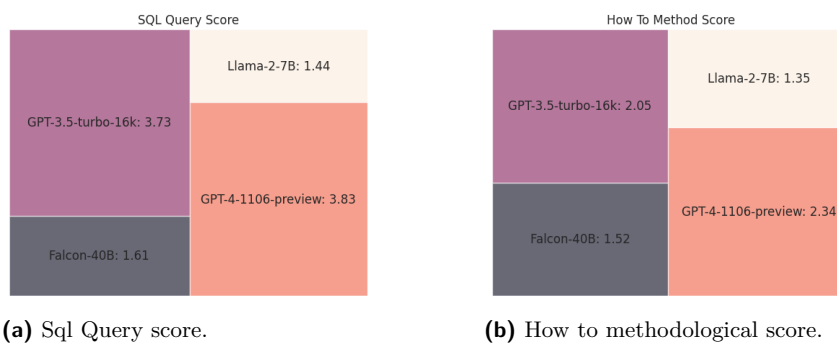
Figure 3 shows the overall performance of each model through a calculated average score model for: a) SQL query questions, and b) ‘How to’ methodology questions. GPT-4 performed better than the other three models for both question types with GPT-3.5 having the next

To check whether two polylines are equal in PostGIS, you can use the `ST_Equals` function, which returns true if the given geometries represent the same geometry structure. Assuming that the polylines are represented as `LINestring` geometries, here's an example of a SQL query that checks for equality:

```
SELECT ST_Equals(
  ST_GeomFromText('LINestring(-150 30, -155 35, -155 25)'),
  ST_GeomFromText('LINestring(-150 30, -155 35, -155 25)')
) AS are_polylines_equal;
```

This query will return a single row with a boolean value: `true` if the two polylines are equal, or `false` if they are not. Since the coordinates of the two polylines you provided are identical, the result of this query should be `true`.

■ **Figure 2** An Example of GPT-3 answer to a question which specifically asked for a sql query.

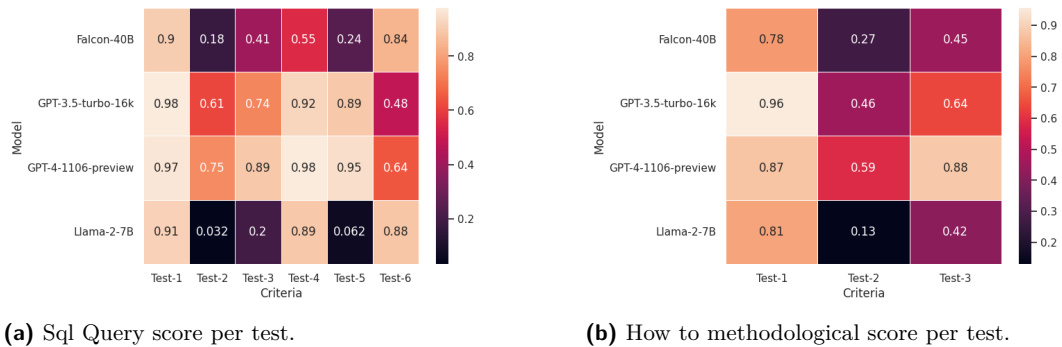


■ **Figure 3** The overall performance of the LLM models for both SQL and methodological questions.

best scores. Interestingly, GPT-3 displayed the greatest difference between the two main question types and was very close to GPT-4’s performance in answering spatial SQL questions. Both Falcon-40B and Llama-2-7B had lower levels of performance, with Llama having the lowest scores for both question types. Overall, all four models performed substantially better (46% in aggregate) at generating valid spatial SQL code than when given less well-defined methodological tasks.

Figure 4 shows disaggregated views of model performance for each test. In Figure 4a), each row reports the scores a model achieved across Tests 1 to 6 as described in the preceding section. On average, all the models provide an answer for 90% or more of the questions (column 1). However, more separation is apparent when examining if the provided answers are correct. GPT-4 answered 75% of the questions correctly, followed by GPT-3.5 at 61%. Falcon and Llama had far fewer correct answers at 18% and 3% respectively. Similarly in the third test which checks if the provided methodology is correct, GPT-4 (89%) and GPT-3.5 (74%) performed best, with Falcon (41%) and Llama (20%) demonstrating somewhat better capabilities to return a valid methodology even if they were less able to produce a correct answer. With respect to generating PostGIS SQL code that ran without edits (column 4), GPT-4 (95%) and GPT-3.5 (89%) were considerably better than the other two models at 24% (Falcon) and 6% (Llama). As one indicator of the ease of porting model outputs to automated processes, the query separator test showed that only Falcon had lagging scores with just 55% of the responses having a proper separator.

## 31:6 Large Language Models' Understanding of Spatial Concepts



**Figure 4** The breakdown of the LLM models per each test case. a) Breakdown for the SQL query questions and b) the breakdown for the methodological query questions.



**Figure 5** The average score of each model per each spatial concept. a) Methodology based questions and b) SQL based questions.

In Panel 4b) of Figure 4, scores for the three methodology questions are reported. Scores for the first methodology criterion, namely does the model return an answer, were relatively close with GPT-3.5 being the most eager to provide an answer (96%), with the other models providing roughly similar counts (78%-87%). In terms of correct answers, Llama (13%) and Falcon (27%) trailed both OpenAI models by a considerable margin. Both Llama and Falcon models provide a better (close to 45%) performance in explaining the correct methodology compared to calculating the correct answer.

Next, we looked into each model's performance relative to each spatial concept in the question set (Figure 5). Overall, the models performed similarly across most of the concepts with a better score in the SQL type of questions than the methodological questions. Perhaps understandably, the models tended to fare more poorly when tasked with explaining how to use spatial relationships in general. However, they were more successful at operationalizing spatial concepts through spatial SQL queries they generated to answer questions.

## 5 Discussion

As was mentioned before, LLMs are prone to errors including hallucinations (e.g., see [10]). In this experiment, all models suffered from hallucination errors. This was especially evident in the queries Falcon and Llama generated as they suggested using non-existent PostGIS functions (e.g., `PostGIS_Distance`, `rhumb_line_distance`). Both GPT models suffered less with only about 1% of such errors in contrast to Falcon (13%). In some cases, Falcon had difficulties distinguishing the question from the real-world issue when it was asked specifically for a SQL query. For example, Falcon tried to find the name of the places in the center of a

town instead of calculating its coordinates as requested via the PostGIS ST\_Centroid function. LLM models are prone to provide more accurate and measurable responses for structured language tasks [26] which is evident in the models's better performance in generating SQL queries. This provides an advantage for automating certain types of tasks as demonstrated by [12].

Overall, all of the models were able to use basic geospatial concepts including ability to distinguish different shapes, direction, angle and topological relations. All of the models are able to distinguish between latitude and longitude space and understand multiple geometry formats such as WKT or Geojson formats. However, they were less able to calculate the expected final correct value. Some of the models, such as GPT-4, were able to construct example datasets to support their own SQL answers or suggest specific data structures for certain tables to be able to answer the questions.

## 6 Conclusion

This study aimed to measure the accuracy of LLM models in responding to the geospatial concepts. Around 100 spatial problems were used to evaluate GPT-3, GPT-4, Llama 2 and Falcon answers by providing a methodology and a SQL query answer. Our results showed that in general GPT-3 and GPT-4 had overall scores that were above the average. In contrast, the open source models failed to provide a considerable number of correct answers. All of the models have better results in writing SQL queries as an structured language.

Finally, as Mooney et al.[17] and Chang and Kidman [4] note, LLM models can play important roles in geospatial training and education too. The evaluation of LLM models and their understanding of geographic concepts allows us to use them for educational purposes such as teaching spatial databases, geographic concern (e.g., see [1, 5]). To date, it appears that considerable care is required for LLM use in these contexts as this study and others highlight the potential for these models to be prone to errors and having incomplete understandings of spatial concepts and tasks at, or above, an intermediate level. Their ability to provide answers to more structured problems, such as writing code, or providing explanatory guidance does appear more promising though.

---

## References

- 1 Alaa Abd-alrazaq, Rawan AlSaad, Dari Alhuwail, Arfan Ahmed, Pdraig Mark Healy, Syed Latifi, Sarah Aziz, Rafat Damseh, Sadam Alabed Alrazak, and Javaid Sheikh. Large language models in medical education: Opportunities, challenges, and future directions. *JMIR Medical Education*, 9:e48291, June 2023. doi:10.2196/48291.
- 2 Mohamed Aghzal, Erion Plaku, and Ziyu Yao. Can large language models be good path planners? a benchmark and investigation on spatial-temporal reasoning. *arXiv preprint arXiv:2310.03249*, 2023.
- 3 Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. Are large language models geospatially knowledgeable? In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, SIGSPATIAL '23*, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3589132.3625625.
- 4 Chew-Hung Chang and Gillian Kidman. The rise of generative artificial intelligence (ai) language models - challenges and opportunities for geographical and environmental education. *International Research in Geographical and Environmental Education*, 32(2):85–89, 2023. doi:10.1080/10382046.2023.2194036.
- 5 Karl de Fine Licht. Integrating large language models into higher education: Guidelines for effective implementation. In *IS4SI Summit 2023*, IS4SI Summit 2023. MDPI, August 2023. doi:10.3390/cmsf2023008065.

- 6 Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. Evaluating factuality in text simplification. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.506.
- 7 Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2023. arXiv:2306.13394.
- 8 Nir Fulman, Abdulkadir Memduhoğlu, and Alexander Zipf. Distortions in judged spatial relations in large language models: The dawn of natural language geographic data?, 2024. arXiv:2401.04218.
- 9 Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics, 2023. arXiv:2310.05694.
- 10 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023. arXiv:2311.05232.
- 11 Fangjun Li, David C. Hogg, and Anthony G. Cohn. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark, 2024. arXiv:2401.03991.
- 12 Zhenlong Li and Huan Ning. Autonomous gis: the next-generation ai-powered gis. *International Journal of Digital Earth*, 16(2):4668–4686, 2023.
- 13 Mengyi Liu, Xieyang Wang, Jianqiu Xu, and Hua Lu. Nalspatial: An effective natural language transformation framework for queries over spatial data. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, SIGSPATIAL '23*, New York, NY, USA, 2023. Association for Computing Machinery. doi:10.1145/3589132.3625600.
- 14 Ollie Liu, Deqing Fu, Dani Yogatama, and Willie Neiswanger. Dellma: A framework for decision making under uncertainty with large language models, 2024. arXiv:2402.02392.
- 15 Gengchen Mai, Chris Cundy, Kristy Choi, Yingjie Hu, Ni Lao, and Stefano Ermon. Towards a foundation model for geospatial artificial intelligence (vision paper). In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '22*, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3557915.3561043.
- 16 Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.173.
- 17 Peter Mooney, Wencong Cui, Boyuan Guan, and Levente Juhász. Towards understanding the geospatial skills of chatgpt: Taking a geographic information systems (gis) exam. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, GeoAI 2023, Hamburg, Germany, 13 November 2023*, November 2023. doi:10.1145/3615886.3627745.
- 18 OpenAI. Openai chatgpt, 2024.
- 19 Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023. arXiv:2306.01116.



- 20 Simon Scheider, Enkhbold Nyamsuren, Han Kruiger, and Haiqi Xu. Geo-analytical question-answering with gis. *International Journal of Digital Earth*, 14(1):1–14, March 2020. doi: 10.1080/17538947.2020.1738568.
- 21 Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158, 2023.
- 22 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.
- 23 Rongxiang Weng, Heng Yu, Xiangpeng Wei, and Weihua Luo. Towards enhancing faithfulness for neural machine translation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2675–2684, Online, November 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.212.
- 24 Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models, 2023. arXiv:2306.09265.
- 25 Yutaro Yamada, Yihan Bao, Andrew Kyle Lampinen, Jungo Kasai, and Ilker Yildirim. Evaluating spatial understanding of large language models. *Transactions on Machine Learning Research*, 2024. URL: <https://openreview.net/forum?id=xkiflfKCw3>.
- 26 Fuheng Zhao, Lawrence Lim, Ishtiyaque Ahmad, Divyakant Agrawal, and Amr El Abbadi. Llm-sql-solver: Can llms determine sql equivalence?, 2024. arXiv:2312.10321.