Understanding LLMs' capabilities to support spatially-disaggregated epidemiological simulations

Anuj Goenka[1 [0000−0002−4508−0361]], Ilya Zaslavsky[1 [0000−0003−4191−8275]], Jiaxi Lei[1 [0009−0009−1113−0058]], Rishi Graham[2 [0000−0003−3532−7313]], and Eliah Aronoff-Spencer[2 [0000−0002−6279−5027]]

[1] San Diego Supercomputer Center, University of California San Diego, La Jolla, CA 92093, USA
[2] School of Medicine, University of California San Diego, La Jolla, CA 92093, USA

Abstract. Evaluating public health interventions during disease outbreaks requires an understanding of the spatial patterns underlying epidemiological processes. In this study, we explore how Large Language Models (LLMs) can leverage spatial understanding and contextual reasoning to support spatially-disaggregated epidemiological simulations. We present an approach in which we query LLM to determine appropriate mitigation strategies, informed by local profiles and the current outbreak status. Through a series of experiments with COVID-19 data from San Diego County, we show how different LLMs perform in tasks requiring spatial adaptation of mitigation strategies, and how incorporating connectivity information through Retrieval-Augmented Generation (RAG) enhances the performance of these customizations. The results reveal significant differences among LLMs in their ability to account for spatial structure and optimize mitigation strategies accordingly. This highlights the importance of selecting the right model and enhancing it with relevant contextual information for effective public health interventions.

Keywords: Large Language Models (LLMs) · Epidemiological Modeling · Spatial Reasoning · System Dynamics · San Diego

1 Introduction

Accurately modeling epidemiological processes is vital for predicting outbreak dynamics, evaluating intervention strategies, and ultimately controlling the spread of infectious disease. Metapopulation models implementing the SEIR (Susceptible, Exposed, Infected, Recovered) framework and utilizing systems dynamics methodology have been a common approach to simulating disease spread. The SEIR model divides the population into compartments, simulating the movement of individuals between these compartments over time, based on factors such as disease transmission rates, incubation periods, and recovery rates [1].

Disease spread is often considered a spatial process, dependent on interactions between individuals within specific geographic contexts. However, compartmental SEIR models, with their multiple stocks, flows, and feedback loops, typically lack the granularity necessary to capture spatial disease patterns and local dynamics. The complexity of SEIR models, along with the vast amount of data required to define compartments and interactions across different spatial units, significantly impacts model performance and sensitivity. This complexity makes spatially-disaggregated SEIR simulations relatively uncommon in the context of System Dynamics (SD) modeling [2], though such simulations are more frequently implemented using agent-based models.

As Large Language Models (LLMs) exhibit remarkable contextual knowledge and reasoning capabilities, leveraging them as a source of local knowledge is an appealing approach. Models like OpenAI's GPT series have demonstrated significant potential in responding to prompts that require locational knowledge [3] [4] [5]. However, it remains unclear whether their local knowledge is sufficient for spatially-disaggregated disease forecasting and whether they can adequately capture spatial relationships to enable realistic simulations of disease spread without the need for explicitly incorporating spatial data and relationships.

In this paper, we develop a series of experiments designed to answer these questions. The next section describes the experimental setup and challenges associated with coupling LLMs and a simulation engine. We then compare public health interventions for managing disease outbreaks, which are selected by different LLMs for different areas, and show how these selections are influenced by adding spatial neighborhood information. A discussion of the observed results is followed by conclusions and future work. Ultimately, our findings aim to highlight the strengths and limitations of LLMs in how they use spatial information for public health applications. We also propose directions for refining these models to better integrate spatial dynamics in future simulations.

2 Methodology

2.1 Modeling context

The initial system dynamics model used in this experiment was developed to support the evaluation of public health interventions within the context of a serious game. Several teams consisting of San Diego public health professionals, county health officials, and researchers were presented with disease outbreak scenarios. Their task was to respond by prioritizing various mitigation measures. The scenarios were primarily based on the COVID-19 pandemic, and the mitigation strategies were generally aligned with recommendations documented in the National COVID-19 Preparedness Plan [6] and included such measures as contact tracing, public health surveillance, public health messaging, medical interventions, improvements in healthcare system preparedness, and investments into scientific infrastructure. During each stage of the game, teams would allocate resources ("game units") to their chosen strategies and submit their decisions

to game judges. The judges would then simulate the effects of these strategies using a model implemented in Stella [7].

The model itself is substantial, comprising 32 stocks, 56 flows, and 209 converters. Additionally, arrays were used to represent infection at four different severity levels across three age groups, resulting in a total of 120 stocks, 254 flows, and 412 converters. While validated with COVID-19 data, the model is flexible enough to simulate future outbreaks by adjusting pathogen characteristics such as infectivity, contact rates, and the duration of asymptomatic contagiousness. However, the model currently lacks spatial information, despite notable variability across San Diego County in factors like population density and socio-demographic and economic characteristics.

2.2 Experimental setup

The integration of LLMs with agent-based models has been discussed in the literature [8] [9] [10] including within the context of epidemiological modeling [11]. In contrast, the integration of LLMs with system dynamics models has received considerably less attention. Several recent papers have demonstrated how models like GPT-4 can assist with problem formulation, simulation design, and analysis of system dynamics models [12] [13]. However, real-time connections between LLMs and Stella-based simulations have yet to be fully explored. In our experience, establishing such a connection poses significant implementation challenges.

Our initial setup involved using an R-based wrapper around Stella, implemented through R's deSolve library and accessed via Python through the RPy2 interface. This approach was hindered by poor performance and limited flexibility in updating the model dynamically. We then explored two Python libraries designed for importing and running system dynamics models: PySD [14] and BPTK-Py [15]. Both attempts were unsuccessful due to the complexity of the initial model and its reliance on Stella-specific features such as dimensions and conveyors, which these libraries either do not support or implement differently.

To overcome these challenges, we re-implemented the model entirely in R and accessed it via Python using the RPy2 wrapper. This new design facilitated easier model updates and smoother communication between the SEIR model and other components of the pipeline, such as the Retrieval-Augmented Generation (RAG) system. The RAG system played a key role in delivering context-aware prompts by accessing spatial and demographic data on demand, allowing the LLMs to tailor their resource allocation recommendations for each specific ZIP code.

Spatial data was retrieved using the GeoPandas library and fed into the RAG system. Compared to the large volumes of data required by the initial SEIR model to simulate each spatial object, the data provided via RAG was considerably smaller. It included ZIP code boundaries, key demographics, contiguity graphs for ZIP codes, and healthcare facility locations.

The only major adjustment to the original system dynamics model was computing it for each ZIP code based on that ZIP code's population, rather than

using the aggregate population of San Diego. All other parameters—including infection rates, recovery rates, and transmission dynamics—remained consistent across ZIP codes, based on the average for San Diego County. This approach allowed for a controlled assessment of how resource allocation and mitigation strategies varied based on population size and geographic location, while maintaining epidemiological consistency.

We utilized two open-source LLMs—Meta Llama 3.1: 70B Instruct (referred to as "Llama" henceforth) and Mixtral 8x7B 32k (referred to as "Mixtral" henceforth)—hosted on dedicated LLM infrastructure at the San Diego Supercomputer Center (SDSC). This infrastructure provided API access to the LLMs and supported large token capacities (up to 32k tokens for Mixtral and up to 120k tokens for Llama). In addition to token capacity, the two models differ in size, internal structure (a single large model for Llama versus a mixture of 8 smaller models in Mixtral), performance, and specialization. The integration between the system dynamics model and the LLMs was managed through a custom Python orchestration module, ensuring seamless communication. Figure1 illustrates the main components of the experimental setup. Further, we compared performance of these models with two closed-source LLMs, OpenAI's GPT-4 and Anthropic's Claude 3.5, accessed via their web interfaces for a subset of prompts.

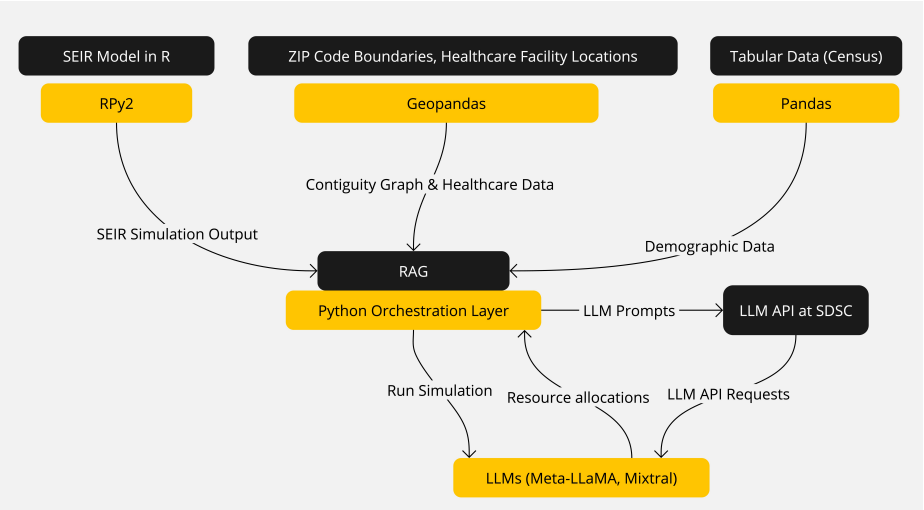

Fig. 1. Main components and information flows in the LLM-SD experimental setup.

The experiments involved running simulations for scenarios where an LLM allocated a total of 10,000 game units across various strategies for each selected ZIP code. LLM prompts were designed to include information about the population and SEIR model outputs such as infection counts, for each ZIP code.

To assess the spatial capabilities of different LLMs, we ran experiments using multiple models.

3 Results

Although the earlier setup (a Stella model wrapped in deSolve and accessed via RPy2, integrated with GPT 3.5 and Claude 2) did not perform well overall, it demonstrated that the LLMs were capable of generating meaningful spatially-disaggregated scenarios. This setup revealed several interesting patterns, such as prioritizing research efforts and vaccine distribution among students in ZIP codes with universities, even though socio-demographic and economic profiles were not included in the system dynamics simulations nor provided to the LLMs via RAG. These initial findings highlight the potential of LLMs to recognize spatial patterns with minimal input data. Building on these results, we present findings for all ZIP codes within San Diego County (Fig. 2). However, for the subsequent experiments, we focus on three specific ZIP codes as case studies. ZIP code 92101, located in downtown San Diego, features a population with diverse socio-demographic characteristics. ZIP code 92091 (Rancho Santa Fe) is characterized by an older population with a higher median income. ZIP code 92093 (University of California, San Diego) is predominantly occupied by younger individuals, mainly students living in dormitories. By focusing on diverse ZIP codes, the experiments evaluate the models' ability to handle a range of demographic variables and spatial contexts, providing a comprehensive assessment of the LLMs' capabilities.

3.1 Comparing LLM Model Performance in Spatial Disaggregation Tasks

The evaluation of LLMs, including GPT, Claude, Llama, and Mixtral, revealed significant differences in making use of spatial information and generating strategies for mitigating disease transmission across ZIP codes in San Diego County. GPT-4 and Claude 3.5 demonstrated superior spatial awareness, accurately recognizing ZIP codes like 92093 as linked to the University of California, San Diego, and tailoring their strategies to the demographic and infrastructural contexts of university areas, such as prioritizing student interventions.

In contrast, Llama and Mixtral struggled with spatially-specific tasks, often providing generic recommendations that overlooked the unique challenges of particular ZIP codes. For example, while GPT-4 suggested targeted public health strategies for university areas, Llama failed to account for their academic and demographic uniqueness, resulting in less effective mitigation plans.

3.2 Limitations in Spatial Reasoning and Data Integration

During an experiment where the LLMs were tasked with allocating specific resource units across multiple stages based on dynamic SEIR model outputs, both

Llama and Mixtral exhibited a tendency to hallucinate. Specifically, the models frequently overshot the available resource units, failing to account for resources allocated in earlier stages. This issue persisted despite detailed prompts, highlighting the models' difficulty in managing cumulative resource allocation over time while adhering to real-world constraints.

Another key limitation was the models' inconsistent spatial reasoning, particularly in balancing complex data inputs. While they demonstrated basic competence in considering single variables, such as recognizing the importance of neighboring ZIP codes when limited local information was provided, they struggled when more intricate spatial relationships were required. For example, when balancing healthcare utilization and population demographics across regions, the models often produced suboptimal allocations.

Prompt tuning was employed to mitigate this issue. By instructing the models to review their initial recommendations and verify that all critical factors—SEIR model outputs, healthcare facilities, neighboring ZIP codes, and spatial context—had been adequately considered, the LLMs showed moderate improvements. This additional layer of review encouraged a more balanced and comprehensive approach, yet the models still struggled to fully integrate complex spatial data and maintain resource allocation constraints.

3.3 Performance Comparison: Explicit vs. Inferred Neighborhood Data

This experiment compared the performance of Llama and Mixtral under two conditions: one where a contiguity graph for ZIP codes was explicitly provided and another where the models were required to infer neighboring ZIP codes using their internal spatial reasoning capabilities. The objective was to assess how effectively the models could integrate spatial context into their resource allocation strategies.

The analysis of percent similarity across resource allocation vectors revealed that Mixtral struggled to utilize the contiguity graph effectively. Its allocations for all three ZIP codes fell within a narrow similarity range (88-90%). In contrast, Llama's performance was more affected by the contiguity information, showing lower similarity (70-75%) for two ZIP codes (92101 and 92091). Interestingly, when explicit contiguity information was provided, the similarity between Llama's and Mixtral's allocations dropped to as low as 40% for ZIP code 92091, suggesting that Llama adjusted its allocations more significantly in response to the added spatial data.

When the models were not provided with a contiguity graph, both Llama and Mixtral generated more similar allocations, with around 80% similarity across all ZIP codes. This suggests that, without explicit guidance, both models defaulted to simpler spatial reasoning, resulting in less variation in their decisions.

An additional observation is that the explicit inclusion of the contiguity graph led Llama to favor interventions that addressed cross-boundary needs, such as Widespread Testing Initiatives, Comprehensive Contact Tracing Programs, and Expansion of Healthcare Capacity. In contrast, omitting the contiguity graph

resulted in a focus on more localized strategies, such as Mental Health Support Services and Telehealth Services Expansion, which are easier to confine to specific ZIP codes.

While there were observed advantages to including the contiguity graph, both Llama and Mixtral still struggled to fully leverage this spatial data. For example, in ZIP code 92093 (UCSD), which lacks long-term healthcare facilities, we expected the models to recognize the availability of prominent UCSD Health resources in neighboring ZIP code 92037. However, the models continued to recommend significant investments in expanding healthcare facilities in 92093, overlooking the proximity of resources in 92037. This misallocation indicates that the models had difficulty accounting for healthcare access across ZIP code boundaries, even with explicit spatial guidance.

Overall, the inclusion of contiguity information enabled better coordination of resources across ZIP code boundaries, though this capability remained limited. Without explicit spatial data, the models frequently failed to consider the healthcare capacity and demographic characteristics of neighboring areas. Additionally, response times were nearly three times longer when contiguity data was omitted, underscoring the models' difficulty in inferring spatial relationships without explicit guidance.

3.4 Resource Allocation Across Multiple Runs

Table 1. Averages and relative range percentages(Avg, % Range) generated by Llama and Mixtral for commonly preferred strategies, for three test zip codes

Strategy	**92091**		**92093**		**92101**	
	Llama	**Mixtral**	**Llama**	**Mixtral**	**Llama**	**Mixtral**
Widespread Testing	2006, 100	1500, 67	2120, 94	1505, 33	2084, 96	1505, 33
Public Health Campaigns	1685, 148	1011, 49	1608, 199	1000, 0	1499, 100	998, 20
Social Distancing	1225, 163	1011, 49	1170, 103	1000, 0	1229, 98	1000, 0
Quarantine Facilities	1162, 112	1885, 27	1450, 103	1900, 26	1000, 200	1474, 102
Healthcare Capacity	1813, 138	2214, 45	1857, 124	2000, 0	1869, 118	2021, 25
Home Isolation	1067, 187	928, 54	977, 154	974, 51	618, 162	712, 70
Telehealth Expansion	1064, 235	536, 93	1095, 110	500, 0	917, 120	500, 0
Vaccination Research	1250, 120	-	1389, 72	-	1317, 228	-
Contact Tracing	1430, 175	1415, 71	1509, 99	1368, 73	1304, 192	1135, 88

The Llama and Mixtral models were each run 100 times to evaluate the patterns of resource allocations they generated. Both models were run under identical conditions, including the same temperature parameter—a key factor influencing the randomness of generated outputs. While both models allocated resources to similar strategies, Llama's results showed notably higher variability in allocations—measured as percent range—especially in strategies like Healthcare Capacity, Contact Tracing, and Public Health Campaigns. Conversely, Mix-

tral exhibited relatively stable allocations, with narrower ranges across the same strategies (Table 1). These differences can be attributed to the underlying architecture and training objectives of the two models. Llama, being a larger model with 70 billion parameters, is designed to handle more complex reasoning and can capture a broader range of potential outcomes, leading to more diverse outputs across multiple runs. In contrast, Mixtral, with fewer parameters, produced less fluctuating results across simulations.

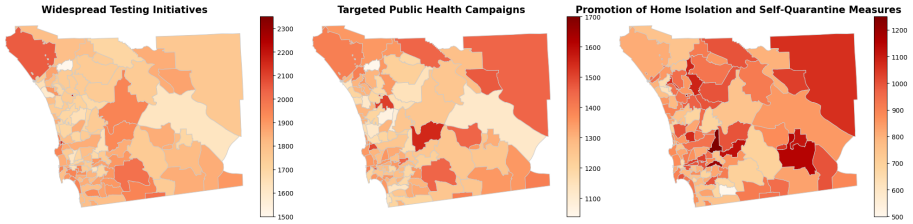

Fig. 2. Resource allocation of three pandemic response strategies preferred by Llama

Fig. 2 illustrates the spatial patterns of allocations for three selected strategies preferred by Llama (as mentioned above, Mixtral simulations demonstrated its limited spatial reasoning capabilities).While the allocations can generally be explained by existing differences in vulnerable populations, average income, median age, population density, and proximity to healthcare facilities, in many cases the patterns appear spurious and may be referred to as hallucinations, as they result from the lack of contextual understanding and incomplete training data about spatial relationships.

4 Discussion

Recent research has significantly advanced the understanding of LLMs' spatial knowledge and reasoning capabilities compared to earlier foundational work with models like BERT and GPT-3, which primarily focused on natural language processing tasks [16] [17]. Subsequent studies have explored various geospatial tasks, providing benchmarks for LLMs' performance in understanding mapping concepts, spatial analysis, and location-based reasoning [18] [3] [5]. Other research has focused on interpreting spatial predicates [4], navigation tasks [19], qualitative spatial reasoning [20], exploring and categorizing geospatial embeddings [21] [22], and translating natural language into spatial SQL queries [23]. While these studies underline LLMs' potential for leveraging spatial knowledge and understanding spatial relationships in a variety of applications, they also highlight current limitations. In this paper, we extended this exploration by integrating LLMs with a system dynamics platform to simulate spatial patterns of disease outbreaks. Our findings were consistent with previous research, showing that while LLMs can process spatial data and leverage spatial relationships to a

certain degree, the results varied significantly across regions. The LLMs demonstrated their ability to integrate non-explicit spatial data into simulations, facilitating a geographically distributed analysis of disease spread and suggesting meaningful and nuanced mitigation measures. At the same time, the models we explored demonstrated different performance under different scenarios, with and without spatial contiguity information explicitly incorporated.

5 Future Directions

In future work, we aim to integrate the system dynamics model more closely with a Python orchestration layer to query LLMs at key decision points. This will allow for dynamic scenario adjustments during multiple continuous runs of the model, enabling more responsive and adaptive decision-making for mitigation strategies.

To further enhance LLMs' performance in spatial contexts, future work should focus on incorporating more sophisticated spatial datasets and training models specifically for spatial reasoning tasks. We are already exploring several improvements to our approach. For example, we are considering adjusting resource allocations by population size or other relevant metrics to better reflect realistic spatial relationships. Additionally, tuning LLM prompts to be more precise and task-specific, rather than general geospatial context prompts, may improve LLMs' performance without relying on model fine-tuning.

Lastly, we are experimenting with additional measures of spatial connectedness beyond the contiguity graph used in our initial experiments. For instance, we plan to leverage data from a local transportation model in San Diego County, which forecasts travel demand between the 4,500 Master Geographic Reference Areas (MGRAs), to enhance the representation of population movement in our simulations. These advancements will help address the current limitations and improve the effectiveness of LLM-assisted spatial simulations.

Acknowledgments. This work was supported by the National Science Foundation (NSF) under awards 2139740 and 2200197, and by the Centers for Disease Control and Prevention under award NU38FT000006. LLM service provided by SDSC LLM at San Diego Supercomputer Center, UC San Diego, is gratefully acknowledged. The code developed for this paper and the SD model is available at:
https://github.com/covABM/generative-seriousgame

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Herbert W. Hethcote. The Mathematics of Infectious Diseases. *SIAM Review*, 42(4):599–653, January 2000. Publisher: Society for Industrial and Applied Mathematics.

2. Faizeh Hatami, Shi Chen, Rajib Paul, and Jean-Claude Thill. Simulating and Forecasting the COVID-19 Spread in a U.S. Metropolitan Region with a Spatial SEIR Model. *International Journal of Environmental Research and Public Health*, 19(23):15771, November 2022.
3. Peter Mooney, Wencong Cui, Boyuan Guan, and Levente Juhász. Towards Understanding the Geospatial Skills of ChatGPT: Taking a Geographic Information Systems (GIS) Exam. June 2023. Publisher: EarthArXiv.
4. 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, pages 1–4, New York, NY, USA, December 2023. Association for Computing Machinery.
5. Wes Gurnee and Max Tegmark. Language Models Represent Space and Time, March 2024. arXiv:2310.02207.
6. The White House. National COVID-19 Preparedness Plan. Technical report, 2022.
7. Stella Architect, 2024.
8. Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.
9. Boyu Wang, Vincent Hess, and Andrew Crooks. Mesa-Geo: A GIS Extension for the Mesa Agent-Based Modeling Framework in Python. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on GeoSpatial Simulation*, GeoSim '22, pages 1–10, New York, NY, USA, November 2022. Association for Computing Machinery.
10. Navid Ghaffarzadegan, Aritra Majumdar, Ross Williams, and Niyousha Hosseinichimeh. Generative agent-based modeling: an introduction and tutorial. *System Dynamics Review*, 40(1):e1761, January 2024.
11. Ross Williams, Niyousha Hosseinichimeh, Aritra Majumdar, and Navid Ghaffarzadegan. Epidemic modeling with generative agents. *arXiv preprint arXiv:2307.04986*, 2023.
12. C. du Plooy and R. Oosthuizen. AI usefulness in systems modelling and simulation: GPT-4 application. *South African Journal of Industrial Engineering*, 34(3):286–303, November 2023. Publisher: The Southern African Institute for Industrial Engineering.
13. Ali Akhavan and Mohammad S. Jalali. Generative AI and simulation modeling: how should you (not) use large language models like ChatGPT. *System Dynamics Review*, 40(3):e1773, July 2024.
14. Eneko Martin-Martinez, Roger Samsó, James Houghton, and Jordi Solé Ollé. PySD: System Dynamics Modeling in Python. *Articles publicats en revistes (Dinàmica de la Terra i l'Oceà)*, October 2022. Accepted: 2023-02-21T12:17:23Z.
15. BPTK-Py: System Dynamics and Agent Based Modeling in Python - Limitations.
16. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

17. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. arXiv:1810.04805.
18. Md Imbesat Hassan Rizvi, Xiaodan Zhu, and Iryna Gurevych. SpaRC and SpaRP: Spatial Reasoning Characterization and Path Generation for Understanding Spatial Reasoning Capability of Large Language Models, June 2024. arXiv:2406.04566.
19. Yutaro Yamada, Yihan Bao, Andrew K. Lampinen, Jungo Kasai, and Ilker Yildirim. Evaluating Spatial Understanding of Large Language Models, March 2024. arXiv:2310.14540.
20. Anthony G. Cohn. An Evaluation of ChatGPT-4's Qualitative Spatial Reasoning Capabilities in RCC-8, September 2023. arXiv:2309.15577.
21. Yuhan Ji and Song Gao. Evaluating the Effectiveness of Large Language Models in Representing Textual Descriptions of Geometry and Spatial Relations, July 2023. arXiv:2307.03678.
22. Sean Tucker. A systematic review of geospatial location embedding approaches in large language models: A path to spatial AI systems, January 2024. arXiv:2401.10279.
23. Yongyao Jiang and Chaowei Yang. Is ChatGPT a Good Geospatial Data Analyst? Exploring the Integration of Natural Language into Structured Query Language within a Spatial Database. *ISPRS International Journal of Geo-Information*, 13(1):26, January 2024. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.