

Evaluating Intrinsic Geospatial Topological Reasoning in LLMs

Shaolin Xie
shaolinx@usc.edu
University of Southern California
Los Angeles, California, USA

Shang-Ling Hsu
hsushang@usc.edu
University of Southern California
Los Angeles, California, USA

Qihan Zhang
qihanzha@usc.edu
University of Southern California
Los Angeles, California, USA

Yiming Gao
gyming@umd.edu
University of Maryland
College Park, Maryland, USA

Cyrus Shahabi
shahabi@usc.edu
University of Southern California
Los Angeles, California, USA

Ibrahim Sabek
sabek@usc.edu
University of Southern California
Los Angeles, California, USA

Abstract

Large language models (LLMs) are increasingly used in geospatial tools, yet their intrinsic topological reasoning, separate from retrieval or geometric encodings, remains unclear. We introduce an evaluation¹ using context-free, natural-language questions about real-world entities. We test three prompting regimes: zero-shot, prompt optimization with definitions, and chain-of-thought, to assess when language alone suffices and when guided reasoning helps. We find that LLMs possess latent geospatial knowledge, but they suffer inherent limitations in spatial understanding. While various prompting methods can improve performance, they also expose fundamental logical inconsistencies and conceptual flaws, revealing that a model’s spatial reasoning ability is limited by its core reasoning power, not just a lack of access to external data.

ACM Reference Format:

Shaolin Xie, Shang-Ling Hsu, Qihan Zhang, Yiming Gao, Cyrus Shahabi, and Ibrahim Sabek. 2025. Evaluating Intrinsic Geospatial Topological Reasoning in LLMs. In *The 1st ACM SIGSPATIAL International Workshop on Generative and Agentic AI for Multi-Modality Space-Time Intelligence (GeoGenAgent '25)*, November 3–6, 2025, Minneapolis, MN, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3764915.3770722>

1 Introduction

As AI agents are tasked with increasingly complex spatial decision-making, their effectiveness hinges on their capacity for robust **geospatial reasoning**, which is crucial for them to be effective in real-world applications. This capability is multifaceted, including reasoning about distance, direction, and scale. One of the most foundational aspects is **topological reasoning**, which challenges models’ qualitative understanding of how entities connect, overlap, and contain one another, independent of precise coordinates or metrics. As we integrate large language models (LLMs) into real-world applications, from logistics to environmental analysis, a critical question emerges: *can LLM agents use their internal model for topological reasoning, without relying on simple recalls or external tools?*

¹https://github.com/shaolin-x/topology_reasoning_queries

While recent studies have introduced benchmarks for geospatial question answering [7, 10, 11, 14] and geospatial-aware LLMs [15, 17, 25], these often rely on access to external information, such as web search, retrieval-augmented generation (RAG), or structured databases, thereby conflating genuine reasoning with simple information retrieval. Although some efforts have begun to explore topological reasoning in LLMs [6, 13], they focus on tasks that require LLMs to manipulate formal geometric data or translate natural language into formal predicates, or on how models grasp symbolic theories. These works shift focus toward retrieval of pre-trained knowledge or manipulation of geometric encoding, rather than emphasizing true topological reasoning. Lack of real-life entities in their tests also shows a lack of grounding in world entities, which is essential to geospatial tasks.

In this work, we focus on context-free and natural-language questions about real-world entities, while forbidding external tools or coordinates. We evaluate topological reasoning by asking about relations between real-world entities (U.S. states and time zones). These tasks can be easily reasoned by humans when given enough context of boundary interactions between the entities.

Our evaluation is therefore designed to assess the internal topological reasoning engine of a potential LLM agent. By focusing on intrinsic geospatial knowledge, rather than on simpler abilities like geometry parsing or symbol look-ups, we measure a critical capability for any agent that must plan, sanity-check tool outputs, or operate robustly when external information is unavailable.

First, we bridge the gap between formal topology and natural language by mapping DE-9IM-style terms [4] to RCC-8 primitives [5], enabling the formulation of queries that resemble everyday questions. Furthermore, to understand an agent’s baseline intuition versus its ability to follow a deliberate reasoning process, we examine three prompting methods: *zero-shot*, *prompt optimization* [20], and *prompt optimization + chain-of-thought* [21].

Our comparative study reveals that while LLMs with strong general reasoning can perform basic spatial tasks, they suffer from inherent limitations. These limitations usually stem from their reliance on non-spatial linguistic associations and a fundamental lack of geometric consistency, preventing true spatial understanding and coherent spatial reasoning.



2 Methodology

2.1 Spatial Formalisms

Region Connection Calculus (RCC-8) [5, 19] defines eight mutually exclusive topological relations (DC, EC, PO, EQ, TPP, NTPP, TPPi, NTPPi). These qualitative relations, unlike the finer-grained Dimensionally Extended 9-Intersection Model (DE-9IM) [4], allow for single, unambiguous answers, making them a robust foundation for benchmarking the reliability and trustworthiness of an agent’s spatial reasoning process.

To bridge formal logic with the natural language interaction essential for agentic systems, we adopt a vernacular mapping of RCC-8 relations, combining the mutual exclusivity of RCC-8 with the conversational phrasing of DE-9IM. This approach allows for the formulation of queries that resemble everyday questions, yet can be rigorously evaluated against single-label criteria without requiring models to understand formal symbols. The mapping shown in Table A1 visualizes these relations.

We also use the conceptual neighborhood of RCC-8 [13] to capture the similarity and transitions between relations through continuous deformation (Figure 1).

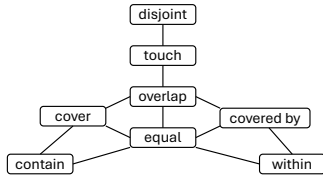


Figure 1: Conceptual neighborhood of topological relations

2.2 Datasets and Domains

To effectively evaluate a model’s intrinsic reasoning, it is crucial to select a domain where the entities are familiar to the model, but the specific relations under investigation are not. U.S. states and time zones meet this requirement perfectly. LLMs are typically pretrained on facts about their interactions and adjacency (e.g., documented in Wikipedia), which ensures that models possess sufficient entity knowledge to support language-only qualitative reasoning, even if the specific topological relations are not explicitly pretrained. For this reason, we focus on these well-structured and widely documented geographies. We benchmark 160 queries that cover all 8 types of RCC-8 interactions between time zones and states (20 randomly sampled queries per relation). We focus on well-structured U.S. geographies: states (polygons), and time zones (polygons), in such a way that the task relies on models’ internal knowledge of widely documented entities rather than obscure datasets. The queried regions are visualized in Appendix Figure A1, reflecting various topological relations. We summarize the formulation of our test queries in Table 1, where the last row shows the complete query template used in evaluation.

2.3 Prompting Methods

We study three prompt settings to understand how LLMs perform topological reasoning by observing how prompting affects performance: **zero-shot**, **PO** (prompt optimization with relations definitions), and **PO+COT** (PO plus chain-of-thought). The prompt optimization strategy [20] is employed to insert explicit operational definitions of key terms directly into the prompt. This approach leverages the model’s internal knowledge while reducing ambiguity and ensuring consistent interpretation. Chain-of-thought reasoning [21] is further applied to the PO prompts to guide LLMs in solving queries step by step.

Prompts used in the experiments are summarized in Figure 2. The model is expected to output a single relation and an explanation of how it reaches the answer [9, 18].

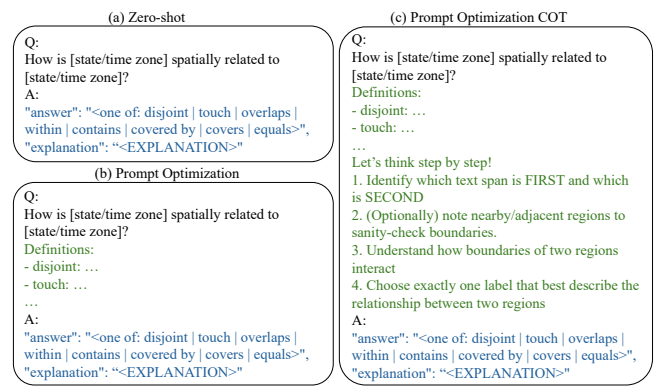


Figure 2: Topological query templates using different prompt techniques.

2.4 Models

We evaluate eight LLM variants from five model families: GPT, Gemini, DeepSeek, Claude, and Llama. This selection was based on their widespread use in prior research [10, 14] and their diverse performance profiles. The GPT series is widely used and has excellent reasoning abilities [3, 23]. The Gemini models have been explicitly fine-tuned for spatial reasoning [12]. We included DeepSeek models to compare general-purpose conversational LLMs with those optimized for logical reasoning [8]. Claude was selected for its low hallucination rate and structured output [1], while Llama models represent widely-used open-source systems for testing reasoning ability [2, 16]. Notably, the specialized “Reasoning” models share a base pre-training corpus but receive an extra stage of supervised chain-of-thought [22] and inference-focused reinforcement fine-tuning [24].

2.5 Evaluation Protocol

We use two complementary metrics to evaluate model performance: **accuracy score** and **difference score**. Together, these provide both binary correctness and graded insights into how close a model’s answer is to the ground truth.

Accuracy Score. This metric reflects the percentage of model responses that match the expected answers. Accuracy is defined

| Formulation | Examples | Output Type |
|-------------------------------------|--|-------------|
| P1 | Mountain Time Zone ; Missouri state ; New York state | P |
| the combined area of P1 and P2 | the combined area of Central and Mountain Time Zones | P |
| P1's area that falls into P2 | the state area that falls into the combined area of Mountain and Central Time Zones | P |
| How is P1 spatially related to P2 ? | How is Missouri state spatially related to its state area that falls into the combined area of Mountain and Central Time Zones ? | Query |

Table 1: Formulation of query templates. P denotes a polygon or a multi-polygon.

strictly based on the spatial relation between two polygons defined in Appendix Table A1. Each query has a single correct relation (e.g., “contains”, “disjoint”) that is derived from polygonal overlays.

Difference Score. The difference score is the shortest path on the conceptual neighborhood shown in Figure 1. The maximum distance between two topological relations is 4; smaller values indicate a more accurate topology understanding.

Our designed tasks enable us to analyze not just whether a model is correct, but how close it is to being correct, providing deeper insight into reasoning and systematic error patterns.

3 Experiments

The comparative study helps us uncover the following questions: 1) How accurately do LLMs resolve everyday spatial terms into correct topological relations without being given explicit definitions? 2) How accurately do LLMs perform the topological reasoning when given definitions of the same labels? 3) Does an LLM perform better with a chain-of-thought prompting when combined with definitions of the relations?

LLMs can use general reasoning to help geospatial reasoning, but with limitations. Table 2 shows that models with strong general reasoning capabilities, such as DeepSeek-Reasoner and Gemini-2.5-Pro, are also superior at geospatial reasoning. However, these models show a nuanced improvement with prompt optimization (PO) and COT. PO with definitions does not consistently improve performance, and sometimes yields only insignificant gains or even dropped performance. This suggests that simply providing definitions is insufficient to overcome a model’s ingrained non-spatial understanding. Conversely, augmenting with COT generally improves performance from a PO baseline (with the notable exception of Gemini-2.5-Pro), indicating that models can leverage explicit reasoning steps to apply definitions more effectively, even if their core understanding of the concepts is flawed.

LLMs struggle to acquire spatial understanding because they rely on non-spatial linguistic reasoning. As seen in Figure 3a, models consistently fail on “covered by” and “covers” relations in a zero-shot setting, often confusing them with “within” or “contains” (as observed in model explanations). This suggests that LLMs’ initial understanding of these terms is not spatially grounded. For example, in everyday language, “contains” and “covers” are often used interchangeably, leading LLMs’ pretraining data to bias their interpretations toward non-spatial semantics.

When explicit spatial definitions are introduced (Figure 3b), performance on “covered by” and “covers” generally improves, but

accuracy on “covers” drops in some models, (e.g., O3-Mini declines from 75% to 35%). This shows that explicit definitions help most models reason better about certain spatial terms, but some models still become more confused about others, likely because their internal language-based associations conflict with the formal spatial meanings being provided.

LLMs’ spatial representations lack geometric consistency. As shown in Figure 3a, 3b, and 3c, models generally perform well on “disjoint”, “within”, and “overlaps”, most achieving more than 80% accuracy in zero-shot setting. However, Figures 3d, 3e, and 3f show that the average difference scores for these relations tend to be large (difference scores > 1), indicating that when they are wrong, they are very wrong. Our analysis further reveals systematic confusion patterns: models mislabel “disjoint” as “overlaps” about 80% of the time, suggesting difficulty distinguishing strict separation from boundary contact. For “within” relations, they most often confuse it with disjoint (22.9%). Likewise, “overlaps” is frequently mistaken for “within” (34.4%). These patterns highlight that models’ internal spatial representations lack geometric consistency, producing severe misclassifications even between topologically distant relations.

Failure modes. From observing models’ explanations to reaching answers, we summarize the following failure modes:

- Incoherent reasoning within the same query: Some models showed inconsistent logic within a single query. For example, models would correctly identify that a state’s boundary touches a time zone’s boundary in their explanation, yet still claim that one is “within” the other instead of “covered by”.
- Incoherent reasoning across different queries: Models struggled to apply consistent logic across different queries, even when the queries involve the same spatial relations.
- Misguided by non-spatial meaning of relations: Some models were misled by the everyday meanings of words such as confusing “within” with “covers”.
- Failure to understand compositional regions: Models struggled when a query involved a combined area of two or more time zones. They often failed to treat the combined area as a single entity and instead reasoned about the individual time zones, leading to incorrect answers.

4 Conclusion and Future Work

In our evaluations, we show that LLMs lack true and grounded understanding of spatial relations, often relying on linguistic patterns that fail on nuanced tasks. This is evident in their inconsistent

| Model | Prompt | Accuracy (%) | Avg. Difference |
|-------------------|-----------|--------------|-----------------|
| Claude-Sonnet-4 | Zero-shot | 47.50 | 0.738 |
| | PO | 46.88 | 0.775 |
| | PO+COT | 51.88 | 0.550 |
| DeepSeek-Chat | Zero-shot | 27.50 | 1.363 |
| | PO | 21.88 | 1.425 |
| | PO+COT | 25.63 | 1.431 |
| DeepSeek-Reasoner | Zero-shot | 66.88 | 0.338 |
| | PO | 79.38 | 0.206 |
| | PO+COT | 88.13 | 0.119 |
| Llama-3.3-70B | Zero-shot | 39.38 | 0.919 |
| | PO | 35.00 | 1.031 |
| | PO+COT | 39.38 | 0.931 |
| GPT-4.1 | Zero-shot | 46.25 | 0.613 |
| | PO | 46.88 | 0.569 |
| | PO+COT | 50.63 | 0.519 |
| Gemini-2.5-Flash | Zero-shot | 65.00 | 0.356 |
| | PO | 60.63 | 0.394 |
| | PO+COT | 62.50 | 0.375 |
| Gemini-2.5-Pro | Zero-shot | 72.50 | 0.275 |
| | PO | 85.63 | 0.144 |
| | PO+COT | 81.76 | 0.182 |
| O3-Mini | Zero-shot | 66.25 | 0.338 |
| | PO | 75.63 | 0.250 |
| | PO+COT | 81.88 | 0.188 |

Table 2: Average accuracy and total average difference by model and prompt.

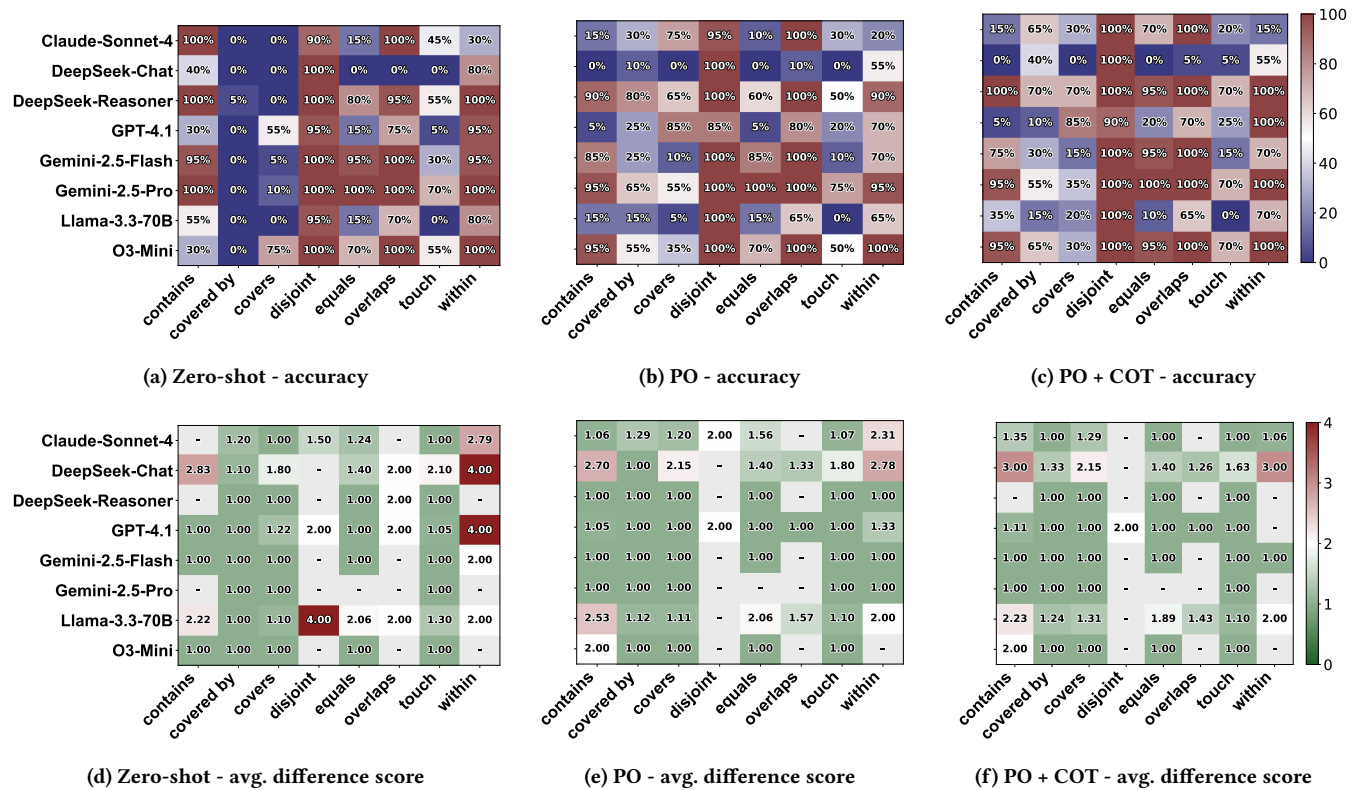


Figure 3: (a)-(c): accuracy percentage by model and relation; (d)-(f): average difference score of incorrect answers.

reasoning and difficulty differentiating between similar topological terms like “within” and “covered by” or “contains” and “covers”, a problem that simple definitions and Chain-of-Thought prompting do not fully solve. This confirms that a model’s ability to reason spatially is fundamentally limited by its general-purpose architecture. For future work, a hybrid approach could be explored, combining LLMs with specialized spatial models like SpaBERT [15] that incorporates geographic signals, or Spatial-RAG [25] which combines

spatial filtering with semantic embeddings for improved spatial reasoning. Such approaches could provide more precise spatial logic needed to overcome the fundamental limitations observed in this experiment.

Acknowledgments

This work is partially funded by Google Data to Insights and Google ML and Systems Junior Faculty research awards.

References

- [1] Anthropic. 2024. *Introducing the Next Generation of Claude*. <https://www.anthropic.com/news/claude-3-family> Highlights reduced hallucination rates and improved structured-output abilities of Claude 3 Sonnet.
- [2] Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. 2023. Are large language models geospatially knowledgeable?. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*. 1–4.
- [3] Ikhyun Cho, Changyeon Park, and Julia Hockenmaier. 2025. The Power of Bullet Lists: A Simple Yet Effective Prompting Approach to Enhancing Spatial Reasoning in Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2025*. 3047–3057.
- [4] Eliseo Clementini, Paolino Di Felice, and Peter van Oosterom. 1993. A small set of formal topological relationships suitable for end-user interaction. In *Advances in Spatial Databases*, David Abel and Beng Chin Ooi (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 277–295.
- [5] Anthony Cohn, Brandon Bennett, John Gooday, and Mark Gotts. 1997. Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. *Geoinformatica* 1 (10 1997). doi:10.1023/A:1009712514511
- [6] Anthony G Cohn and Robert E Blackwell. 2024. Can Large Language Models Reason about the Region Connection Calculus? arXiv:2411.19589 [cs.CL]
- [7] Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. 2024. GEOBench-VLM: Benchmarking Vision-Language Models for Geospatial Tasks. *arXiv preprint arXiv:2411.19325* (2024).
- [8] DeepSeek. 2025. *Models & Pricing Documentation*. https://api-docs.deepseek.com/quick_start/pricing Lists capabilities and context limits of deepseek-chat and deepseek-reasoner.
- [9] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. ERASER: A benchmark to evaluate rationalized NLP models. *arXiv preprint arXiv:1911.03429* (2019).
- [10] Mahir Labib Dihan, Md Tanvir Hassan, Md Tanvir Parvez, Md Hasebul Hasan, Md Almash Alam, Muhammad Aamir Cheema, Mohammed Eunus Ali, and Md Rizwan Parvez. 2024. MapEval: A Map-Based Evaluation of Geo-Spatial Reasoning in Foundation Models. *arXiv preprint arXiv:2501.00316* (2024).
- [11] Andres Garcia-Silva, Cristian Berrio, Jose Manuel Gomez-Perez, Jose Antonio Martinez-Heras, Alessandro Donati, and Ilaria Roma. 2022. Spaceqa: Answering questions about the design of space missions and space craft concepts. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3306–3311.
- [12] Google DeepMind. 2023. Gemini: A Family of Highly Capable Multimodal Models. https://deepmind.google/gemini/gemini_1_report.pdf. Describes fine-tuning and evaluation on spatial-embodied reasoning tasks.
- [13] Yuhan Ji, Song Gao, Ying Nie, Ivan Majić, and Krzysztof Janowicz. 2025. 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, 9 (2025), 1866–1903.
- [14] Zekun Li, Malcolm Grossman, Mihir Kulkarni, Muhao Chen, Yao-Yi Chiang, et al. 2025. MapQA: Open-domain Geospatial Question Answering on Map Data. *arXiv preprint arXiv:2503.07871* (2025).
- [15] Zekun Li, Jina Kim, Yao-Yi Chiang, and Muhao Chen. 2022. SpaBERT: A Pre-trained Language Model from Geographic Data for Geo-Entity Representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 2757–2769.
- [16] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2023. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213* (2023).
- [17] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David B Lobell, and Stefano Ermon. [n. d.]. GeoLLM: Extracting Geospatial Knowledge from Large Language Models. In *The Twelfth International Conference on Learning Representations*.
- [18] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361* (2019).
- [19] David A. Randell, Zhan Cui, and Anthony G. Cohn. 1992. A spatial logic based on regions and connection. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning (Cambridge, MA) (KR'92)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 165–176.
- [20] Zirui Song, Bin Yan, Yuhan Liu, Miao Fang, Mingzhe Li, Rui Yan, and Xiuying Chen. 2025. Injecting Domain-Specific Knowledge into Large Language Models: A Comprehensive Survey. *CoRR abs/2502.10708* (2025). arXiv:2502.10708
- [21] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1800, 14 pages.
- [22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
- [23] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Mind’s Eye of LLMs: Visualization-of-Thought Elicits Spatial Reasoning in Large Language Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- [24] Fengli Xu, Qianye Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. 2025. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. *arXiv preprint arXiv:2501.09686* (2025).
- [25] Dazhou Yu, Riyang Bao, Gengchen Mai, and Liang Zhao. 2025. Spatial-rag: Spatial retrieval augmented generation for real-world spatial reasoning questions. *arXiv preprint arXiv:2502.18470* (2025).

A Appendix

| Vernacular | RCC-8 | Visual | Vernacular | RCC-8 | Visual |
|------------|-------|--------|------------|-------|--------|
| disjoint | DC | | covered by | TPP | |
| touch | EC | | cover | TPPi | |
| overlap | PO | | within | NTPP | |
| equal | EQ | | contain | NTPPi | |

Table A1: Vernacular and corresponding RCC-8 relations.

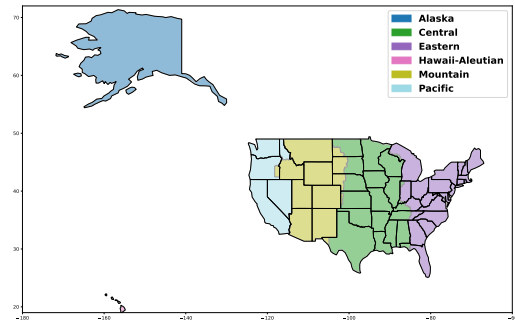


Figure A1: Distribution of areas of evaluated queries, classified as U.S. time zones and U.S. states.