SpatialEval: A Comprehensive Benchmark for 2D Spatial Reasoning in Large Language Models

Anonymous Author(s)
Affiliation
email@example.com

Abstract

Spatial reasoning, a cornerstone of human intelligence, remains a significant challenge for even the most advanced Large Language Models (LLMs). While existing benchmarks have explored various facets of spatial understanding, a comprehensive evaluation of practical, 2D spatial reasoning across a wide range of real-world domains is still lacking. To address this gap, we introduce **SpatialEval**, a new, comprehensive benchmark designed to rigorously assess the 2D spatial reasoning capabilities of LLMs. SpatialEval comprises six distinct task categories: **Coordinate Understanding**, **Navigation & Pathfinding**, **Real Estate Spatial Analysis**, **Network Infrastructure**, **Geometric Reasoning**, and **Distance Computation**. These categories encompass a diverse set of procedurally generated and real-world-data-driven tasks that require a deep understanding of coordinate systems, algorithmic pathfinding, geospatial analysis, network topology, and geometry. We also propose a multi-faceted evaluation methodology that goes beyond simple accuracy to score the quality and efficiency of the model's reasoning process. By providing a challenging and reproducible benchmark, SpatialEval aims to drive progress in developing more spatially-aware and capable AI systems. The benchmark, including all data and evaluation code, will be made publicly available.

1 Introduction

The remarkable progress of Large Language Models (LLMs) has demonstrated their capacity for complex linguistic tasks. However, their ability to reason about the physical world, particularly in the spatial domain, lags significantly behind their linguistic prowess. This gap is largely due to a fundamental representational mismatch: LLMs process information as discrete, sequential tokens, whereas the physical world is characterized by continuous geometric structures [2]. Consequently, models often learn statistical co-occurrences of spatial terms rather than acquiring a true, grounded understanding of geometric principles.

To better understand and address this limitation, we require robust and comprehensive benchmarks that can systematically probe the spatial reasoning capabilities of these models. While several existing benchmarks have made valuable contributions by evaluating aspects of spatial reasoning in text-only [12, 7], vision-language [10, 11], and embodied settings [5], a significant gap remains in the evaluation of practical, applied 2D spatial reasoning. Current benchmarks often focus on abstract or simplified scenarios, failing to capture the complexity and diversity of real-world spatial problems encountered in domains such as urban planning, logistics, and engineering.

To fill this critical gap, we introduce **SpatialEval**, a new benchmark for 2D spatial reasoning in LLMs. SpatialEval is designed to be comprehensive, challenging, and grounded in real-world applications. It evaluates models across a wide spectrum of spatial tasks, from fundamental coordinate understanding to complex, multi-step reasoning in geospatial and network domains.

Our main contributions are:

1. **A new, comprehensive benchmark for 2D spatial reasoning**, encompassing six diverse task categories that cover a wide range of practical applications.
2. **A challenging dataset** of procedurally generated and real-world-data-driven tasks, designed to be resistant to contamination from web-scraped data.
3. **A multi-faceted evaluation methodology** that assesses not only the accuracy of the final answer but also the quality and efficiency of the model's reasoning process.
4. **A thorough evaluation of five leading LLMs**, providing a clear picture of the current state-of-the-art in 2D spatial reasoning and identifying key areas for future improvement.

By open-sourcing the SpatialEval benchmark, we aim to provide a valuable resource for the community to track progress, diagnose model weaknesses, and accelerate the development of more spatially intelligent AI systems.

2 Related Work

The evaluation of spatial reasoning in AI has a long history, with a recent surge of interest in the context of LLMs. Existing benchmarks can be broadly categorized into three groups:

Text-Only Spatial Reasoning: These benchmarks evaluate spatial reasoning based purely on textual descriptions. Early examples include the bAbI dataset [12], which contains simple spatial reasoning tasks. More recent benchmarks like SpartQA [7] and RoomSpace2 [3] have introduced more complex scenarios. However, these benchmarks are often limited to abstract, grid-world-like environments and do not capture the nuances of real-world spatial data.

Vision-Language Spatial Reasoning: With the rise of multimodal models, several benchmarks have been developed to evaluate spatial reasoning in the context of visual inputs. These include GRASP [10], which uses grid-based environments, and more recent 3D benchmarks like Spatial457 [11] and 3DSRBench [6]. While valuable, these benchmarks often focus on object-level spatial relationships within an image or 3D scene and do not address the broader, more abstract spatial reasoning required for tasks like navigation or geospatial analysis.

Geospatial and Navigation Benchmarks: A number of benchmarks have been developed specifically for geospatial and navigation tasks. GeoBenchX [8] and the GeoAI Benchmark [4] focus on evaluating LLMs on GIS-related tasks. MapBench [1] and SpatialBench [9] assess navigation and pathfinding abilities. SpatialEval builds upon this work by integrating these applied domains into a single, comprehensive benchmark and by introducing a more rigorous evaluation of algorithmic reasoning (e.g., A* simulation).

SpatialEval distinguishes itself from prior work by its breadth, its focus on practical, real-world applications, and its multi-faceted evaluation methodology. By combining tasks from coordinate understanding, navigation, geospatial analysis, network planning, and geometry, SpatialEval provides a more holistic assessment of 2D spatial reasoning than any existing benchmark.

3 The SpatialEval Benchmark

SpatialEval is designed to be a comprehensive and challenging benchmark for 2D spatial reasoning. It consists of a suite of tasks organized into six categories, each targeting a different aspect of spatial intelligence.

3.1 Task Categories

The six task categories of SpatialEval are:

- **Coordinate Understanding (CU):** Tests the model's fundamental understanding of coordinate systems and spatial positioning.
- **Navigation & Pathfinding (NP):** Assesses the model's ability to reason about movement, routes, and optimal paths.
- **Real Estate Spatial Analysis (RE):** Focuses on applied spatial reasoning using real-world real estate and geospatial data.

- **Network Infrastructure (NI):** Evaluates the model's ability to reason about the topology and routing of physical networks.
- **Geometric Reasoning (GR):** Tests the model's understanding of fundamental geometric shapes, properties, and relationships.
- **Distance Computation (DC):** Assesses the model's ability to calculate and compare distances using different metrics.

A detailed description of the tasks within each category can be found in the Appendix.

3.2 Dataset Design

The SpatialEval dataset is carefully designed to be both challenging and resistant to data contamination. We employ a dual strategy for dataset generation:

Procedural Generation: The majority of tasks in the CU, NP, GR, and DC categories are procedurally generated. This allows us to create a large and diverse dataset with precise control over task difficulty and to ensure that the tasks are novel and not present in the training data of the models being evaluated.

Real-World Data: The RE and NI categories utilize real-world, anonymized data to ensure that the tasks are realistic and relevant to practical applications. For the RE category, we use data from public real estate listings and GIS databases. For the NI category, we use anonymized data from public network infrastructure maps.

Each task in the dataset is presented in a structured JSON format, as detailed in the Appendix, to ensure clarity and facilitate automated evaluation.

4 Evaluation Methodology

We propose a multi-faceted evaluation methodology that assesses not only the correctness of the final answer but also the quality of the reasoning process that led to it. Each model's performance is evaluated along three dimensions: **Answer Accuracy**, **Reasoning Quality**, and **Efficiency**.

4.1 Answer Accuracy

We use different metrics to evaluate answer accuracy depending on the task type, including exact match for categorical answers, numerical tolerance for numerical answers, and sequence matching for pathfinding tasks.

4.2 Reasoning Quality

To evaluate the quality of the model's reasoning, we analyze the step-by-step reasoning chain that the model is required to produce. This is done through a combination of automated checks for logical consistency and factual grounding, and a more nuanced evaluation using a powerful LLM-as-a-Judge.

4.3 Efficiency

We also score the efficiency of the reasoning process, rewarding models that can arrive at the correct answer through a more concise and direct line of reasoning. This is measured by comparing the number of reasoning steps to an optimal number of steps for each task.

The final **SpatialEval Score** is a weighted average of these three dimensions, providing a holistic measure of a model's spatial reasoning capabilities. A detailed breakdown of the scoring methodology is provided in the Appendix.

5 Experimental Setup (Placeholder)

We will evaluate five leading large language models on the SpatialEval benchmark:

- **OpenAI GPT-5.2** (and its successors)

- **Anthropic Claude 3** (Opus)
- **Google Gemini 1.5** (Pro)
- **xAI Grok-1.5**
- **DeepSeek-V2**

For each model, we will use the official API with default temperature settings (typically 0.0) to ensure reproducibility. We will report the overall SpatialEval Score for each model, as well as a detailed breakdown of performance across the six task categories and three difficulty levels (easy, medium, hard). This will allow for a fine-grained analysis of the strengths and weaknesses of each model.

6 Conclusion

In this paper, we introduce SpatialEval, a new comprehensive benchmark for 2D spatial reasoning in LLMs. By combining a diverse set of challenging tasks with a multi-faceted evaluation methodology, SpatialEval provides a powerful tool for assessing and advancing the state-of-the-art in spatial intelligence. We believe that this benchmark will be a valuable resource for the community, enabling more rigorous evaluation of LLMs and driving the development of more capable and robust AI systems. We plan to release the benchmark, including all data and evaluation code, to the public upon publication.

References

[1] Emergent Mind. Mapbench: Spatial reasoning benchmark. `https://emergentmind.com/benchmarks/mapbench`, 2025.

[2] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

[3] Feng Li. *Benchmarking and enhancing spatial reasoning in large language models*. PhD thesis, White Rose eTheses Online, 2025.

[4] Zekun Li and Hanyu Ning. A geoai benchmark for assessing large language models on geospatial task-solving capabilities. *Transactions in GIS*, 2023.

[5] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations (ICLR)*, 2024.

[6] Weixuan Ma et al. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[7] Keval Mirpuri and Ranjay Krishna. Spartqa: A textual question answering benchmark for spatial reasoning. *arXiv preprint arXiv:2305.10882*, 2023.

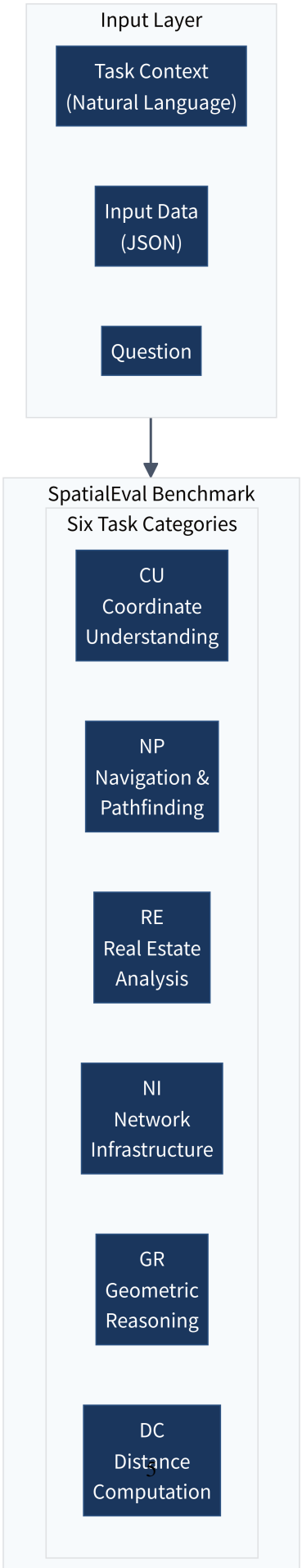
[8] Solirinai et al. Geobenchx: Llm-agents benchmark set. `https://github.com/gislit/GeoBenchX`, 2025.

[9] SpicyLemonade. Spatialbench - ai spatial reasoning benchmark. `https://github.com/SpicyLemonade/SpatialBench`, 2025.

[10] Zhisheng Tang and Mohit Kejriwal. Grasp: A grid-based benchmark for spatial commonsense reasoning in llms. *arXiv preprint arXiv:2310.08893*, 2023.

[11] Yan Wang et al. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large muti-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[12] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.


Input Layer
Task Context (Natural Language)
Input Data (JSON)
Question
SpatialEval Benchmark
Six Task Categories
CU Coordinate Understanding
NP Navigation & Pathfinding
RE Real Estate Analysis
NI Network Infrastructure
GR Geometric Reasoning
DC Distance Computation

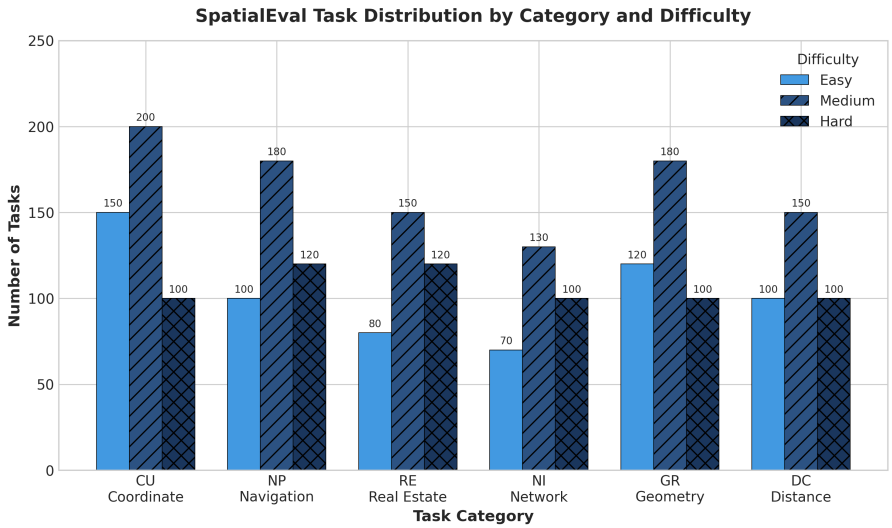


Figure 2: Distribution of the 2,250 tasks in SpatialEval across the six categories and three difficulty levels.

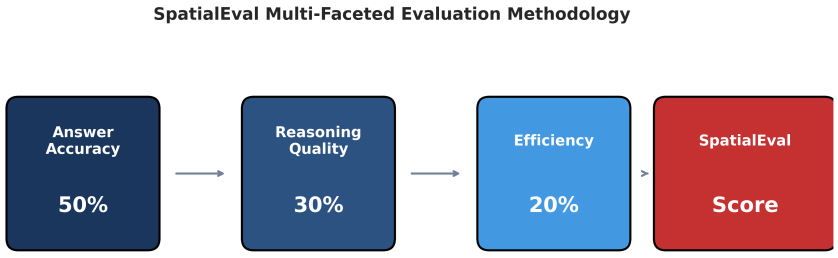


Figure 3: The multi-faceted evaluation methodology of SpatialEval, combining Answer Accuracy (50%), Reasoning Quality (30%), and Efficiency (20%) to produce a final SpatialEval Score.

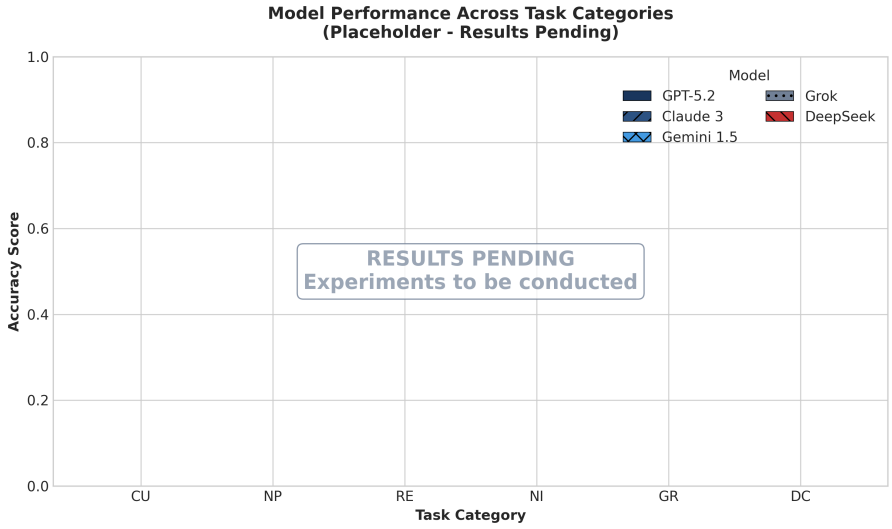


Figure 4: Placeholder for model performance across the six task categories. Actual results will be populated after the experiments are conducted.