
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 [3]. 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 [13, 8], vision-language [11, 12], and embodied settings [6], 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 [13], which contains simple spatial reasoning tasks. More recent benchmarks like SpartQA [8] and RoomSpace2 [4] 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 [11], which uses grid-based environments, and more recent 3D benchmarks like Spatial457 [12] and 3DSRBench [7]. 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 [9] and the GeoAI Benchmark [5] focus on evaluating LLMs on GIS-related tasks. MapBench [1] and SpatialBench [10] 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).

Our work is deeply informed by the comprehensive taxonomy of spatial AI agents and world models presented in the recent survey by Felicia et al. [2]. That work provides a unified framework for understanding the capabilities of spatial AI agents, and we adopt their three-axis taxonomy (Spatial Task, Agentic Capability, Spatial Scale) as a foundational guide for the design of SpatialEval. While their survey provides the theoretical framework, SpatialEval provides the practical, large-scale benchmark to measure and drive progress within that framework.

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 Design Principles

We designed SpatialEval with four core principles:

- **Real-World Grounding:** Tasks are derived from documented, high-value industry use cases to ensure practical relevance and applicability.
- **Comprehensive Coverage:** The benchmark spans six distinct categories of spatial reasoning, from fundamental geometry to complex, multi-step geospatial analysis.
- **Controlled Difficulty:** A mix of procedural generation and real-world data allows for precise control over task difficulty, enabling fine-grained analysis of model capabilities.
- **Multi-Faceted Evaluation:** We move beyond simple accuracy to evaluate the quality and efficiency of the reasoning process, providing a more holistic view of a model’s spatial intelligence.

3.2 Benchmark Task Taxonomy

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.

3.3 Dataset Composition

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 Metrics

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 Experiments

5.1 Baseline Models (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.

5.2 Main Results (Placeholder)

Results will be populated after conducting experiments with the five baseline models. We will present a comprehensive analysis including overall SpatialEval Scores, per-category breakdowns, and performance across difficulty levels.

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] Gloria Felicia, Nolan Bryant, Handi Putra, Ayaan Gazali, Eliel Lobo, and Esteban Rojas. From perception to action: Spatial ai agents and world models. *arXiv preprint arXiv:2602.01644*, 2026.

- [3] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.
- [4] Feng Li. *Benchmarking and enhancing spatial reasoning in large language models*. PhD thesis, White Rose eTheses Online, 2025.
- [5] Zekun Li and Hanyu Ning. A geoai benchmark for assessing large language models on geospatial task-solving capabilities. *Transactions in GIS*, 2023.
- [6] 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.
- [7] Weixuan Ma et al. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.
- [8] Keval Mirpuri and Ranjay Krishna. Spartqa: A textual question answering benchmark for spatial reasoning. *arXiv preprint arXiv:2305.10882*, 2023.
- [9] Solirina et al. Geobenchx: Llm-agents benchmark set. <https://github.com/gislit/GeoBenchX>, 2025.
- [10] SpicyLemonade. Spatialbench - ai spatial reasoning benchmark. <https://github.com/SpicyLemonade/SpatialBench>, 2025.
- [11] Zhisheng Tang and Mohit Kejriwal. Grasp: A grid-based benchmark for spatial commonsense reasoning in llms. *arXiv preprint arXiv:2310.08893*, 2023.
- [12] Yan Wang et al. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [13] 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.

A Appendix A: Real-World Grounding in AtlasPro Use Cases

To ensure SpatialEval is grounded in practical, high-value applications, we derived our task categories from an analysis of 60 validated industry use cases from AtlasPro AI, a leading platform for spatial and network intelligence. These use cases span five major industries: Telecom/Fiber, Utilities, Government/Smart Cities, Retail, and Construction/Industrial. The table below maps these real-world use cases to the six core spatial reasoning capabilities evaluated in our benchmark, demonstrating the direct link between SpatialEval and the demands of real-world AI agent applications.

B Appendix B: Dataset Statistics

The SpatialEval benchmark comprises 2,250 tasks distributed evenly across six categories and three difficulty levels. Table 2 provides a detailed breakdown of the dataset composition.

C Appendix C: Task Type Details

Each category contains multiple task types designed to probe different aspects of spatial reasoning:

Coordinate Understanding (CU):

- Quadrant identification
- GPS coordinate transformation
- Polygon containment testing

Navigation & Pathfinding (NP):

- Direction following
- Manhattan distance shortest path
- A* pathfinding with obstacles

Real Estate Analysis (RE):

- Property area calculation
- Proximity analysis to amenities
- Zoning compliance verification

Network Infrastructure (NI):

- Cable length computation
- Network topology analysis
- Failure cascade impact assessment

Geometric Reasoning (GR):

- Basic shape area calculation
- Spatial relationship determination
- Polygon area via shoelace formula

Distance Computation (DC):

- Euclidean distance calculation
- Manhattan distance comparison
- Geodesic distance via Haversine formula

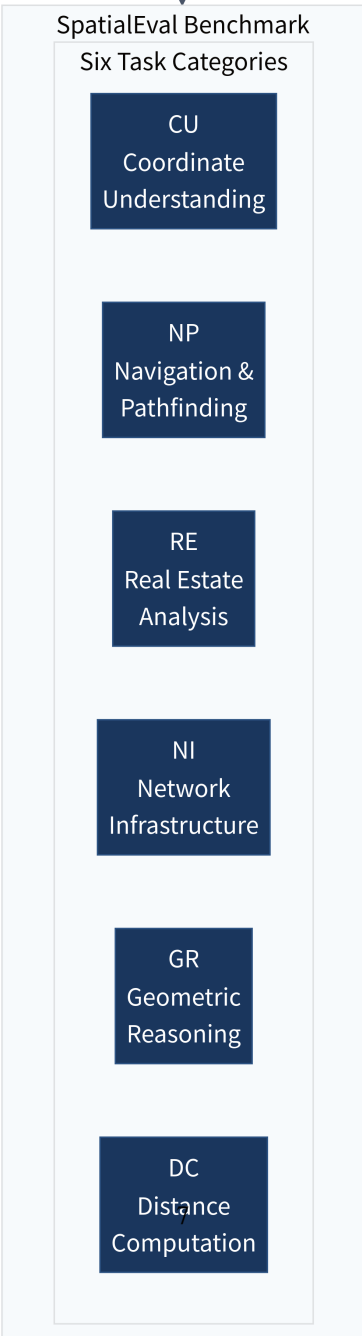
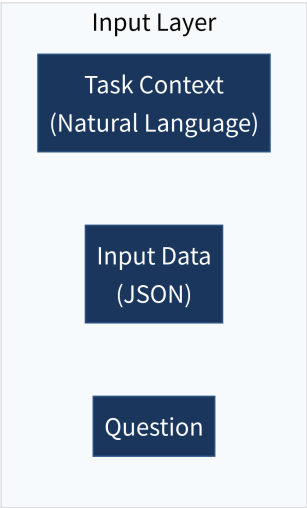
D Appendix D: Evaluation Score Calculation

The final SpatialEval Score is computed as a weighted average:

$$\text{SpatialEval Score} = 0.50 \times \text{Accuracy} + 0.30 \times \text{Reasoning Quality} + 0.20 \times \text{Efficiency} \quad (1)$$

where:

- **Accuracy** is the percentage of tasks with correct final answers
- **Reasoning Quality** is scored on a 0-100 scale by an LLM-as-Judge evaluator
- **Efficiency** is computed as $\min(1, \frac{\text{optimal steps}}{\text{actual steps}}) \times 100$



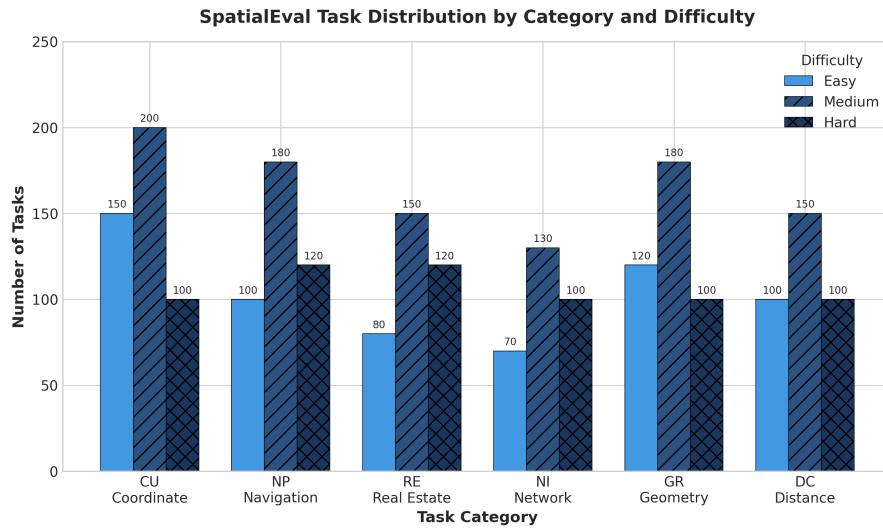


Figure 2: Distribution of the 2,250 tasks in SpatialEval across the six categories and three difficulty levels.

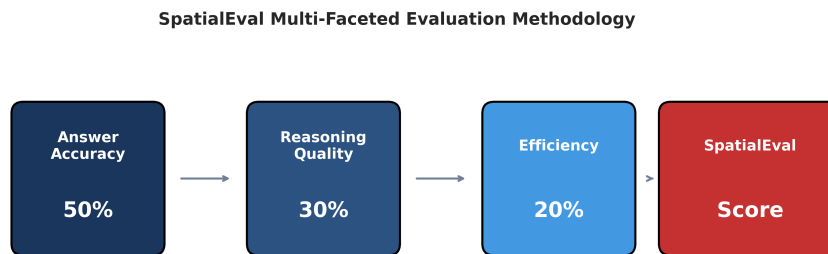


Figure 3: The multi-faceted evaluation methodology of SpatialEval, combining Answer Accuracy (50%), Reasoning Quality (30%), and Efficiency (20%) to produce a final SpatialEval Score.

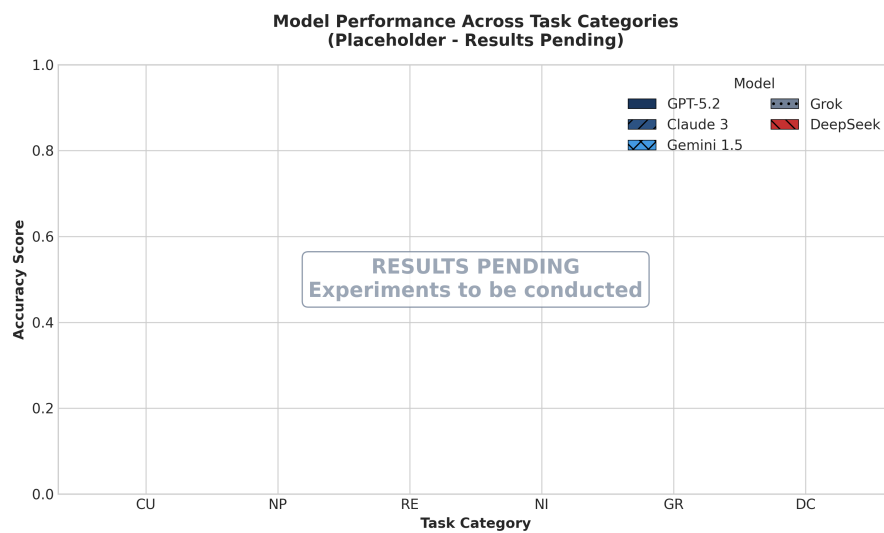


Figure 4: Placeholder for model performance across the six task categories. Actual results will be populated after the experiments are conducted.

Table 1: Mapping of AtlasPro AI Use Cases to SpatialEval Task Categories

AtlasPro AI Use Case Query	CU	NP	RE	NI	GR	DC
Telecom / Fiber						
Breakdown of premises in a polygon	✓		✓		✓	
Check for existing fiber in a neighborhood			✓	✓		✓
Cost to build fiber to a polygon	✓		✓	✓		✓
Find unserved locations in a county	✓		✓			✓
Identify usable utility poles for aerial fiber	✓			✓		✓
Generate a splice point plan for a route		✓		✓		
Identify single point of failure exposure				✓		
Find optimal route from CO to subdivision		✓		✓		✓
Utilities — Electric/Gas/Water						
Assess wildfire risk for transmission lines				✓		✓
Identify customers affected by de-energizing a circuit				✓		
Analyze cascade risk from substation failure				✓		
Detect vegetation encroachment in a corridor	✓				✓	✓
Predict outage probability for the next week				✓		
Optimize crew dispatch for an outage		✓				✓
Government / Smart Cities						
Analyze traffic impact of a road closure		✓				
Optimize signal timing for a corridor		✓				
Find best route for emergency response		✓				✓
Assess impact of a water main break				✓		
Identify crime hotspots in a district	✓					
Determine optimal patrol routes		✓				✓
Retail						
Identify camera blind spots in a store	✓				✓	

Table 2: SpatialEval Dataset Statistics

Category	Code	Easy	Medium	Hard	Total
Coordinate Understanding	CU	125	125	125	375
Navigation & Pathfinding	NP	125	125	125	375
Real Estate Analysis	RE	125	125	125	375
Network Infrastructure	NI	125	125	125	375
Geometric Reasoning	GR	125	125	125	375
Distance Computation	DC	125	125	125	375
Total		750	750	750	2,250