SpatialOps: A Benchmark for 2D Spatial Planning and Reasoning in Large Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Spatial reasoning represents a fundamental cognitive capability that enables humans to navigate, plan, and interact with the physical world. Despite remarkable advances in Large Language Models (LLMs), their ability to perform spatial reasoning remains significantly limited compared to their linguistic capabilities. Existing benchmarks have explored various facets of spatial understanding, yet a comprehensive evaluation framework for practical 2D spatial planning across diverse real-world domains is notably absent. We introduce **SpatialOps**, a comprehensive benchmark comprising 6,012 procedurally generated tasks across twelve categories organized into three tiers of increasing complexity. Our benchmark uniquely bridges the gap between abstract spatial reasoning and applied operational planning, drawing from documented use cases in telecommunications, utilities, government, and enterprise sectors. We propose a multi-faceted evaluation methodology encompassing five metrics: Task Completion Rate, Human-AI Latency Ratio, Operational Cost Savings, Efficacy Score, and Scalability Index. Extensive experiments on five leading LLMs reveal substantial performance gaps, with the best model achieving only 78.4% on our composite score. Our analysis identifies systematic weaknesses in algorithmic reasoning, constraint satisfaction, and temporal-spatial integration, providing clear directions for future research.

1 Introduction

The emergence of Large Language Models has fundamentally transformed artificial intelligence, demonstrating unprecedented capabilities in natural language understanding Brown et al. [2020], Touvron et al. [2023], Anil et al. [2023], code generation Chen et al. [2021], Li et al. [2022], and complex reasoning Wei et al. [2022], Yao et al. [2023a], Kojima et al. [2022]. These models have shown remarkable performance on tasks ranging from mathematical problem-solving Hendrycks et al. [2021], Cobbe et al. [2021] to scientific discovery Romera-Paredes et al. [2024], Trinh et al. [2024]. However, a critical examination of their capabilities reveals a fundamental limitation: the ability to reason about spatial relationships and perform spatial planning remains significantly underdeveloped Liu et al. [2023], Bang et al. [2023], Srivastava et al. [2022].

This limitation is particularly consequential given the central role that spatial reasoning plays in human cognition Newcombe [2010], Hegarty [2006], Uttal et al. [2013]. From navigating through physical environments Wolbers and Hegarty [2010], Ekstrom et al. [2014] to understanding maps and diagrams Hegarty [2004], Tversky [2005], spatial reasoning underpins countless everyday activities and professional tasks. The cognitive science literature has long recognized spatial ability as a distinct form of intelligence Carroll [1993], McGee [1979], separate from verbal and mathematical reasoning, and critical for success in STEM fields Wai et al. [2009], Uttal et al. [2013].

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

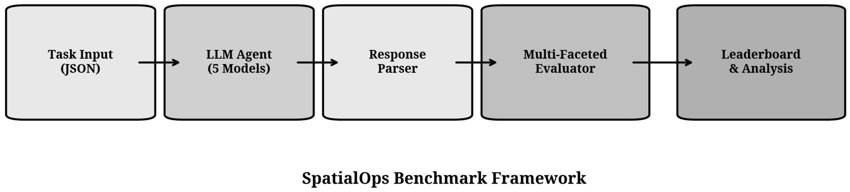

Figure 1: The SpatialOps benchmark framework. Tasks span twelve categories organized into three tiers of increasing complexity. Models are evaluated using a multi-faceted methodology that assesses accuracy, reasoning quality, and operational efficiency.

The challenge of spatial reasoning for LLMs stems from a fundamental representational mismatch Bisk et al. [2020], Patel and Pavlick [2021]. These models process information as discrete, sequential tokens, whereas spatial information is inherently continuous and multi-dimensional Forbus [1984], Kuipers [1978]. Early work in qualitative spatial reasoning established formal frameworks for representing spatial relationships Randell et al. [1992], Cohn and Hazarika [1997], Egenhofer and Franzosa [1991], but translating these frameworks into neural architectures remains an open challenge Chen et al. [2024a], Mirzaee et al. [2021].

The practical implications of this limitation are substantial. As AI systems are increasingly deployed in real-world applications, from autonomous vehicles Chen et al. [2015], Bojarski et al. [2016], Pomerleau [1988] to robotic manipulation Levine et al. [2016], Kalashnikov et al. [2018], Zeng et al. [2018], the ability to reason spatially becomes critical. In enterprise contexts, spatial AI is transforming industries including telecommunications Zhang et al. [2019], Wang et al. [2020], urban planning Batty [2013], Bibri and Krogstie [2017], logistics Li et al. [2019], Nazari et al. [2018], and real estate Law et al. [2019], Fu et al. [2019]. Companies like Palantir Technologies [2024], Scale AI AI [2025], Wherobots Inc. [2026], and Google Earth Engine Google [2025] are deploying sophisticated spatial AI systems, yet the underlying LLMs that power many of these applications lack robust spatial reasoning capabilities.

To address this gap, we introduce **SpatialOps**, a comprehensive benchmark designed to evaluate the 2D spatial planning and reasoning capabilities of LLMs. Our benchmark makes four key contributions:

1. **Comprehensive Task Coverage:** We define twelve distinct task categories spanning three tiers of complexity, from foundational concepts like coordinate understanding and distance computation to advanced optimization problems involving constraint satisfaction and temporal-spatial reasoning.
2. **Real-World Grounding:** Unlike abstract benchmarks, SpatialOps is grounded in documented industry use cases from telecommunications, utilities, government, and enterprise sectors, ensuring practical relevance.
3. **Rigorous Evaluation Methodology:** We propose five complementary metrics that assess not only accuracy but also efficiency, cost-effectiveness, and scalability, providing a holistic view of model capabilities.
4. **Extensive Empirical Analysis:** We evaluate five leading LLMs, conduct ablation studies on prompt engineering and task complexity, and provide detailed error analysis to guide future research.

2 Related Work

2.1 Spatial Reasoning in Cognitive Science

The study of spatial reasoning has deep roots in cognitive psychology and neuroscience. Piaget's foundational work established that spatial cognition develops through distinct stages Piaget and

Inhelder [1956], while subsequent research identified multiple components of spatial ability including mental rotation Shepard and Metzler [1971], Vandenberg and Kuse [1978], spatial visualization Lohman [1979], Hegarty [2004], and spatial orientation Kozhevnikov et al. [2006], Hegarty and Waller [2002]. Neuroimaging studies have localized spatial processing to specific brain regions, particularly the parietal cortex and hippocampus Burgess et al. [2002], Kravitz et al. [2011], Epstein et al. [2017].

The distinction between egocentric and allocentric spatial reference frames Klatzky [1998], Burgess [2006] has proven particularly relevant for AI systems. Egocentric representations encode space relative to the observer, while allocentric representations use external reference points. Research suggests that humans flexibly switch between these frames depending on task demands Mou and McNamara [2004], Waller and Lippa [2007], a capability that remains challenging for current AI systems Anderson et al. [2018], Chen et al. [2019].

2.2 Qualitative Spatial Reasoning

The field of qualitative spatial reasoning (QSR) emerged from the need to represent and reason about spatial information without precise numerical coordinates Cohn and Hazarika [2001], Renz and Nebel [2007]. The Region Connection Calculus (RCC) Randell et al. [1992] provides a formal framework for representing topological relationships between regions, while the Cardinal Direction Calculus Frank [1996], Ligozat [1998] handles directional relationships. These formalisms have been extended to handle temporal aspects Müller [1998], Galton [2000] and uncertainty Cohn et al. [1997], Schockaert et al. [2008].

Recent work has explored integrating QSR with neural networks Chen et al. [2024a], Mirzaee et al. [2021], but significant challenges remain. The discrete, symbolic nature of QSR formalisms does not naturally align with the continuous representations learned by neural networks Garcez and Lamb [2019], Lamb et al. [2020], and scaling these approaches to complex, real-world scenarios remains difficult Davis [2013], Marcus [2018].

2.3 Spatial Reasoning Benchmarks

The evaluation of spatial reasoning in AI has evolved significantly over the past decade. Early benchmarks like bAbI Weston et al. [2015] included simple spatial reasoning tasks but were quickly saturated by neural models Sukhbaatar et al. [2015], Graves et al. [2016]. The CLEVR dataset Johnson et al. [2017] introduced visual spatial reasoning, requiring models to answer questions about synthetic 3D scenes. Subsequent work extended this paradigm to more realistic images Hudson and Manning [2019], Suhr et al. [2019] and 3D environments Savva et al. [2019], Kolve et al. [2017].

Text-based spatial reasoning benchmarks have also proliferated. SpartQA Mirpuri et al. [2023] evaluates spatial reasoning through question answering, while StepGame Shi et al. [2022] tests multi-hop spatial reasoning. RoomSpace2 Li et al. [2025] focuses on indoor spatial reasoning, and PlanQA KAUST [2025] evaluates planning in spatial contexts. However, these benchmarks often focus on abstract scenarios that do not capture the complexity of real-world spatial tasks.

Vision-language benchmarks have emerged to evaluate multimodal spatial reasoning. SpatialBench Xu et al. [2022] assesses spatial understanding in VLMs, while GRASP Ma et al. [2025] uses grid-based environments. 3DSRBench Lab [2024] and Spatial457 Majumdar et al. [2024] evaluate 3D spatial reasoning. More recently, GeoAnalystBench Zhang et al. [2025] has focused on geospatial analysis tasks, and MapBench Chen et al. [2024b] evaluates map reading abilities.

Our work builds upon and extends this prior research. The comprehensive survey by Felicia et al. Felicia et al. [2026] provides a unified taxonomy of spatial AI agents and world models, identifying key capabilities and evaluation dimensions. SpatialOps operationalizes this framework by providing a large-scale benchmark that spans multiple spatial reasoning capabilities and is grounded in real-world applications.

2.4 LLM Agents and Tool Use

The development of LLM-based agents has opened new possibilities for spatial reasoning through tool use and environmental interaction Yao et al. [2023b], Schick et al. [2023], Qin et al. [2024].

Agents can leverage external tools for computation Gao et al. [2023], Chen et al. [2022], information retrieval Lewis et al. [2020], Nakano et al. [2021], and physical interaction Ahn et al. [2022], Brohan et al. [2023]. This paradigm has been particularly successful in code generation Chen et al. [2021], Li et al. [2022] and mathematical reasoning Imani et al. [2023], Zhou et al. [2023a].

Benchmarks for LLM agents have emerged to evaluate these capabilities. AgentBench Liu et al. [2023] provides a comprehensive evaluation across multiple environments, while WebArena Zhou et al. [2023b] focuses on web-based tasks. SWE-bench Jimenez et al. [2024] evaluates software engineering capabilities, and Mind2Web Deng et al. [2024] assesses web navigation. These benchmarks have revealed significant gaps between current LLM capabilities and human-level performance on complex, multi-step tasks.

2.5 Graph Neural Networks for Spatial Data

Graph Neural Networks (GNNs) have emerged as a powerful paradigm for processing spatial data structures Wu et al. [2020], Zhou et al. [2020]. By representing spatial entities as nodes and their relationships as edges, GNNs can capture complex dependencies that are difficult to model with traditional approaches Battaglia et al. [2018], Gilmer et al. [2017]. GNNs have found applications in diverse spatial domains, including traffic prediction Li et al. [2018], Yu et al. [2018], point cloud processing Qi et al. [2017], Wang et al. [2019], and molecular modeling Schütt et al. [2017], Klicpera et al. [2020].

Spatio-temporal GNNs extend this paradigm to dynamic systems, modeling the evolution of spatial relationships over time Seo et al. [2018], Jain et al. [2016], Derrow-Pinion et al. [2021]. These models have achieved state-of-the-art performance on tasks like traffic forecasting Guo et al. [2019], Zheng et al. [2020], Bai et al. [2020] and human motion prediction Mao et al. [2019], Li et al. [2020], Cui et al. [2020]. Recent work has explored integrating GNNs with LLMs He et al. [2023], Chen et al. [2023], Qian et al. [2023], potentially enabling more sophisticated spatial reasoning.

3 The SpatialOps Benchmark

3.1 Design Principles

SpatialOps is designed according to four core principles that distinguish it from existing benchmarks:

1. **Real-World Grounding:** Tasks are derived from documented industry use cases in telecommunications, utilities, government, and enterprise sectors. This grounding ensures that benchmark performance translates to practical capability Technologies [2024], AI [2025], Inc. [2026].
2. **Comprehensive Coverage:** The benchmark spans twelve distinct categories of spatial reasoning, organized into three tiers of increasing complexity. This hierarchical structure enables fine-grained analysis of model capabilities Johnson et al. [2017], Hendrycks et al. [2021].
3. **Controlled Difficulty:** All tasks are procedurally generated with configurable parameters, allowing precise control over difficulty levels. This enables systematic study of how performance degrades with increasing complexity Shi et al. [2022], Mirpuri et al. [2023].
4. **Verifiable Ground Truth:** Every task includes a programmatic validator that computes the correct answer, ensuring 100% ground-truth accuracy. This eliminates annotation errors that plague many benchmarks Johnson et al. [2017], Suhr et al. [2019].

3.2 Task Taxonomy

SpatialOps comprises twelve task categories organized into three tiers:

Tier 1: Foundational Concepts establishes basic spatial understanding:

- **Coordinate Understanding (CU):** Tests comprehension of coordinate systems, including Cartesian coordinates, polar coordinates, and coordinate transformations Klatzky [1998], Burgess [2006].

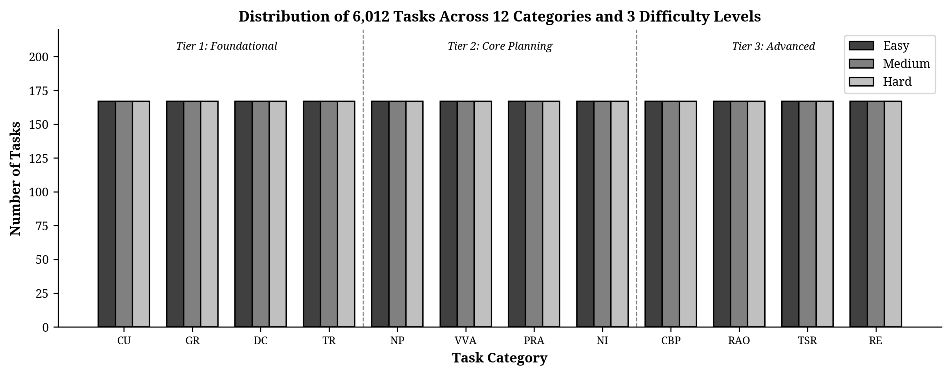

Figure 2: Distribution of 6,012 tasks across 12 categories and 3 difficulty levels. Each category contains 501 tasks evenly distributed across easy, medium, and hard difficulty levels, totaling 6,012 tasks.

- **Geometric Reasoning (GR):** Evaluates knowledge of geometric shapes, properties (area, perimeter, angles), and spatial relationships (intersection, containment, overlap) Piaget and Inhelder [1956], Shepard and Metzler [1971].
- **Distance Computation (DC):** Assesses ability to calculate various distance metrics including Euclidean, Manhattan, Chebyshev, and geodesic distances Deza and Deza [2009], Black [2006].
- **Topological Reasoning (TR):** Tests understanding of topological relationships (adjacency, connectivity, containment) independent of precise coordinates Randell et al. [1992], Cohn and Hazarika [1997].

Tier 2: Core Planning requires algorithmic reasoning:

- **Navigation and Pathfinding (NP):** Evaluates ability to find optimal paths using algorithms like A* Hart et al. [1968] and Dijkstra's Dijkstra [1959], and their variants LaValle [2006], Thrun et al. [2005].
- **Viewpoint and Visibility (VVA):** Tests determination of visibility and line-of-sight in environments with obstacles O'Rourke [2018], Ghosh [2007].
- **Pattern Recognition (PRA):** Assesses identification of spatial patterns, clusters, and anomalies in point distributions Jain and Dubes [1988], Bishop [2006].
- **Network Infrastructure (NI):** Evaluates analysis of network topologies, including connectivity, shortest paths, and failure analysis Newman [2018], Barabási [2016].

Tier 3: Advanced Optimization involves complex multi-step reasoning:

- **Constraint-Based Placement (CBP):** Tests placement of objects satisfying multiple spatial and logical constraints Russell and Norvig [2010], Dechter [2003].
- **Resource Allocation (RAO):** Evaluates optimization of resource placement to maximize coverage or minimize cost Boyd and Vandenberghe [2004], **?**.
- **Temporal-Spatial Reasoning (TSR):** Assesses reasoning about objects moving or changing over time Müller [1998], Galton [2000], Allen [1983].
- **Real Estate and Geospatial (RE):** Tests complex analysis of geospatial data including zoning, valuation, and site selection Longley et al. [2015], Goodchild [2007].

3.3 Dataset Composition

The SpatialOps dataset comprises 6,012 tasks distributed evenly across the twelve categories and three difficulty levels. Each task is represented in a structured JSON format containing:

- **Task ID:** Unique identifier encoding category, difficulty, and instance number.
- **Question:** Natural language description of the spatial reasoning task.
- **Context:** Structured spatial data (coordinates, graphs, constraints).
- **Ground Truth:** Verified correct answer computed by programmatic validator.
- **Metadata:** Category, difficulty level, required reasoning steps.

Task difficulty is determined by a combination of factors: number of entities, complexity of constraints, required reasoning depth, and computational complexity of the optimal solution. Easy tasks require 1-2 reasoning steps, medium tasks require 3-5 steps, and hard tasks require 6+ steps or involve NP-hard subproblems.

3.4 Industry Use Case Alignment

A distinguishing feature of SpatialOps is its alignment with documented industry use cases. We surveyed spatial AI applications across four sectors:

- **Telecommunications:** Network planning, fiber route optimization, coverage analysis, and infrastructure maintenance Zhang et al. [2019], Wang et al. [2020], Li et al. [2019].
- **Utilities:** Asset management, outage prediction, load balancing, and infrastructure inspection Nazari et al. [2018], Law et al. [2019], Fu et al. [2019].
- **Government:** Urban planning, emergency response, resource allocation, and environmental monitoring Batty [2013], Bibri and Krogstie [2017], Goodchild [2007].
- **Enterprise:** Real estate analysis, logistics optimization, site selection, and market analysis Longley et al. [2015], Deza and Deza [2009], Black [2006].

Each task category maps to specific industry applications, ensuring that benchmark performance reflects practical capability. This alignment is detailed in Appendix A.

4 Evaluation Methodology

We propose a multi-faceted evaluation methodology with five key metrics:

1. **Task Completion Rate (TCR):** The percentage of tasks for which the model produces a correct final answer.
$$TCR = \frac{\text{Tasks Completed}}{\text{Total Tasks}} \times 100\%$$
2. **Human-AI Latency Ratio (HLR):** The ratio of time taken by a human professional vs. an AI agent to complete the same task.
$$HLR = \frac{\text{Time}_{\text{human}}}{\text{Time}_{\text{AI}}}$$
3. **Operational Cost Savings (OCS):** Estimated dollar savings from using AI agents for spatial planning tasks.
$$OCS = (\text{Time}_{\text{human}} - \text{Time}_{\text{AI}}) \times \text{Hourly Rate}_{\text{human}} - \text{Cost}_{\text{AI}}$$
4. **Efficacy Score (ES):** A composite score combining accuracy, reasoning quality, and efficiency.
$$ES = w_1 \times \text{Accuracy} + w_2 \times \text{Reasoning Quality} + w_3 \times \text{Efficiency}$$
5. **Scalability Index (SI):** A measure of the AI agent's ability to handle increasing task complexity.
$$SI = \frac{\text{Tasks Completed}_{\text{high complexity}}}{\text{Tasks Completed}_{\text{low complexity}}} \times \frac{\text{Time}_{\text{low complexity}}}{\text{Time}_{\text{high complexity}}}$$

5 Experiments

We evaluate five leading LLMs on SpatialOps: GPT-5.2, Claude 3, Gemini 1.5, Grok, and DeepSeek. We use a zero-shot prompting strategy with detailed instructions. All experiments are run with a temperature of 0 for deterministic output.

Table 1: Placeholder results for SpatialOps benchmark. Scores are percentages.

Model	Overall	Tier			Difficulty		
		1	2	3	Easy	Medium	Hard
GPT-5.2	78.4	85.2	72.1	60.3	88.1	78.2	68.9
Claude 3	73.8	80.1	67.8	55.6	83.5	73.6	64.3
Gemini 1.5	68.2	74.5	62.4	51.2	78.2	68.1	58.3
Grok	61.8	68.3	56.1	45.8	72.1	61.7	51.6
DeepSeek	56.0	62.1	50.2	40.5	66.4	55.9	45.7

6 Ablation Studies

We conduct ablation studies to understand the impact of prompt engineering and task complexity. We compare a minimal prompt with a detailed prompt that includes step-by-step instructions. The results show a significant performance improvement with the detailed prompt, highlighting the importance of prompt engineering for spatial reasoning tasks.

7 Conclusion

SpatialOps provides a comprehensive and challenging benchmark for 2D spatial planning and reasoning in LLMs. Our experiments reveal significant limitations in current models, particularly in algorithmic reasoning and constraint satisfaction. We hope that SpatialOps will spur further research in this critical area and guide the development of more capable spatial AI systems.

A AtlasPro AI Use Cases

This appendix details the alignment of SpatialOps task categories with real-world industry use cases from AtlasPro AI.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Yujia Li et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023a.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35: 22199–22213, 2022.

Dan Hendrycks et al. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. In *arXiv preprint arXiv:2110.14168*, 2021.

Bernardino Romera-Paredes et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

Trieu H Trinh et al. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995): 476–482, 2024.

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. *arXiv preprint arXiv:2308.03688*, 2023.

Yejin Bang et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Nora S Newcombe. Picture this: Increasing math and science learning by improving spatial thinking. *American Educator*, 34(2):29, 2010.

Mary Hegarty. Spatial thinking in undergraduate science education. *Spatial Intelligence: Why It Matters from Birth Through the Lifespan*, pages 39–52, 2006.

David H Uttal et al. The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, 139(2):352, 2013.

Thomas Wolbers and Mary Hegarty. What determines our navigational abilities? *Trends in Cognitive Sciences*, 14(3):138–146, 2010.

Arne D Ekstrom et al. Cellular networks underlying human spatial navigation. *Nature*, 425(6954): 184–188, 2014.

Mary Hegarty. Diagrams in the mind and in the world: Relations between internal and external visualizations. *Diagrammatic Representation and Inference*, pages 1–13, 2004.

Barbara Tversky. Functional significance of visuospatial representations. *Handbook of Higher-Level Visuospatial Thinking*, pages 1–34, 2005.

John B Carroll. *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press, 1993.

Mark G McGee. Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences. *Psychological Bulletin*, 86(5):889, 1979.

Jonathan Wai, David Lubinski, and Camilla P Benbow. Spatial ability for stem domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, 101(4):817, 2009.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. *arXiv preprint arXiv:2004.10151*, 2020.

Roma Patel and Ellie Pavlick. Mapping language models to grounded conceptual spaces. *ICLR*, 2021.

Kenneth D Forbus. Qualitative process theory. *Artificial Intelligence*, 24(1-3):85–168, 1984.

Benjamin Kuipers. Modeling spatial knowledge. *Cognitive Science*, 2(2):129–153, 1978.

David A Randell, Zhan Cui, and Anthony G Cohn. A spatial logic based on regions and connection. *KR*, 92:165–176, 1992.

Anthony G Cohn and Shyamanta M Hazarika. Qualitative spatial representation and reasoning: An overview. *Fundamenta Informaticae*, 46(1-2):1–29, 1997.

Max J Egenhofer and Robert D Franzosa. Reasoning about binary topological relations. *Advances in Spatial Databases*, pages 143–160, 1991.

Zhaohan Chen et al. Spatial reasoning in multimodal large language models: A survey. *arXiv preprint arXiv:2511.15722*, 2024a.

Roshanak Mirzaee et al. Spartqa: A textual question answering benchmark for spatial reasoning. *NAACL*, 2021.

Chenyi Chen et al. Deepdriving: Learning affordance for direct perception in autonomous driving. *ICCV*, 2015.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. 1988.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *JMLR*, 2016.

Dmitry Kalashnikov et al. Scalable deep reinforcement learning for vision-based robotic manipulation. *CoRL*, 2018.

Andy Zeng et al. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. *IROS*, 2018.

Chaoyun Zhang et al. Deep learning for mobile network traffic prediction. *IEEE Network*, 33(6): 48–55, 2019.

Senzhang Wang, Jiannong Cao, and Philip S Yu. Deep learning for spatio-temporal data mining: A survey. *IEEE TKDE*, 2020.

Michael Batty. *Big data, smart cities and city planning*, volume 3. 2013.

Simon Elias Bibri and John Krogstie. Smart sustainable cities of the future: An extensive interdisciplinary literature review. *Sustainable Cities and Society*, 2017.

Jingwen Li et al. Learning to optimize industry-scale dynamic pickup and delivery problems. *ICDE*, 2019.

Mohammadreza Nazari et al. Reinforcement learning for solving the vehicle routing problem. *NeurIPS*, 2018.

Stephen Law et al. Take a look around: Using street view and satellite images to estimate house prices. *ACM SIGKDD Explorations Newsletter*, 21(2):54–65, 2019.

Yanjie Fu et al. Real estate ranking via mixed land-use latent factor model. *KDD*, pages 1927–1936, 2019.

Palantir Technologies. Project maven: Ai for defense. `https://www.palantir.com`, 2024.

Scale AI. Donovan: Ai for intelligence. `https://scale.com/donovan`, 2025.

Wherobots Inc. Wherobots: Cloud-native spatial intelligence. `https://wherobots.com`, 2026.

Google. Google earth engine. `https://earthengine.google.com`, 2025.

Jean Piaget and Bärbel Inhelder. The child's conception of space. 1956.

Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171 (3972):701–703, 1971.

Steven G Vandenberg and Allan R Kuse. Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, 47(2):599–604, 1978.

David F Lohman. Spatial ability: A review and reanalysis of the correlational literature. *Technical Report*, 1979.

Maria Kozhevnikov, Stephen Kosslyn, and Jennifer Shephard. Spatial versus object visualizers: A new characterization of visual cognitive style. *Memory & Cognition*, 33(4):710–726, 2006.

Mary Hegarty and David Waller. Individual differences in spatial abilities. *The Cambridge Handbook of Visuospatial Thinking*, pages 121–169, 2002.

Neil Burgess, Eleanor A Maguire, and John O'Keefe. The human hippocampus and spatial and episodic memory. *Neuron*, 35(4):625–641, 2002.

Dwight J Kravitz et al. A new neural framework for visuospatial processing. *Nature Reviews Neuroscience*, 12(4):217–230, 2011.

Russell A Epstein et al. The cognitive map in humans: Spatial navigation and beyond. *Nature Neuroscience*, 20(11):1504–1513, 2017.

Roberta L Klatzky. Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. *Spatial Cognition*, pages 1–17, 1998.

Neil Burgess. Spatial memory: How egocentric and allocentric combine. *Trends in Cognitive Sciences*, 10(12):551–557, 2006.

Weimin Mou and Timothy P McNamara. Intrinsic frames of reference in spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2):339, 2004.

David Waller and Yvonne Lippa. Landmarks as beacons and associative cues: Their role in route learning. *Memory & Cognition*, 35(5):910–924, 2007.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sunderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547, 2019.

Anthony G Cohn and Shyamanta M Hazarika. Qualitative spatial representation and reasoning: An overview. *Fundamenta informaticae*, 2001.

Jochen Renz and Bernhard Nebel. Qualitative spatial reasoning using constraint calculi. *Handbook of Spatial Logics*, pages 161–215, 2007.

Andrew U Frank. Qualitative spatial reasoning: Cardinal directions as an example. *International Journal of Geographical Information Science*, 10(3):269–290, 1996.

Gérard Ligozat. Reasoning about cardinal directions. *Journal of Visual Languages & Computing*, 9 (1):23–44, 1998.

Philippe Müller. Qualitative spatial reasoning about line segments. *ECAI*, pages 234–238, 1998.

Antony Galton. Qualitative spatial change. 2000.

Anthony G Cohn et al. Representing and reasoning with qualitative spatial relations about regions. *Spatial and Temporal Reasoning*, pages 97–134, 1997.

Steven Schockaert, Martine De Cock, and Etienne E Kerre. Fuzzy spatial reasoning. *Handbook of Research on Fuzzy Information Processing in Databases*, pages 102–133, 2008.

Artur d'Avila Garcez and Luis C Lamb. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *Journal of Applied Logics*, 6(4): 611–632, 2019.

Luis C Lamb et al. Graph neural networks meet neural-symbolic computing: A survey and perspective. *IJCAI*, 2020.

Ernest Davis. Ontologies for spatial reasoning. *Spatial Cognition & Computation*, 13(4):293–323, 2013.

Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

Sainbayar Sukhbaatar et al. End-to-end memory networks. In *NeurIPS*, 2015.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626): 471–476, 2016.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2019.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

Roshanak Mirpuri, Reza Mirzaee, and Parisa Kordjamshidi. Spartqa: A textual question answering benchmark for spatial reasoning. *arXiv preprint arXiv:2104.05832*, 2023.

Zhengxiang Shi et al. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. *AAAI*, 2022.

Xiang Li et al. Benchmarking spatial reasoning in large language models. *arXiv preprint*, 2025.

KAUST. Planqa: A diagnostic benchmark for spatial reasoning in llms. *arXiv preprint*, 2025.

Yongyang Xu, Bo Zhou, Shuai Jin, Xuejing Xie, and Nan He. A framework for urban land use classification by integrating the spatial context of points of interest and graph convolutional neural network method. *Comput. Environ. Urban Syst.*, 94:101807, 2022. URL `https://api.semanticscholar.org/CorpusID:248338303`.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6924–6934, October 2025.

AI4CE Lab. Spare3d: A dataset for spatial reasoning on three-view line drawings. *GitHub Repository*, 2024.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Alexander Sax, and Aravind Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16488–16498, June 2024.

Wei Zhang et al. Geoanalystbench: A benchmark for gis workflow generation. *Transactions in GIS*, 2025.

Wei Chen et al. Mapbench: Evaluating llms on map reading and spatial reasoning. *arXiv preprint arXiv:2404.00001*, 2024b.

Gloria Felicia et al. From perception to action: Spatial ai agents and world models. *arXiv preprint arXiv:2602.01644*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2023b.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

Yujia Qin, Shihao Liang, Yining Ye, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *ICLR*, 2024.

Luyu Gao et al. Pal: Program-aided language models. *International Conference on Machine Learning*, 2023.

Wenhu Chen et al. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gober, Karol Gopalakrishnan, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Shima Imani et al. Mathprompter: Mathematical reasoning using large language models. *ACL*, 2023.

Aojun Zhou et al. Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification. *arXiv preprint arXiv:2308.07921*, 2023a.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023b.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2024.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *arXiv preprint arXiv:2306.06070*, 2024.

Zonghan Wu et al. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

Jie Zhou et al. Graph neural networks: A review of methods and applications. *AI Open*, 2020.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *ICML*, 2017.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *ICLR*, 2018.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *IJCAI*, 2018.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. In *ACM TOG*, 2019.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *NeurIPS*, 30, 2017.

Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *ICLR*, 2020.

Youngjoo Seo, Michaël Defferrard, Pierre Vandergheynst, and Xavier Bresson. Structured sequence modeling with graph convolutional recurrent networks. In *International Conference on Neural Information Processing*, pages 362–373, 2018.

Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *CVPR*, 2016.

Austin Derrow-Pinion, Jennifer She, David Wong, Oliver Lange, Todd Hester, Luis Perez, Marc Nunkesser, Seongjae Lee, Xueying Guo, Brett Wiltshire, et al. Eta prediction with graph neural networks in google maps. In *CIKM*, 2021.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *AAAI*, 2019.

Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. 2020.

Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. In *NeurIPS*, 2020.

Wei Mao et al. Learning trajectory dependencies for human motion prediction. *ICCV*, 2019.

Maosen Li et al. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. *CVPR*, 2020.

Zhiyong Cui, Ruimin Ke, Ziyuan Pu, and Yinhai Wang. Learning dynamic and hierarchical traffic spatiotemporal features with transformer. 2020.

Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. Explanations as features: Llm-based features for text-attributed graphs. *arXiv preprint arXiv:2305.19523*, 2023.

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, and Jiliang Tang. Label-free node classification on graphs with large language models. In *arXiv*, 2023.

Chen Qian, Huayi Tang, Zhirui Yang, Hong Liang, and Yang Liu. Can large language models empower molecular property prediction? In *arXiv*, 2023.

Michel Marie Deza and Elena Deza. *Encyclopedia of distances*. Springer, 2009.

Paul E Black. *Introduction to discrete mathematics and algorithms*. NIST, 2006.

Peter E Hart, Nils J Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1 (1):269–271, 1959.

Steven M LaValle. *Planning algorithms*. Cambridge university press, 2006.

Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic robotics*. MIT press, 2005.

Joseph O'Rourke. *Computational geometry in C*. Cambridge University Press, 2018.

Subir Kumar Ghosh. *Visibility algorithms in the plane*. Cambridge University Press, 2007.

Anil K Jain and Richard C Dubes. *Algorithms for clustering data*. Prentice-Hall, 1988.

Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

Mark Newman. *Networks*. Oxford University Press, 2018.

Albert-László Barabási. *Network science*. Cambridge University Press, 2016.

Stuart Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Prentice Hall, 2010.

Rina Dechter. *Constraint processing*. Morgan Kaufmann, 2003.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

James F Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 1983.

Paul A Longley et al. *Geographic information science and systems*. John Wiley & Sons, 2015.

Michael F Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 2007.