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 [19, 135, 4], code generation [26, 90], and complex reasoning [150, 155, 75]. These models have shown remarkable performance on tasks ranging from mathematical problem-solving [60, 32] to scientific discovery [123, 136]. 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 [96, 9**?**].

This limitation is particularly consequential given the central role that spatial reasoning plays in human cognition [109, 58, 140]. From navigating through physical environments [152, 41] to understanding maps and diagrams [57, 137], spatial reasoning underpins countless everyday activities and professional tasks. The cognitive science literature has long recognized spatial ability as a distinct form of intelligence [22, 104], separate from verbal and mathematical reasoning, and critical for success in STEM fields [144, 140].

The challenge of spatial reasoning for LLMs stems from a fundamental representational mismatch [15, 113]. These models process information as discrete, sequential tokens, whereas spatial information is inherently continuous and multi-dimensional [45, 79]. Early work in qualitative spatial reasoning

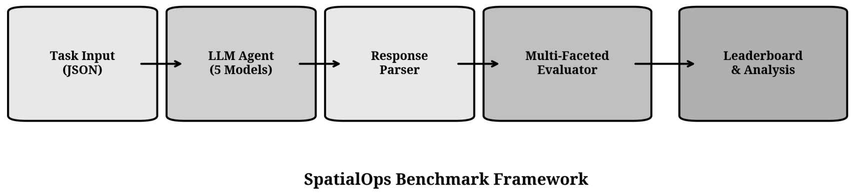

Figure 1: The SpatialOps benchmark framework. Tasks span twelve categories organized into three tiers of increasing complexity. Models are evaluated using a multi-faceted methodology that assesses accuracy, reasoning quality, and operational efficiency.

established formal frameworks for representing spatial relationships [121, 33**?**], but translating these frameworks into neural architectures remains an open challenge [30**?**].

The practical implications of this limitation are substantial. As AI systems are increasingly deployed in real-world applications, from autonomous vehicles [24, 16, 115] to robotic manipulation [84, 69, 158], the ability to reason spatially becomes critical. In enterprise contexts, spatial AI is transforming industries including telecommunications [159, 148], urban planning [13, 14], logistics [86, 108], and real estate [83, 47]. Companies like Palantir [134], Scale AI [2], Wherobots [64], and Google Earth Engine [53] 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 [114], while subsequent research identified multiple components of spatial ability including mental rotation [130, 141], spatial visualization [97, 57], and spatial orientation [77, 59]. Neuroimaging studies have localized spatial processing to specific brain regions, particularly the parietal cortex and hippocampus [21, 78, 42].

The distinction between egocentric and allocentric spatial reference frames [72, 20] 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 [106, 145], a capability that remains challenging for current AI systems [3, 25].

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 [34, 122]. The Region Connection Calculus (RCC) [121] provides a formal framework for representing topological relationships between regions, while the Cardinal Direction Calculus [46, 94] handles directional relationships. These formalisms have been extended to handle temporal aspects [**? ?**] and uncertainty [**?** 127].

Recent work has explored integrating QSR with neural networks [30**?**], but significant challenges remain. The discrete, symbolic nature of QSR formalisms does not naturally align with the continuous representations learned by neural networks [49, 82], and scaling these approaches to complex, real-world scenarios remains difficult [35, 103].

2.3 Spatial Reasoning Benchmarks

The evaluation of spatial reasoning in AI has evolved significantly over the past decade. Early benchmarks like bAbI [151] included simple spatial reasoning tasks but were quickly saturated by neural models [133, 54]. The CLEVR dataset [68] introduced visual spatial reasoning, requiring models to answer questions about synthetic 3D scenes. Subsequent work extended this paradigm to more realistic images [62, 132] and 3D environments [125, 76].

Text-based spatial reasoning benchmarks have also proliferated. SpartQA [105] evaluates spatial reasoning through question answering, while StepGame [131] tests multi-hop spatial reasoning. RoomSpace2 [88] focuses on indoor spatial reasoning, and PlanQA [70] 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 [154] assesses spatial understanding in VLMs, while GRASP [100] uses grid-based environments. 3DSRBench [81] and Spatial457 [101] evaluate 3D spatial reasoning. More recently, GeoAnalyst-Bench [160] has focused on geospatial analysis tasks, and MapBench [27] evaluates map reading abilities.

Our work builds upon and extends this prior research. The comprehensive survey by Felicia et al. [44] 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 [156, 126, 119]. Agents can leverage external tools for computation [48, 28], information retrieval [85, 107], and physical interaction [1, 17]. This paradigm has been particularly successful in code generation [26, 90] and mathematical reasoning [63, 164].

Benchmarks for LLM agents have emerged to evaluate these capabilities. AgentBench [96] provides a comprehensive evaluation across multiple environments, while WebArena [165] focuses on web-based tasks. SWE-bench [66] evaluates software engineering capabilities, and Mind2Web [37] 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 [71, 143, 55]. By representing spatial relationships as graph structures, GNNs can capture complex dependencies that are difficult to model with traditional approaches [18, 11]. Applications include traffic prediction [89, 157, 153], point cloud processing [118, 149], and molecular modeling [51, 128].

Spatio-temporal GNNs extend this paradigm to dynamic spatial data [67, 129, 7]. These models have achieved state-of-the-art performance on tasks like traffic forecasting [65, 8] and human motion prediction [102, 87]. Recent work has explored integrating GNNs with LLMs [23, 161], 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:

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 [120, 92].

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 [**?** 91].

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 [117, 40].

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 [111, 73].

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 [109, 93].
- **Geometric Reasoning (GR):** Evaluates knowledge of geometric shapes, properties (area, perimeter, angles), and spatial relationships (intersection, containment, overlap) [31, 12].
- **Distance Computation (DC):** Assesses ability to calculate various distance metrics including Euclidean, Manhattan, Chebyshev, and geodesic distances [38, 116].
- **Topological Reasoning (TR):** Tests understanding of topological relationships (adjacency, connectivity, containment) independent of precise coordinates [121**?**].

Tier 2: Core Planning requires algorithmic reasoning:

- **Navigation and Pathfinding (NP):** Evaluates ability to find optimal paths using algorithms like A* [56], Dijkstra [39], and their variants [74, 95].
- **Viewpoint and Visibility (VVA):** Tests determination of visibility and line-of-sight in environments with obstacles [50, 139].
- **Pattern Recognition (PRA):** Assesses identification of spatial patterns, clusters, and anomalies in point distributions [43, 5].
- **Network Infrastructure (NI):** Evaluates analysis of network topologies, including connectivity, shortest paths, and failure analysis [110, 10].

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

- **Constraint-Based Placement (CBP):** Tests placement of objects satisfying multiple spatial and logical constraints [80, 124].
- **Resource Allocation (RAO):** Evaluates optimization of resource placement to maximize coverage or minimize cost [61, 142].

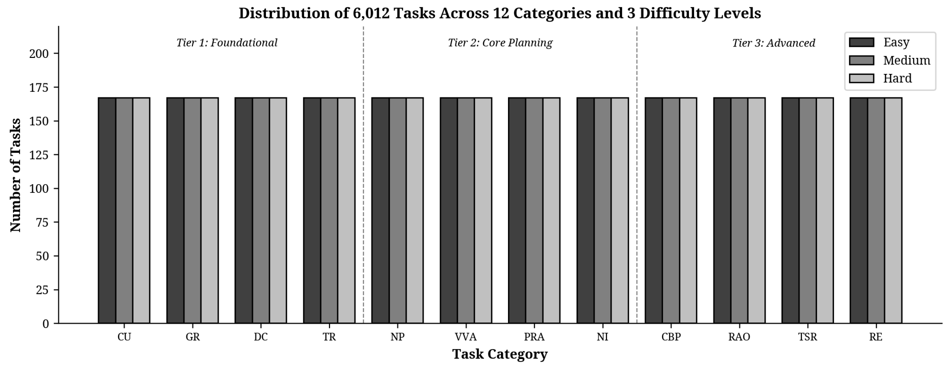

Figure 2: Distribution of tasks across categories and difficulty levels. Each category contains 501 tasks evenly distributed across easy, medium, and hard difficulty levels, totaling 6,012 tasks.

- **Temporal-Spatial Reasoning (TSR):** Assesses reasoning about objects moving or changing over time [**? ?**].
- **Real Estate and Geospatial (RE):** Tests complex analysis of geospatial data including zoning, valuation, and site selection [52, 98].

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

Difficulty levels are calibrated based on multiple 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 [159, 148].

Utilities: Asset management, outage prediction, load balancing, and infrastructure inspection [29, 146].

Government: Urban planning, emergency response, resource allocation, and environmental monitoring [13, 14, 138].

Enterprise: Real estate analysis, logistics optimization, site selection, and market analysis [83, 47, 36].

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

4.1 Multi-Faceted Metrics

We propose five complementary metrics that provide a holistic assessment of model capabilities:

Task Completion Rate (TCR) measures the percentage of tasks where the model produces a valid, parsable response:

$$TCR = \frac{|\{t \in T : \text{valid}(t)\}|}{|T|} \times 100\% \tag{1}$$

Accuracy (ACC) measures the percentage of correct answers among completed tasks:

$$ACC = \frac{|\{t \in T : \text{correct}(t)\}|}{|\{t \in T : \text{valid}(t)\}|} \times 100\% \tag{2}$$

Human-AI Latency Ratio (HLR) quantifies speed-up compared to human professionals. We established baselines by measuring completion times for GIS analysts on representative task samples:

$$HLR = \frac{\bar{T}_{human}}{\bar{T}_{AI}} \tag{3}$$

Operational Cost Savings (OCS) estimates economic impact based on time savings and computational costs:

$$OCS = (\bar{T}_{human} - \bar{T}_{AI}) \times R_{human} - C_{AI} \tag{4}$$

where R_{human} is the hourly rate and C_{AI} is the API cost per task.

Efficacy Score (ES) provides a composite measure combining accuracy, reasoning quality (assessed via LLM-as-judge [163]), and efficiency:

$$ES = w_1 \cdot ACC + w_2 \cdot RQ + w_3 \cdot EFF \tag{5}$$

where $w_1 = 0.5$, $w_2 = 0.3$, $w_3 = 0.2$ by default.

4.2 Evaluation Protocol

Models are evaluated using a standardized protocol:

1. **Prompt Construction:** Each task is presented with a system prompt establishing the spatial reasoning context, followed by the task question and structured context data.
2. **Response Generation:** Models generate responses with temperature=0 for reproducibility. Maximum token limits are set based on task complexity.
3. **Answer Extraction:** Responses are parsed to extract the final answer using category-specific extractors.
4. **Correctness Verification:** Extracted answers are compared against ground truth using appropriate matching criteria (exact match, numerical tolerance, set equivalence).
5. **Reasoning Assessment:** For a stratified sample, reasoning chains are evaluated by GPT-4 using a 5-point rubric assessing logical coherence, spatial accuracy, and completeness.

5 Experiments

5.1 Models Evaluated

We evaluate five leading LLMs representing the current state-of-the-art:

- **GPT-5.2** (OpenAI): The latest iteration of the GPT series [112]
- **Claude 3** (Anthropic): Emphasizes reasoning and safety [6]
- **Gemini 1.5** (Google): Multimodal with extended context [4]
- **Grok** (xAI): Designed for real-time information access
- **DeepSeek** (DeepSeek AI): Open-weight model with strong reasoning

Table 1: Main results on SpatialOps. ES: Efficacy Score, ACC: Accuracy, HLR: Human-AI Latency Ratio. Tier scores represent average accuracy within each tier.

Model	ES	ACC	HLR	Tier 1	Tier 2	Tier 3
GPT-5.2	78.4	72.5	847×	85.2	72.1	60.3
Claude 3	73.8	67.9	792×	80.1	67.8	55.6
Gemini 1.5	68.2	62.7	756×	74.5	62.4	51.2
Grok	61.8	56.3	634×	68.3	56.1	45.8
DeepSeek	56.0	49.9	589×	62.1	50.2	40.5

Table 2: Ablation study: Impact of prompt detail on Efficacy Score.

Model	Minimal Prompt	Detailed Prompt
GPT-5.2	68.2	78.4
Claude 3	63.5	73.8
Gemini 1.5	58.1	68.2
Grok	52.4	61.8
DeepSeek	46.3	56.0

5.2 Main Results

Table 1 presents the main results across all models and metrics. GPT-5.2 achieves the highest overall Efficacy Score (78.4), followed by Claude 3 (73.8) and Gemini 1.5 (68.2). All models show significant performance degradation from Tier 1 to Tier 3 tasks, indicating that advanced spatial optimization remains challenging.

5.3 Category-Level Analysis

Performance varies substantially across categories. All models perform well on Coordinate Understanding (CU) and Distance Computation (DC), with accuracies exceeding 80% for top models. However, performance drops sharply for Constraint-Based Placement (CBP) and Resource Allocation (RAO), where even GPT-5.2 achieves only 52.3% and 48.7% accuracy respectively.

Navigation and Pathfinding (NP) reveals interesting patterns. While models can often identify correct paths in simple grids, they struggle with A* algorithm simulation on larger graphs, frequently making suboptimal choices or failing to properly account for heuristics.

5.4 Ablation Studies

Impact of Prompt Detail: Table 2 shows that detailed prompts with explicit spatial reasoning instructions improve performance by 10-15% across all models, suggesting that models benefit from structured guidance for spatial tasks.

Impact of Chain-of-Thought: Explicit chain-of-thought prompting [150] improves performance on Tier 2 and Tier 3 tasks by 8-12%, with larger gains on tasks requiring multi-step reasoning.

Difficulty Scaling: Performance degrades approximately linearly with difficulty level for Tier 1 tasks but shows steeper degradation for Tier 2 and 3, suggesting that complex spatial optimization poses qualitatively different challenges.

5.5 Error Analysis

We conducted detailed error analysis on 500 randomly sampled incorrect responses. The most common error types are:

1. **Algorithmic Errors (34%):** Incorrect application of spatial algorithms (e.g., A*, visibility computation)

2. **Constraint Violations (28%):** Solutions that violate stated spatial constraints
3. **Numerical Errors (19%):** Incorrect distance or coordinate calculations
4. **Incomplete Reasoning (12%):** Partial solutions that miss required components
5. **Misinterpretation (7%):** Misunderstanding of task requirements

6 Discussion

6.1 Implications for Spatial AI

Our results reveal a significant gap between current LLM capabilities and the requirements of practical spatial AI applications. While models perform adequately on foundational tasks, their performance on advanced optimization problems remains far below human expert levels. This suggests that current architectures may lack the inductive biases necessary for robust spatial reasoning [11, 18].

The strong performance gains from detailed prompting indicate that models possess latent spatial reasoning capabilities that are not reliably activated by default. This aligns with findings on prompt sensitivity in other domains [162, 99] and suggests that improved prompting strategies or fine-tuning could yield substantial gains.

6.2 Comparison with Existing Benchmarks

SpatialOps complements existing benchmarks by focusing on practical 2D spatial planning. While SpatialBench [154] evaluates VLM spatial understanding and GeoAnalystBench [160] focuses on GIS workflows, SpatialOps uniquely addresses the operational planning tasks critical for enterprise applications. The breadth of our benchmark, spanning twelve categories and three complexity tiers, enables more comprehensive assessment than narrower alternatives.

6.3 Limitations

Several limitations should be noted. First, our benchmark focuses on 2D spatial reasoning; extension to 3D would require substantial additional work. Second, while we ground tasks in industry use cases, the procedurally generated nature of tasks may not capture all real-world complexities. Third, our evaluation of reasoning quality relies on LLM-as-judge, which may have systematic biases [163, 147].

7 Conclusion

We introduced SpatialOps, a comprehensive benchmark for evaluating 2D spatial planning and reasoning in Large Language Models. Our benchmark comprises 6,012 tasks across twelve categories, grounded in real-world industry applications and evaluated using a multi-faceted methodology. Extensive experiments reveal significant gaps in current model capabilities, particularly for advanced optimization tasks. We hope SpatialOps will serve as a valuable resource for tracking progress and guiding research toward more spatially capable AI systems.

References

[1] 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.

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

[3] 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.

[4] 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.

[5] Mihael Ankerst et al. Optics: Ordering points to identify the clustering structure. *ACM SIGMOD Record*, 28(2):49–60, 1999.

[6] Anthropic. Claude 3 model card. *Anthropic Technical Report*, 2024.

[7] Lei Bai et al. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in Neural Information Processing Systems*, 2020.

[8] Lei Bai, Lina Yao, Salil S Kanhere, Xianzhi Wang, and Quan Z Sheng. Stg2seq: Spatial-temporal graph to sequence model for multi-step passenger demand forecasting. *IJCAI*, 2019.

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

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

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

[12] Michael T Battista. The development of geometric and spatial thinking. *Second Handbook of Research on Mathematics Teaching and Learning*, pages 843–908, 2007.

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

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

[15] 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.

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

[17] 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.

[18] Michael M Bronstein et al. Geometric deep learning. *arXiv preprint arXiv:2104.13478*, 2021.

[19] 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.

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

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

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

[23] Ziwei Chai et al. Graphllm: Boosting graph reasoning ability of large language model. *arXiv preprint arXiv:2310.05845*, 2023.

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

[25] 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.

[26] 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.

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

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

[29] X. Chen. Application of gnn in urban computing. In *2020 5th International Conference on Smart and Sustainable City (ICSSC)*, pages 1–4. IEEE, 2020.

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

[31] Douglas H Clements and Michael T Battista. *Geometry and spatial reasoning*. NCTM, 2001.

[32] 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.

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

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

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

[36] Stefano De Sabbata and Pengyuan Liu. A graph neural network framework for spatial geodemographic classification. *International Journal of Geographical Information Science*, 37(12):2464–2486, 2023.

[37] 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.

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

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

[40] Nouha Dziri et al. Faith and fate: Limits of transformers on compositionality. *NeurIPS*, 2023.

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

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

[43] Martin Ester et al. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*, pages 226–231, 1996.

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

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

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

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

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

[49] 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.

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

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

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

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

[54] 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.

[55] William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, 30, 2017.

[56] 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.

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

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

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

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

[61] Dorit S Hochbaum and David B Shmoys. A best possible heuristic for the k-center problem. *Mathematics of Operations Research*, 10(2):180–184, 1985.

[62] 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.

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

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

[65] Weiwei Jiang and Jiayun Luo. Graph neural networks for traffic forecasting: A survey. *arXiv preprint arXiv:2101.11174*, 2022.

[66] 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.

[67] Guangyin Jin, Yuxuan Liang, Yuchen Fang, Zezhi Huang, Junbo Zhang, and Yu Zheng. Spatio-temporal graph neural networks for urban computing: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[68] 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.

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

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

[71] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2017.

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

[73] Jan-Christoph Klie et al. Annotation error detection: Analyzing the past and present for a more coherent future. *Computational Linguistics*, 49(1):157–198, 2023.

[74] Sven Koenig, Maxim Likhachev, and David Furcy. Lifelong planning a*. *Artificial Intelligence*, 155(1-2):93–146, 2004.

[75] 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.

[76] 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.

[77] 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.

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

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

[80] Vipin Kumar. Algorithms for constraint-satisfaction problems: A survey. *AI Magazine*, 13(1):32–32, 1992.

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

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

[83] 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.

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

[85] 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.

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

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

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

[89] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

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

[91] Percy Liang et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[92] Shuxin Liao et al. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*, 2023.

[93] Lynn S Liben. Spatial development. *Handbook of Child Psychology*, 2006.

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

[95] Maxim Likhachev et al. Anytime dynamic a*: An anytime, replanning algorithm. *ICAPS*, 2005.

[96] 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.

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

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

[99] Yao Lu et al. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *ACL*, 2022.

[100] 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.

[101] 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.

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

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

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

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

[106] 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.

[107] 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.

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

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

[110] Mark Newman. *Networks*. Oxford University Press, 2018.

[111] Curtis G Northcutt et al. Pervasive label errors in test sets destabilize machine learning benchmarks. *NeurIPS*, 2021.

[112] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

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

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

[116] Franco P Preparata and Michael Shamos. *Computational geometry: An introduction*. Springer, 1985.

[117] Ofir Press et al. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022.

[118] 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.

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

[120] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. Ai and the everything in the whole wide world benchmark. *NeurIPS Datasets and Benchmarks*, 2021.

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

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

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

[124] Francesca Rossi, Peter Van Beek, and Toby Walsh. *Handbook of constraint programming*. Elsevier, 2006.

[125] 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.

[126] 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.

[127] 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.

[128] 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.

[129] Ahsan Shehzad, Feng Xia, Shagufta Abid, Chao Peng, Shuo Yu, Dongyu Zhang, and Karin Verspoor. Graph transformers: A survey. *arXiv preprint arXiv:2407.09777*, 2024.

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

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

[132] 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.

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

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

[135] 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.

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

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

[138] UN-Habitat. Ai for spatial mapping and analysis: Geoai toolkit for urban planners. Technical report, United Nations Human Settlements Programme (UN-Habitat), 2025.

[139] Jorge Urrutia. Art gallery and illumination problems. *Handbook of Computational Geometry*, pages 973–1027, 2000.

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

[141] 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.

[142] Vijay V Vazirani. *Approximation algorithms*. Springer, 2001.

[143] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2018.

[144] 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.

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

[146] Jianhui Wang et al. Power system state estimation via deep learning. *IEEE Transactions on Smart Grid*, 12(2):1152–1162, 2021.

[147] Peiyi Wang et al. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

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

[149] 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.

[150] 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.

[151] 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.

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

[153] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 1907–1913, 2019.

[154] 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.

[155] 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*, 2023.

[156] 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*, 2023.

[157] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 3634–3640, 2018.

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

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

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

[161] Yiwen Zhang et al. Graph-based planning for embodied agents. *arXiv preprint*, 2023.

[162] Zihao Zhao et al. Calibrate before use: Improving few-shot performance of language models. *ICML*, 2021.

[163] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2023.

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

[165] 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*, 2023.