UrbanGeoEval: A City-Scale Benchmark for Evaluating Large Language Models in Geospatial Reasoning

Anonymous ACL submission

Abstract

Evaluating the geospatial reasoning of Large Language Models (LLMs) in complex urban environments remains a significant challenge, as existing benchmarks often overlook the city scale and fail to disentangle urban environment memory recall from geospatial reasoning abilities. To address this, we introduce UrbanGeoEval, a comprehensive benchmark featuring an innovative dual-module framework. The Knowledge Module uses open-source map data to automatically generate scalable multiple-choice questions that assess memory recall. The Reasoning Module evaluates pure reasoning through 3,148 open-ended questions across seven distinct tasks, with necessary knowledge injected into the context to isolate reasoning capabilities. For evaluation, we employ a LLM-as-a-Judge pipeline for fine-grained analysis of both the reasoning process and final answers. Our extensive experiments on 16 popular LLMs reveal that model scale and high-quality reasoning are critical for performance. We also uncover geographical biases and common weaknesses in quantitative and multi-hop geospatial relationship tasks. UrbanGeoEval provides a robust framework for diagnosing LLM capabilities in urban intelligence and paving the way for future advancements.

1 Introduction

Large Language Models (LLMs) have emerged as a promising tool for Geographic Question Answering (GeoQA; Mai et al., 2021), demonstrating notable potential in understanding about geographic information (Roberts et al., 2023; Huang et al., 2019). The ability to understand and reason about complex urban environments is crucial for maps, navigation, and location-based services. Urban environments concentrate dense points of interest (POIs), road networks, and heterogeneous areas of interest (AOIs), along with fine-grained spatial relations and diverse constraints (Wang et al., 2020).

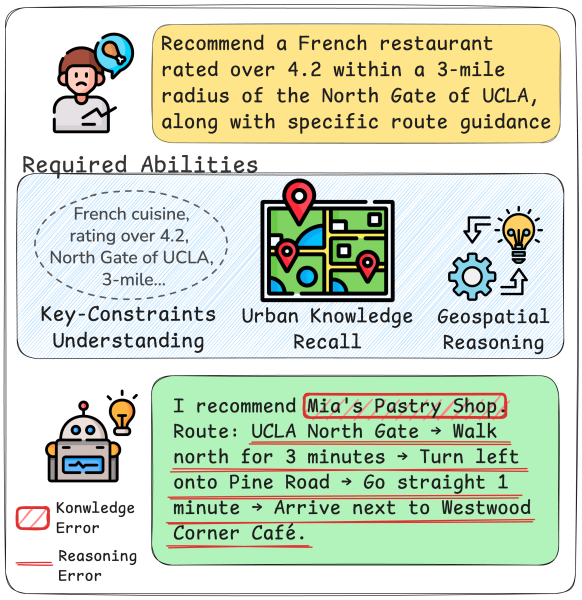

Figure 1: A typical LLM failure on a city-scale geospatial query, where a knowledge error (hallucinating a POI) leads to a subsequent reasoning error (generating an invalid path). This confounding of error types motivates our decoupled benchmark design.

Many real user queries—*Which entrance should I take to minimize walking from this metro stop at rush hour? Can you plan an itinerary to visit these sites with minimal backtracking? From my current coordinates, which café is closest and still open?*—require models to integrate precise coordinates, spatial topology, and attribute semantics under natural-language specifications.

However, existing evaluations rarely operate at this resolution, most GeoQA benchmarks use multiple-choice formats at coarse geographic scales or entangle factual knowledge recall with reasoning process, making it difficult to attribute errors (Roberts et al., 2023; Mai et al., 2024; Feng et al., 2025). For example, if an LLM cannot determine the closest restaurant to a given bus station, is it because it lacks knowledge of their specific loca-

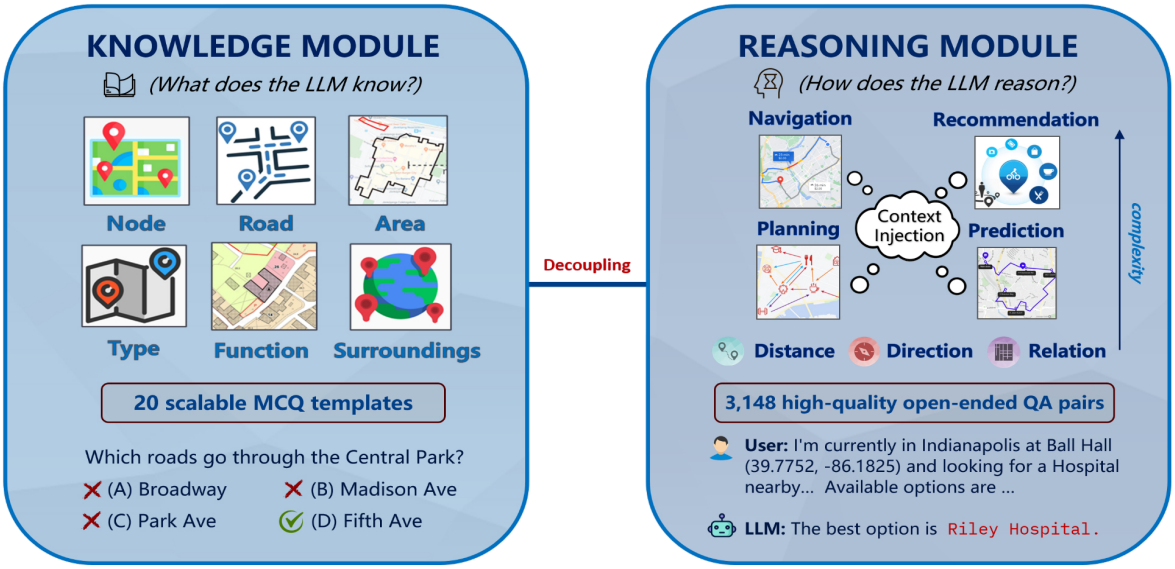

Figure 2: The overall architecture of the UrbanGeoEval benchmark. The Knowledge Module (left) uses an automated pipeline to generate multiple-choice questions from OSM data. The Reasoning Module (right) uses human-annotated scenarios to create open-ended questions that test pure reasoning with in-context knowledge.

tions (a knowledge deficit), or because it cannot correctly calculate and compare multiple routes in geographical space (a reasoning failure)? As shown in Figure 1, we lack a reliable picture of what LLMs can and cannot do in urban geospatial understanding.

To address this, we introduce UrbanGeoEval, a comprehensive benchmark for assessing LLMs on urban geospatial tasks. The design is motivated by *dual-process cognitive theory* (Kahneman, 2011) and recent work that disentangle LLM inference into distinct steps of knowledge retrieval and subsequent reasoning (Jin et al., 2025; Yang et al., 2025b). Following this principle, UrbanGeoEval employs a dual-module architecture (Figure 2) to systematically evaluate these separate facets of geospatial intelligence:

The Knowledge Module probes the model's static representation of the urban geography. We designed a rigorous set of tasks across six core dimensions of urban knowledge including POI attributes, road network properties and area functionalities to test the breadth and precision of its factual memory via automatically generated template-driven multiple-choice questions. This module goes beyond simple fact-checking by using strategically generated distractors to challenge the model's fine-grained perception.

The Reasoning Module evaluates the model's ability to dynamically manipulate geospatial information. To isolate reasoning capabilities, we ensure all necessary geo-entity background (e.g., coordinates, names, attributes) are provided within the context, eliminating the need for latent world knowledge. This frees the model from the burden of recall, allowing us to evaluate model's pure geospatial reasoning skills across seven realistic open-ended tasks comprising 3,148 high quality human annotated problems—ranging from foundational abilities like distance and direction calculation—to advanced, multi-hop complex tasks like itinerary planning and navigation. A structured prompting protocol separates models' chain-of-thought from final answers, enabling process-aware evaluation with rubric-based scoring. Under a unified evaluation framework, we conduct a comprehensive study covering both non-reasoning and reasoning LLMs. Our contributions are threefold:[1]

- We introduce UrbanGeoEval, the first city-scale benchmark to systematically decouple the evaluation of an LLM's factual knowledge from its geospatial reasoning ability.
- We propose a process-aware evaluation framework using an LLM-as-a-Judge with fine-grained scoring rubrics to assess not just the final answer, but the entire reasoning chain.
- We conduct a large-scale analysis of 16 popular LLMs, providing a detailed diagnosis

[1]Dataset and Codebase will be made publicly available upon acceptance.

of their capabilities and revealing critical insights, including geospatial biases and task-specific weaknesses.

2 Related Work

2.1 LLMs for Geospatial Question Answering

Current methods to enhance LLM performance in GeoQA include prompt engineering (Manvi et al., 2024), retrieval-augmented generation (RAG) (Yu et al., 2025; Li et al., 2023), and agent-based systems that use external tools (Gupta et al., 2024; Zhang et al., 2025). While effective, these approaches often offload the core geospatial task to external knowledge sources or pretrained foundation models (Huang et al., 2022; Ding et al., 2023; Balsebre et al., 2024), bypassing an evaluation of the LLM's intrinsic, end-to-end reasoning capabilities. This leaves a gap in understanding the model's core cognitive capacity for geospatial tasks.

2.2 Benchmarks for Geospatial Abilities

Existing benchmarks for the geospatial abilities of LLMs have notable limitations. Text-based benchmarks like CityEval (Feng et al., 2025), STBench (Li et al., 2024), and MapEval (Dihan et al., 2025) often feature simple question formats that fail to capture the underlying reasoning process. Meanwhile, multimodal benchmarks such as MapQA (Chang et al., 2022), MapIQ (Srivastava et al., 2025) and MapWise (Mukhopadhyay et al., 2025) are hindered by the high cost and difficulty of acquiring quality geo-referenced data. Consequently, a critical need remains for a robust, scalable, text-based benchmark for city-scale reasoning, a gap UrbanGeoEval directly addresses.

2.3 The Rise of Reasoning LLMs

A recent frontier is the development of reasoning-specialized LLMs, such as OpenAI's o-series (OpenAI, 2025) and DeepSeek-R1 (DeepSeek-AI et al., 2025), which leverage Reinforcement Learning Fine-Tuning (RFT) to excel at complex, multi-step tasks. This has led to expert-level performance in domains like STEM (Hendrycks et al., 2021; Rein et al., 2023). However, whether these advanced reasoning skills can transfer from abstract domains to noisy, real-world physical spaces remains under-explored, highlighting the need for our benchmark.

Statistic	**Value**
Knowledge Dimensions	6
Question Templates	20
Questions per City	3,000
Total Questions	**27,000**
Evaluated Cities (9 Total)	
Berlin, Dubai, London, New Delhi, New York, Paris, Rome, San Francisco, Tokyo	

Table 1: Data statistics for the Knowledge Module.

3 Benchmark Construction

The UrbanGeoEval benchmark targets realistic urban scenarios where dense POIs, AOIs, and road networks interact with natural-language queries. We disentangle knowledge and reasoning through a dual-module architecture, allowing us to:

(i) testing factual memory recall with carefully controlled multiple-choice items and

(ii) testing pure reasoning with open-ended realistic geospatial problems that include all necessary background as context.

3.1 Knowledge Module

The Knowledge Module is conceived as a series of "knowledge probes" designed to meticulously map the LLM's internal, static representation of the urban world, akin to querying a geographic information system (GIS).

3.1.1 Data Collection and Generation.

Our methodology begins with an automated pipeline that extracts and structures geospatial data (POIs, AOIs, road network) from the rich, open-source OpenStreetMap[2] (OSM) database. Geo-entities information appear as "tags" in OSM and do not include natural-language address information. To bridge the gap between structured data and natural language, we integrate the Nominatim[3] reverse geocoding engine, transforming raw coordinates into human-readable addresses. This curated knowledge base serves as the ground truth for an automated generation pipeline that produces a large-scale, extensible set of multiple-choice questions. For our experiments, we constructed a dataset spans 9 international cities across different continents to ensure geographic diversity, with statistics presented in Table 1.

[2] https://openstreetmap.org/
[3] https://nominatim.org/

Dimension	Template ID	Description
Node	Retrieve_POI_Address	Given a POI name, query its address.
	Retrieve_POI_Name	Given a POI address, query its name.
	Retrieve_Coordinates	Given a POI name, query its coordinates.
	Calculate_Distance	Given two POIs, calculate the great-circle distance between them.
	Calculate_Distance_Context	Same as Calculate_Distance, but includes surrounding context as a hint.
Area	Retrieve_AOI_Address	Given an AOI name, query its address.
	Retrieve_AOI_Name	Given an AOI address, query its name.
	Find_Cover_AOI	Given a POI, find the AOI that spatially contains it.
	Find_Cover_AOI_All	Given a set of POIs, find the minimal AOI that covers all of them.
Road	Query_Length	Given a road name, query its length.
	Find_Junctions	Given a road and an area, find the road's start/end junctions within the area.
	Identify_Connections	Identify roads directly connected to a target road within an area.
Type	Classify_POI	Given a POI name, query its category.
	Landuse_AOI	Given an AOI name, query its land-use type.
	Find_Instances_By_Type	Given a category, select POIs belonging to it.
Function	Compare_POI_Count	Among several AOIs, find the one with the most POIs.
	Compare_Area_Size	Among several AOIs, find the one with the largest area.
	Compare_Type_Diversity	Among several AOIs, find the one with the most diverse POI types.
Surroundings	Infer_POI_From_Scene	Given an environment description, infer the most likely POI type to appear.
	Find_Nearest_POI	Given a POI, find the nearest other POI.

Table 2: Task dimensions and question templates in the Knowledge Module, designed to probe different facets of an LLM's static urban knowledge.

3.1.2 Taxonomy of Knowledge Tasks.

These dimensions cover the full vector geometry stack—POI points ***(Node)***, lines ***(Road)***, and AOI polygons ***(Area)***—together with their attributes and topological relations. Furthermore, we evaluate semantic categorization abilities ***(Type)***, measure basic aggregation skills ***(Function)***, and test contextual environmental awareness ***(Surroundings)***. The data generation process results in a diverse pool of questions covering 20 distinct templates across the six dimensions, as shown in Table 2.

3.1.3 Strategic Distractor Generation.

To truly challenge the models and prevent them from relying on superficial pattern matching, we designed four strategies for generating plausible yet incorrect "distractor" options. These strategies are crafted to test specific aspects of a model's knowledge resolution by creating options with spatial proximity to test location precision, semantic similarity to test disambiguation, numerical perturbation for quantitative accuracy, and similar attribute substitution to verify entity-property linkage.

3.2 Reasoning Module

The Reasoning Module measures process-level geospatial reasoning once relevant facts are provided in-context, so that the model need not recall world facts from pretraining. A central goal is to assess the entire chain-of-thought, capturing not just if a model's answer is correct, but how it arrived at its conclusion, thereby distinguishing genuine inference from lucky guesses.

3.2.1 Background Context Schema

To formalize the contextual knowledge, we define a schema for each core geospatial entity:

- **POI:** Represented as Info(P), which includes its name, coordinates, address, type and other attributes such as ratings and open-hour.
- **Road Segment:** Represented as Info(R), containing its name and a polyline sequence of WGS84 coordinates sampled along its centerline.
- **AOI:** Represented as Info(A), which includes its name, boundary, land-use type, and other necessary attributes. The boundary is defined by a closed ring of 6–10 uniformly sampled WGS84 coordinates.

This structured information is rendered into prompts as either compact key-value blocks or JSON-like fragments.

3.2.2 Taxonomy of Reasoning Tasks

We developed a diverse set of seven realistic reasoning scenarios structured into two tiers of increasing

complexity. The first tier, *Foundational Spatial Reasoning*, covers the building blocks of spatial intelligence, including Distance Calculation, Direction Finding, and Spatial Relation Identification. The second tier, *Advanced Composite Reasoning*, requires multi-hop, complex inference that simulates real-world problem-solving, such as Itinerary Planning, Trajectory Prediction, Navigation, and Personalized POI Recommendation. Each task's question example can be found in Appendix D.

Distance *(DS)* Asking model to compute the great-circle (Haversine) or Euclidean distance between two sets of POI with coordinates provided in the context. Accurate distance calculation is a fundamental quantitative reasoning ability that underpins more complex tasks like proximity-based recommendations and path planning.

Direction *(DR)* Evaluates the ability to perform trigonometric calculations and map the resulting numerical bearing (azimuth) to a descriptive cardinal direction (e.g., Northeast). This is crucial for understanding spatial layouts and generating human-readable navigation instructions.

Relation *(RE)* Determine the spatial topology relation among entities, instantiated as three subtypes: road–road connectivity (intersecting/parallel/disconnected), road–area interaction (pass through/touch/lie outside), and area–area relation (overlapping/adjacent/separate). Beyond simple geometry, the ability to correctly perceive the underlying spatial structure forms the structural foundation upon which more complex, semantic, and goal-oriented reasoning can be performed.

Planning *(PN)* Simulating a complex, NP-hard optimization problem, requires the model to determine an optimal visiting order for a set of locations to minimize total travel distance. Given a starting point and multiple unordered destinations with their pairwise spatial relationships (distances and bearings), this tests the model's multi-hop reasoning, sequential optimization, and use of heuristic strategies.

Prediction *(PR)* Evaluates probabilistic geospatial inference. Given a partial movement trajectory and a set of candidate destinations, the model must infer the most likely final destination, requiring it to capture movement patterns and understand user intent within spatial constraints.

Navigation *(NV)* Assesses the ability to understand road network topology and translate a planned route into actionable instructions. The model must first identify a shortest path between start and goal on a simplified road network and then verbalize this path as a sequence of clear, turn-by-turn directions a human can follow.

Recommendation *(RC)* Tests the model's ability to recommend one or more POIs satisfying a complex combination of spatial constraints (e.g., within a 3-mile radius), semantic preferences (e.g., cuisine type, user rating), and contextual filters (e.g., currently open), requiring it to integrate and weigh diverse geo-entity information to make a final decision.

3.2.3 Human Annotation and Quality Control

We first had 15 trained annotators collect structured geospatial scenarios from various global cities on publicly available information. These scenarios were then processed by an automated pipeline to generate natural language question-answer pairs. The distribution of these questions across the seven task categories is shown in Table 3. For quality control, we apply dual-pass reviews: an initial validator verifies geometric correctness and constraint sufficiency, a second validator checks linguistic clarity and ambiguity. Instances failing determinism (e.g., multiple equally valid answers without tie-breaking rules) are revised or discarded.

Task	DS	DR	RE	PN	PR	NV	RC
Count	498	497	637	497	494	392	133

Table 3: Distribution of the 3,148 questions across the seven tasks in the Reasoning Module.

4 Evaluation

For the **Knowledge Module**, we use *Accuracy* on its multiple-choice questions to measure a model's ability to recall static geospatial facts when challenged by our strategically designed distractors.

Evaluating the **Reasoning Module** is more complex, as a correct answer can hide a flawed logical process. Our framework therefore assesses the entire reasoning chain. To achieve this, we first enforce a guided output format, instructing models to separate their step-by-step logic (`<reason></reason>`) from their final answer (`<answer></answer>`). Following MT-Bench-101 (Bai et al., 2024), we then employ GPT-4o (OpenAI

Model	Node	Area	Road	Type	Function	Surroundings	Overall
GPT-4o	41.59	69.99	59.21	71.87	51.97	26.33	54.49
Llama-3.1-70B-Instruct	31.43	58.43	39.98	77.42	49.02	24.05	46.72
Llama-3.1-8B-Instruct	22.22	44.19	28.13	66.95	27.47	18.44	34.57
Qwen2.5-72B-Instruct	31.40	62.16	39.80	79.17	54.56	26.33	48.90
Qwen2.5-7B-Instruct	24.34	48.10	29.37	76.88	50.74	25.94	42.56
Average	**30.20**	**56.57**	**39.30**	**74.46**	**46.75**	**24.22**	**45.45**

Table 4: Overall accuracy (%) on the six dimensions across 9 cities of the Knowledge Module.

	Fundamental Reasoning						**Advanced Reasoning**							
	DS		DR		RE		PN		PR		NV		RC	
Model	**R**	**A**	**R**	**A**	**R**	**A**	**R**	**A**	**R**	**A**	**R**	**A**	**R**	**A**
GPT-4o	13.71	8.55	11.68	1.80	13.96	6.84	10.82	4.69	18.81	6.64	16.43	7.63	19.09	6.66
o3-mini	14.66	9.90	13.80	8.94	13.95	7.27	18.88	9.70	17.35	4.54	17.41	8.55	18.65	7.12
Qwen3-32B	14.27	9.78	12.51	7.74	14.05	6.80	17.90	9.57	17.91	3.59	16.63	8.05	19.04	6.74
QwQ-32B	14.32	9.67	12.97	8.13	13.97	6.48	18.68	9.70	17.88	4.24	14.05	6.15	17.71	6.82
DeepSeek-R1	14.65	9.75	12.77	7.52	14.15	7.16	13.08	6.75	18.41	6.07	16.35	7.89	18.26	6.81
DeepSeek-V3	13.53	9.39	11.39	3.83	14.13	7.25	10.41	5.10	18.23	5.20	17.78	8.77	19.27	7.38
Qwen3-8B	13.53	9.19	11.43	7.09	13.77	6.00	13.86	7.69	16.56	3.00	15.28	6.93	18.68	6.40
DeepSeek-R1-32B	12.33	8.43	9.87	4.46	13.12	6.30	10.66	6.27	14.73	3.11	14.03	6.66	18.17	6.71
Llama3.1-70B-Ins.	11.13	7.98	9.01	2.85	12.92	4.84	9.85	5.18	16.98	3.58	14.45	6.48	18.92	7.24
Llama3.1-8B-Ins.	8.42	6.16	7.92	0.74	12.00	3.30	7.73	4.12	16.60	2.67	12.82	4.06	18.19	6.15
Qwen2.5-72B-Ins.	11.89	7.18	10.85	2.29	13.86	6.34	11.43	5.20	18.05	4.05	15.78	7.08	19.01	7.41
Qwen2.5-7B-Ins.	11.66	5.14	9.01	1.49	13.14	4.44	10.29	5.08	16.68	2.47	11.23	3.01	17.97	6.20
Gemma-3-27B	13.90	7.67	10.28	1.73	13.33	4.40	9.88	4.93	18.30	4.33	16.38	7.31	18.07	6.44
Phi-4	13.58	7.97	11.50	3.56	13.76	6.06	11.39	5.18	17.63	3.67	14.89	6.88	18.44	6.26
DeepSeek-R1-7B	11.42	7.83	7.09	1.80	12.08	4.84	5.70	3.89	14.03	2.29	8.45	2.24	15.28	5.10
DeepSeek-R1-1.5B	8.96	4.44	5.49	1.14	9.73	3.91	6.78	3.47	10.82	2.05	7.45	0.94	11.26	3.32
Task Avg.	**12.62**	**8.06**	**10.47**	**4.07**	**13.25**	**5.76**	**11.71**	**6.03**	**16.81**	**3.84**	**14.34**	**6.16**	**17.88**	**6.42**

Table 5: Detailed evaluation results for the Reasoning Module. For each task, we report the average score for the Reasoning process (R) and the final Answer (A).

et al., 2024) as an automated evaluator. It uses detailed, task-specific **Scoring Rubrics** with carefully crafted checklists (see Appendix B) to perform a fine-grained assessment of both components. This process yields a *Reason Score* (S_{reason}) for the logic and an *Answer Score* (S_{answer}) for the outcome. To enable fair cross-task comparisons, we normalize the total score ($S_{\text{total}} = S_{\text{reason}} + S_{\text{answer}}$) to a 0-100% scale.

5 Experiments

We conduct a comprehensive evaluation on UrbanGeoEval to assess the urban geospatial capabilities of different LLMs. Our analysis is structured to first present the overall performance landscape and then delve into a fine-grained analysis of model strengths and weaknesses across both knowledge and reasoning dimensions.

5.1 Experimental Setup

We evaluated a diverse set of 16 popular LLMs, including both instruction-tuned and reasoning-specialized models. For each LLM, we apply the corresponding chat format while setting the temperature and sampling parameter as their official configures. Additional details on the full model list, experimental setup and implementation can be found in Appendix A.

5.2 Main Results

The evaluation results on the UrbanGeoEval reveals a clear hierarchy of capabilities and yields critical insights into the determinants of geospatial intelligence. The results demonstrate a significant divergence between performance on factual knowledge recall and complex procedural reasoning.

Knowledge Module Performance On the Knowledge Module, which measures factual recall, GPT-4o achieves the highest average accu-

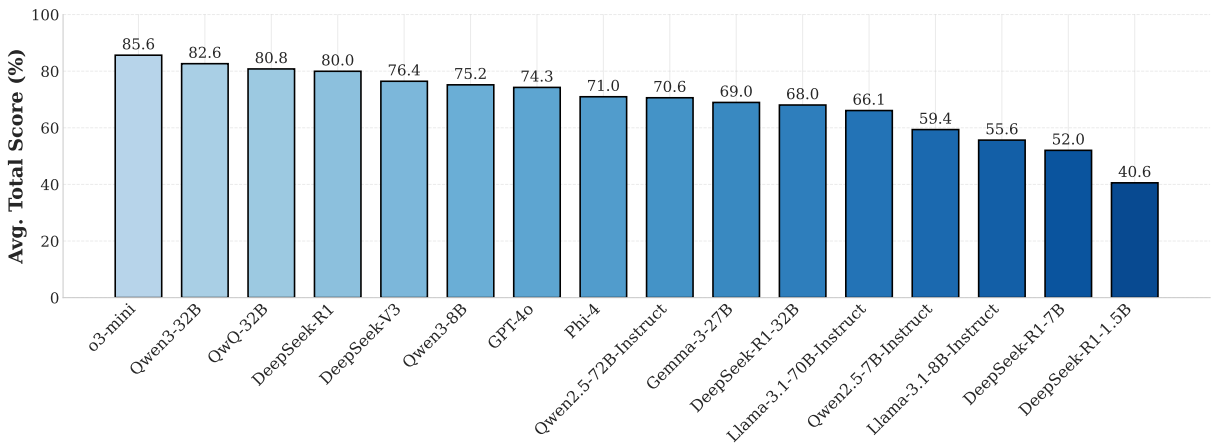

Figure 3: Overall performance of all 16 evaluated models on the Reasoning Module.

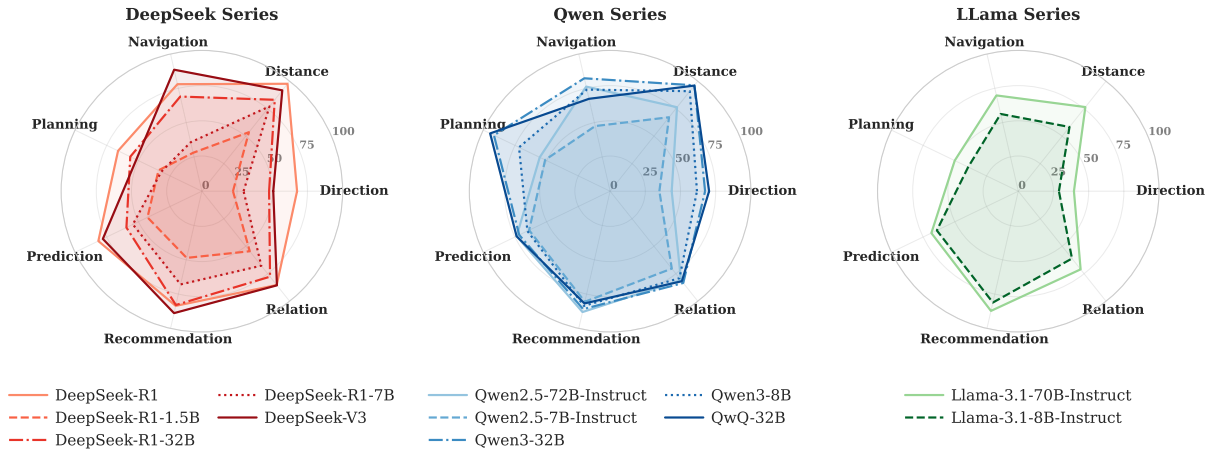

Figure 4: Performance breakdown across seven reasoning tasks for the DeepSeek, Qwen, and Llama model series.

racy (54.49%), establishing it as the most knowledgeable model evaluated, as shown in Table 4. This superior performance is likely attributable to its massive parameter scale and extensive, high-quality training data. A consistent trend across all model families is that larger models outperform their smaller counterparts, confirming that model scale is positively correlated with geospatial knowledge capacity. However, these gains exhibit diminishing returns, as a nearly tenfold increase in model scale (e.g., from Qwen2.5-7B to 72B) yields only a modest 6-point accuracy improvement (from 42.56% to 48.90%). Among open-source models, the Qwen series consistently surpasses the Llama series in accuracy. A plausible explanation for this performance gap could be a greater diversity and richness in the geospatial data within the Qwen series' pre-training corpus, potentially offering more comprehensive global coverage.

Reasoning Module Performance The results from the Reasoning Module, presented in Figure 3 and Table 5, are even more striking. A key finding is that the top-performing models are overwhelmingly those specifically optimized for reasoning, such as o3-mini, Qwen3-32B, QwQ-32B, and DeepSeek-R1, with 32B reasoning models significantly outperforming even much larger instruct-tuned models like GPT-4o (e.g., Qwen3-32B at 82.6% vs. GPT-4o at 74.3%). This result strongly suggests that complex, multi-step geospatial problem-solving—such as itinerary planning and constrained navigation—benefits more from sophisticated training paradigms like Reinforcement Learning Fine-Tuning (RFT) than from scale alone.

5.3 In-depth Analysis

Geospatial Biases and Granularity Bottlenecks A fine-grained breakdown of models' performance by city on the Knowledge Module (Figure 5) reveals a consistent geospatial bias. While GPT-4o's average accuracy is the highest, it exhibits a larger

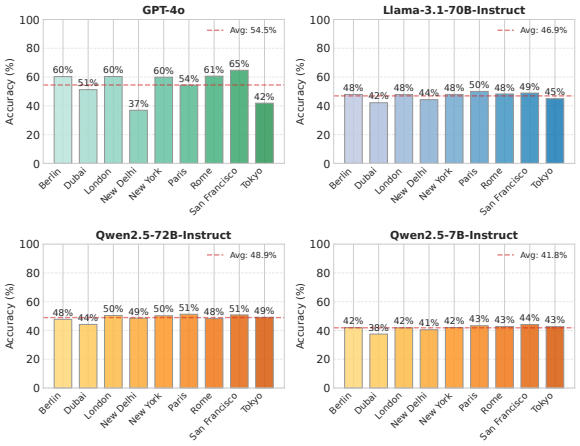

Figure 5: Accuracy breakdown by city for five models on the Knowledge Module.

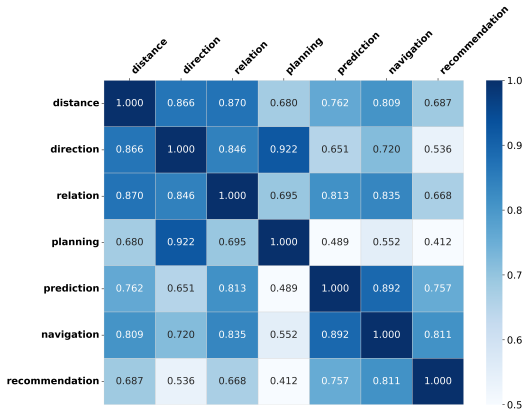

Figure 6: Pearson correlation matrix of model average total scores across the seven reasoning tasks.

performance variance across cities, with notable accuracy drops in Asian cities such as New Delhi and Tokyo, likely reflecting its geographic training corpus may be less globally balanced than that of some open-source models like Qwen, leading to a deep but uneven distribution of world knowledge.

Furthermore, Table 4 shows even top models struggle with tasks requiring precise quantitative recall (e.g., *Node*, *Road*), but excel at tasks reliant on semantic cues (e.g., *Type*). This indicates LLMs operate as probabilistic pattern matchers rather than symbolic databases, making them inherently less reliable for recalling fine-grained facts like address or road lengths. See Appendix C for more detailed results.

The Primacy of the Reasoning Process Our process-aware evaluation framework confirms that a high-quality reasoning chain is a strong prerequisite for a correct final answer. Models with high S_{total} scores, like o3-mini, consistently produced logical, step-by-step thinking. In contrast, lower-scoring models frequently exhibited flawed logic, hallucinations, or calculation errors in their reasoning, even if they occasionally arrived at the correct answer by chance. Detailed case studies can be found in Appendix D, to show that GPT-4o evaluation aligns well with the corresponding abilities and scoring criteria.

Brittle and Non-Uniform Reasoning As illustrated in Figure 4, model reasoning capabilities are highly task-dependent. Most models found structured tasks like *Recommendation* and *Distance* relatively easy. Conversely, the majority struggled with *Planning* (an NP-hard problem) and *Direction* (requiring precise trigonometric calculations), revealing that current LLMs' geospatial reasoning are often heuristic shortcuts learned from textual patterns rather than robust, generalizable computational competence. Their capabilities fracture when faced with problems that demand algorithmic complexity or high numerical precision.

Foundational Skills as Building Blocks A key hypothesis of our benchmark design is that elementary spatial abilities are foundational to more complex ones. We validated this by computing the Pearson correlation matrix of model scores across the seven reasoning tasks (Figure 6). The heatmap reveals a strong positive correlation (mostly > 0.6) between performance on foundational tasks (e.g., *Distance*, *Relation*) and advanced composite tasks (e.g., *Planning*, *Navigation*).

This empirically validates our design, suggesting that a curriculum learning approach—mastering foundational skills before tackling complex problems may be a more effective path toward developing sophisticated and reliable geo-aware models.

6 Conclusion

This paper introduces a benchmark that decouples urban geospatial knowledge from reasoning with a comprehensive task taxonomy, revealing that superior reasoning is unlocked by specialized training over sheer model scale. Our framework diagnoses widespread challenges in current LLMs, including geospatial biases, knowledge granularity bottlenecks and brittle performance on complex reasoning tasks. Furthermore, extensive case studies indicate that tasks in our benchmark effectively measure the urban geospatial abilities.

Limitations

The current benchmark is text-only, presenting opportunities to incorporate real-time multimodal inputs, and to expand the geographic scope to a more diverse range of city environments. Furthermore, our scalable evaluation framework, which relies on OpenStreetMap and an LLM-as-a-Judge, could be further enhanced by integrating varied data sources and complementing it with human validation studies. We intend to regularly update our benchmark, from UrbanGeoEval to future iterations, to incorporate new developments.

Ethical Considerations

We have conducted our work with a strong commitment to ethical and responsible research practices. The data for our Knowledge Module is sourced from OpenStreetMap under ODbL license, a public, open-source dataset aiming at building and maintaining a free, editable map of the world. For our Reasoning Module, the foundational data was manually collected and annotated by human workers based on publicly available information from sources such as Google Maps. All annotators were recruited through a professional data annotation platform and were compensated fairly for their labor. The annotation team was composed of a diverse group of well-educated individuals of both genders, aged 18-30. Across both modules, our data collection process did not involve any private or personally identifiable information. All benchmark questions are synthetically generated for the scientific purpose of evaluating AI models and are free of any offensive or sensitive content. A key ethical contribution of this work is the explicit identification of geospatial biases in current LLMs, which we believe is a crucial first step toward mitigation. UrbanGeoEval is intended solely for research use and will be publicly released to foster transparency and responsible innovation.

References

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand. Association for Computational Linguistics.

Pasquale Balsebre, Weiming Huang, Gao Cong, and Yi Li. 2024. City foundation models for learning general purpose representations from openstreetmap. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 87–97, New York, NY, USA. Association for Computing Machinery.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. 2022. Mapqa: A dataset for question answering on choropleth maps. *Preprint*, arXiv:2211.08545.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Mahir Labib Dihan, MD Tanvir Hassan, MD TANVIR PARVEZ, Md Hasebul Hasan, Md Almash Alam, Muhammad Aamir Cheema, Mohammed Eunus Ali, and Md Rizwan Parvez. 2025. Mapeval: A map-based evaluation of geo-spatial reasoning in foundation models. In *Forty-second International Conference on Machine Learning*.

Ruixue Ding, Boli Chen, Pengjun Xie, Fei Huang, Xin Li, Qiang Zhang, and Yao Xu. 2023. Mgeo: Multimodal geographic language model pre-training. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 185–194, New York, NY, USA. Association for Computing Machinery.

Jie Feng, Tianhui Liu, Yuwei Du, Siqi Guo, Yuming Lin, and Yong Li. 2025. Citygpt: Empowering urban spatial cognition of large language models. *Preprint*, arXiv:2406.13948.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Devashish Vikas Gupta, Azeez Syed Ali Ishaqui, and Divya Kiran Kadiyala. 2024. Geode: A zero-shot geospatial question-answering agent with explicit reasoning and precise spatio-temporal retrieval. *Preprint*, arXiv:2407.11014.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Jizhou Huang, Haifeng Wang, Yibo Sun, Yunsheng Shi, Zhengjie Huang, An Zhuo, and Shikun Feng. 2022. Ernie-geol: A geography-and-language pre-trained model and its applications in baidu maps. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3029–3039, New York, NY, USA. Association for Computing Machinery.

Zixian Huang, Yulin Shen, Xiao Li, Yuang Wei, Gong Cheng, Lin Zhou, Xinyu Dai, and Yuzhong Qu. 2019. GeoSQA: A benchmark for scenario-based question answering in the geography domain at high school level. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5866–5871, Hong Kong, China. Association for Computational Linguistics.

Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, and Yongfeng Zhang. 2025. Disentangling memory and reasoning ability in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1681–1701, Vienna, Austria. Association for Computational Linguistics.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Haonan Li, Martin Tomko, and Timothy Baldwin. 2023. Location aware modular biencoder for tourism question answering. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 95–109, Nusa Dua, Bali. Association for Computational Linguistics.

Wenbin Li, Di Yao, Ruibo Zhao, Wenjie Chen, Zijie Xu, Chengxue Luo, Chang Gong, Quanliang Jing, Haining Tan, and Jingping Bi. 2024. Stbench: Assessing the ability of large language models in spatio-temporal analysis. *Preprint*, arXiv:2406.19065.

Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, Chris Cundy, Ziyuan Li, Rui Zhu, and Ni Lao. 2024. On the opportunities and challenges of foundation models for geoai (vision paper). *ACM Trans. Spatial Algorithms Syst.*, 10(2).

Gengchen Mai, Krzysztof Janowicz, Rui Zhu, Ling Cai, and N. Lao. 2021. Geographic question answering: Challenges, uniqueness, classification, and future directions. In *AGILE Conf.*

Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David B. Lobell, and Stefano Ermon. 2024. Geollm: Extracting geospatial knowledge from large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Srija Mukhopadhyay, Abhishek Rajgaria, Prerana Khatiwada, Manish Shrivastava, Dan Roth, and Vivek Gupta. 2025. MAPWise: Evaluating vision-language models for advanced map queries. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9348–9378, Albuquerque, New Mexico. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Openai o3-mini system card. Technical report, OpenAI. Accessed: 2025-10-05.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof qa benchmark. *Preprint*, arXiv:2311.12022.

Jonathan Roberts, Timo Luddecke, Sowmen Das, Kai Han, and Samuel Albanie. 2023. Gpt4geo: How a language model sees the world's geography. *ArXiv*, abs/2306.00020.

Varun Srivastava, Fan Lei, Srija Mukhopadhyay, Vivek Gupta, and Ross Maciejewski. 2025. Mapiq: Benchmarking multimodal large language models for map question answering. *Preprint*, arXiv:2507.11625.

Zhecheng Wang, Haoyuan Li, and Ram Rajagopal. 2020. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1013–1020.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Mutian Yang, Jiandong Gao, and Ji Wu. 2025b. Decoupling knowledge and reasoning in llms: An exploration using cognitive dual-system theory. *Preprint*, arXiv:2507.18178.

Dazhou Yu, Riyang Bao, Ruiyu Ning, Jinghong Peng, Gengchen Mai, and Liang Zhao. 2025. Spatial-rag: Spatial retrieval augmented generation for real-world geospatial reasoning questions. *Preprint*, arXiv:2502.18470.

Cong Zhang, Xin Deik Goh, Dexun Li, Hao Zhang, and Yong Liu. 2025. Planning with multi-constraints via collaborative language agents. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10054–10082, Abu Dhabi, UAE. Association for Computational Linguistics.

A Details on Experimental Setup

A.1 Evaluated Models

Our evaluation includes closed-source models (GPT-4o OpenAI et al., 2024, o3-mini OpenAI, 2025) and a diverse range of open-source models. The open-source set includes instruct-tuned models from the Llama3 (Grattafiori et al., 2024) and Qwen2.5 (Qwen et al., 2025) series, as well as reasoning-tuned models from the DeepSeek-R1[4] (DeepSeek-AI et al., 2025) and Qwen3 (Yang et al., 2025a) series. All details about the evaluated models are presented in Table 6.

A.2 Implementation Details

We accessed closed-source models via their official APIs. Open-source models were deployed locally using the vLLM framework (Kwon et al., 2023) on a server equipped with 8 NVIDIA A100 GPUs. Following best practices, we set the temperature to 0.3 for the fact-based Knowledge Module to encourage deterministic outputs. For the generative Reasoning Module, we used model-specific recommended configurations for the models under evaluation and a temperature of 0.6 for our LLM-as-a-Judge, GPT-4o.

B Evaluation of Reasoning Module Tasks

This section provides details on the prompts used to guide model outputs and the scoring rubrics used by the Judge-LLM for evaluating the geospatial reasoning tasks.

B.1 Guided Output Prompt

We use a two-part prompt structure. An initial system prompt defines the LLM's role and output format (Figure 7), followed by a task-specific prompt containing the context and question.

B.2 Judge-LLM Prompt and Scoring Rubrics

The GPT-4o judge is provided with the original question, the model's response, the ground truth answer, and a detailed scoring rubric. The initial Judge-LLM prompt is shown in Figure 8. The detailed scoring rubrics for each of the seven reasoning tasks are shown in Figure 9 to 17.

[4]Smaller models in the DeepSeek-R1 series are knowledge-distilled from Qwen2.5 teacher models, which is reflected in their full names (e.g., `DeepSeek-R1-Distill-Qwen-1.5B`).

C Details on Evaluation Results

This appendix provides a more granular analysis of the evaluation results presented in the main paper, as shown in Figures 18 to 32 to offer deeper insights into their specific strengths and weaknesses in urban geospatial understanding.

C.1 Detailed Performance on the Knowledge Module

C.1.1 Cross-Model Performance on San Francisco

Figures 18 through 21 illustrate the performance of four different models (GPT-4o, Llama-3.1-70B-Instruct, Llama-3.1-8B-Instruct, and Qwen2.5-72B-Instruct) on the 20 question templates for the city of San Francisco. This comparative analysis reveals several consistent patterns:

Across all models, there is a clear performance gap between tasks requiring semantic understanding and those demanding precise quantitative recall. Models consistently achieve high accuracy on tasks like `Classify_POI` and `Find_Instances_By_Type`, which rely on recognizing and categorizing entities. Conversely, all models struggle significantly with tasks that require fine-grained, precise data retrieval. Performance is notably low on templates such as `Retrieve_Coordinates`, `Query_Length`, and `Find_Nearest_POI`.

C.1.2 Cross-City Performance

Figures 22 through 30 show the performance of a single model, Qwen2.5-7B-Instruct, across all nine cities in the benchmark. This analysis highlights the presence of geospatial bias in the model's knowledge base.

C.2 Detailed Performance on the Reasoning Module

The Reasoning Module evaluates the ability to perform multi-step logical inference on geospatial data provided in-context. The following analysis delves into the overall reasoning landscape and the computational efficiency of the models.

C.2.1 Overall Reasoning Capability Landscape

Figure 32 provides a consolidated radar chart of the total scores for all 16 evaluated models across the seven reasoning tasks.

Provider/Series	Model Name	Model Link
OpenAI	GPT-4o (2024-11-20)	https://platform.openai.com/docs/models/gpt-4o
	o3-mini	https://platform.openai.com/docs/models/o3-mini
DeepSeek	DeepSeek-V3	https://huggingface.co/deepseek-ai/DeepSeek-V3
	DeepSeek-R1	https://huggingface.co/deepseek-ai/DeepSeek-R1
	DeepSeek-R1-1.5B	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B
	DeepSeek-R1-7B	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
	DeepSeek-R1-32B	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B
Llama	Llama-3.1-8B-Instruct	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
	Llama-3.1-70B-Instruct	https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct
Qwen	QwQ-32B	https://huggingface.co/Qwen/QwQ-32B
	Qwen2.5-7B-Instruct	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
	Qwen2.5-72B-Instruct	https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
	Qwen3-8B	https://huggingface.co/Qwen/Qwen3-8B
	Qwen3-32B	https://huggingface.co/Qwen/Qwen3-32B
Google	Gemma-3-27B	https://huggingface.co/google/Gemma-3-27B-it
Microsoft	Phi-4	https://huggingface.co/microsoft/Phi-4

Table 6: Detailed information of the evaluated models.

C.2.2 Inference Efficiency and Performance Trade-offs

Figure 31 presents the average inference time per query for each model on the Reasoning Module. The results reveal a critical trade-off between reasoning performance and computational efficiency.

D Case Study

Figure 33 to 42 show cases corresponding to each task, each reflecting the classical error of the model response. These cases show that our task design can accurately assess the corresponding ability of LLMs.

You are an expert urban geospatial reasoning assistant. Your task is to answer questions with adherence to the following structure and rules.

Your response MUST consist of ONLY two sections formatted as follows:

Section 1: Reasoning Process
Provide a detailed, step-by-step explanation of your reasoning process enclosed within <reason> and </reason> tags. This section should capture all relevant considerations and intermediate steps, ensure that your reasoning is comprehensive yet focused to the question.

[Steps to follow]

Section 2: Final Answer
Present your concise and definitive final answer enclosed within <answer> and </answer> tags.
Format strictly as: [Formatted final answer]

Figure 7: Initial system prompt and output format instruction template for guiding evaluated LLMs in geospatial reasoning.

You are an impartial and meticulous judge. Your task is to evaluate a model's response to a question. The response is divided into two parts: <reason> (the model's reasoning process) and <answer> (the model's final answer). You will be provided with:

1. The original question
2. The model's response (<reason>...</reason><answer>...</answer>)
3. A checklist that defines the scoring criteria
4. A ground truth final answer for reference in evaluating <answer>

Instructions:
1. Evaluate the <reason> section using the checklist:
2. Analyze how well the reasoning aligns with each criterion in the checklist.
3. Refer to specific points from the checklist in your explanation.
4. Note strengths, weaknesses, and any missing or incorrect reasoning steps.
5. Evaluate the <answer> section by comparing it to the ground truth.
6. Be objective and avoid personal bias.
7. Give your clear explanations, finally end your evaluation with the scores in this exact format:

reason:[[score]], answer:[[score]], total:[[score]]

Figure 8: The initial scoring prompt for Judge-LLM on evaluation.

Scoring Rubric – Total: 25 Points
Problem Type: Basic Straight-Line Distance Calculation
Objective: Given two geographic coordinates, calculate the straight-line (great-circle) distance between them in kilometers, showing a step-by-step explanation.

Part 1: Reasoning Process (<reason>) – Max 15 Points
Coordinate Extraction (2 points)
2: Clearly and correctly states the latitude and longitude of both locations in decimal degrees.
1: Minor errors in conversion from degrees-minutes-seconds (DMS) to decimal degrees.
0: Coordinates are missing, incorrect, or not in decimal format.
Conversion to Radians (3 points)
3: Correctly converts all decimal degree values to radians.
1~2: Minor errors in conversion or no calculaton results given(e.g., wrong precision or wrong formula for one coordinate).
0: Conversion is missing or significantly incorrect.
Difference Calculation (3 points)
3: Accurately calculates the differences in latitude ($\Delta\varphi$) and longitude ($\Delta\lambda$) in radians.
1~2: Minor mistake in one of the difference values or no calculaton results given.
0: Differences are missing or clearly incorrect.
Haversine Formula Application (5 points)
5: Correctly applies the Haversine formula to compute intermediate values a and c.
2~3: Minor computation error in a or c or no calculaton results given.
0: Formula not applied or completely incorrect.
Clarity & Structure (2 points)
2: The response is clearly separated into <reason> and <answer> sections and follows a logical structure.
1: Slight structural confusion or disorganization.
0: No clear separation or poorly structured response.

Part 2: Final Answer Accuracy (<answer>) – Max 10 Points
The final answer will be compared against the ground truth distance.
Scoring Rule: Absolute Error Penalty
Formula: Score = max(0, 10 - |model_answer - ground_truth|), round the score to two decimal places.

Important Notes:
The maximum total score is 25 points:
- Up to 15 points from reasoning (<reason>)
- Up to 10 points from final answer accuracy (<answer>)
The minimum total score is 0 points.

Figure 9: The unique scoring prompt for the Distance task.

Scoring Rubric – Total: 25 Points
Problem Type: Direction Calculation (Bearing between two points)
Objective: Given two geographic coordinates (latitude and longitude of two points of interest), calculate the bearing (angle) from one point to the other, measured clockwise from true north. Additionally, provide the direction in descriptive terms (e.g., Northeast, East-Southeast).

Part 1: Reasoning Process (<reason>) – Max 15 Points
Coordinate Extraction (2 points)
2: Clearly and correctly states the latitude and longitude of both locations in decimal degrees.
1: Minor errors in conversion from degrees-minutes-seconds (DMS) to decimal degrees.
0: Coordinates are missing, incorrect, or not in decimal format.
Conversion to Radians (3 points)
3: Correctly converts all decimal degree values to radians.
1~2: Minor errors in conversion or no calculation results given (e.g., wrong precision or wrong formula for one coordinate).
0: Conversion is missing or significantly incorrect.
Difference Calculation (3 points)
3: Accurately calculates the differences in longitude (Δlong) in radians.
1~2: Minor mistake in the difference value or no calculation results given.
0: Difference is missing or clearly incorrect.
Bearing Formula Application (5 points)
5: Correctly applies the bearing formula to compute intermediate values.
2~3: Minor computation error in the bearing formula or no calculation results given.
0: Formula not applied or completely incorrect.
Clarity & Structure (2 points)
2: The response is clearly separated into and sections and follows a logical structure.
1: Slight structural confusion or disorganization.
0: No clear separation or poorly structured response.

Part 2: Final Answer Accuracy (<answer>) – Max 10 Points
The final answer will be compared against the ground truth bearing and cardinal direction.
Scoring Rule: Absolute Error Penalty
Formula: Score = max(0, 10 - (|model answer - ground truth bearing|) * 0.25), round the score to two decimal places.
If cardinal direction in model answer does not match model answer bearing, Score -2 points, else +0 points.

Important Notes:
The maximum total score is 25 points:
- Up to 15 points from reasoning (<reason>)
- Up to 10 points from final answer accuracy (<answer>)
The minimum total score is 0 points.

Figure 10: The unique scoring prompt for the Direction task.

Scoring Rubric – Total: 25 Points
Problem Type: Road-to-Road Spatial Relationship
Objective: Given two roads with sampled coordinates, determine their spatial relationship (Intersecting, Parallel, or Disconnected).

Part 1: Reasoning Process (<reason>) – Max 15 Points
Coordinate Extraction (1 points)
1: Correctly states the coordinates for both roads.
0: Coordinates are missing or incorrect.
Intersection/Parallel/Disconnected Criteria (3 points)
3: Correctly identifies the criteria for determining spatial relations (intersection, parallelism, or disconnection).
1~2: Minor misunderstanding of the criteria or misinterpretation of road relationships.
0: Criteria are applied incorrectly or not at all.
Analysis of Spatial Relationship (8 points)
8: Correctly analyzes the roads' relationship by considering sampled coordinates, using geometric or mathematical methods such as line intersection analysis, or by comparing directional patterns and slopes of the segments.
3~7: Minor mistake recognizing road orientations or relation or no computational/spatial reasoning steps are shown.
0~2: Incorrect analysis or application of incorrect spatial relationships.
Clarity & Structure (3 points)
3: The response is clearly separated into <reason> and <answer> sections, logically structured, and easy to follow.
1~2: Slight structural confusion or lack of clarity.
0: No clear separation or poorly structured response.

Part 2: Final Answer Accuracy (<answer>) – Max 10 Points
The final answer will be compared against the ground truth spatial relationship (Intersecting, Parallel, or Disconnected).
Scoring Rule:
Score = 10 if answer is correct else 0.

Important Notes:
The maximum total score is 25 points:
- Up to 15 points from reasoning (<reason>)
- Up to 10 points from final answer accuracy (<answer>)
The minimum total score is 0 points.

Figure 11: The unique scoring prompt for the Relation task (road-road).

Scoring Rubric – Total: 25 Points
Problem Type: Road-to-Area Spatial Relationship
Objective: Given a road and an area with coordinates, determine the spatial relationship between them (Pass through the area, Touch the area, or Lie outside the area).

Part 1: Reasoning Process (<reason>) – Max 15 Points
Coordinate Extraction (1 points)
1: Correctly states the coordinates for both the road and area.
0: Coordinates are missing or incorrect.
Boundary/Path Analysis (3 points)
3: Correctly identifies the criteria for the spatial relationship based on the position of the road and area (intersection, touching boundary, or lying outside).
1~2: Minor misunderstanding of the criteria or misinterpretation of road and area interactions.
0: Criteria are applied incorrectly or not at all.
Analysis of Spatial Relationship (8 points)
8: Correctly analyzes the road's relationship to the area by considering sampled coordinates, using geometric or mathematical methods such as point-in-polygon or line-to-polygon tests.
3~7: Minor mistakes in determining the relationship or no computational/spatial reasoning steps are shown.
0~2: Incorrect analysis or application of incorrect spatial relationships.
Clarity & Structure (3 points)
3: The response is clearly separated into <reason> and <answer> sections, logically structured, and easy to follow.
1~2: Slight structural confusion or lack of clarity.
0: No clear separation or poorly structured response.

Part 2: Final Answer Accuracy (<answer>) – Max 10 Points
The final answer will be compared against the ground truth spatial relationship (Pass through the area, Touch the area, or Lie outside the area).
Scoring Rule:
Score = 10 if answer is correct else 0.

Important Notes:
The maximum total score is 25 points:
- Up to 15 points from reasoning (<reason>)
- Up to 10 points from final answer accuracy (<answer>)
The minimum total score is 0 points.

Figure 12: The unique scoring prompt for the Relation task (road-area).

Scoring Rubric – Total: 25 Points
Problem Type: Area-to-Area Spatial Relationship
Objective: Given two areas with boundary coordinates, determine the spatial relationship between them (Overlapping, Adjacent, or Separate).

Part 1: Reasoning Process (<reason>) – Max 15 Points
Coordinate Extraction (1 points)
1: Correctly states the coordinates for both areas.
0: Coordinates are missing or incorrect.
Overlap/Adjacency/Separation Criteria (3 points)
3: Correctly identifies the criteria for determining spatial relationships (overlap, adjacency, or separation).
1~2: Minor misunderstanding of the criteria or misinterpretation of area interactions.
0: Criteria are applied incorrectly or not at all.
Analysis of Spatial Relationship (8 points)
8: Correctly analyzes the areas' relationship by considering boundary coordinates and applying geometric or mathematical methods (e.g., intersection, adjacency check, or area overlap tests).
3~7: Minor mistake in determining the relationship or no computational/spatial reasoning steps are shown.
0~2: Incorrect analysis or application of incorrect spatial relationships.
Clarity & Structure (3 points)
3: The response is clearly separated into <reason> and <answer> sections, logically structured, and easy to follow.
1~2: Slight structural confusion or lack of clarity.
0: No clear separation or poorly structured response.

Part 2: Final Answer Accuracy (<answer>) – Max 10 Points
The final answer will be compared against the ground truth spatial relationship (Overlapping, Adjacent, or Separate).
Scoring Rule:
Score = 10 if answer is correct else 0.

Important Notes:
The maximum total score is 25 points:
- Up to 15 points from reasoning (<reason>)
- Up to 10 points from final answer accuracy (<answer>)
The minimum total score is 0 points.

Figure 13: The unique scoring prompt for the Relation task (area-area).

Scoring Rubric – Total: 30 Points
Problem Type: Itinerary Planning
Objective: Given a set of locations and their relative position description, determine the optimal visiting order to minimize the total straight-line travel distance.

Part 1: Reasoning Process (<reason>) – Max 20 Points
Coordinate Extraction (2 points)
2: Correctly extracts the distances and bearings between consecutive locations and organizes them logically.
1: Minor errors in extracting data or organizing information.
0: Missing or incorrect data extraction.
Farthest Location Identification (3 points)
3: Correctly identifies the farthest location based on direction and distances and logically concludes that this location should be visited last.
0: Incorrect or missing identification of the farthest location.
Coordinate Conversion and Calculation (5 points)
5: Correctly converts bearings and distances to Cartesian coordinates for all locations.
2~3: Minor mistakes in the conversion or failure to account for the correct bearing angle or distance.
0~1: Incorrect or missing conversion of bearings to Cartesian coordinates.
Path Evaluation (8 points)
8: Correctly consider the reduced search space (e.g., only permutations of locations excluding the farthest one) and correctly identifies and calculates the total distance for the optimal visiting order.
4~6: Minor mistakes in evaluating paths, or considering unnecessary permutations.
0~3: Incorrect or incomplete evaluation of paths.
Clarity & Structure (2 points)
2: The response is clearly separated into <reason> and <answer> sections and logically structured.
1: Slight structural confusion or lack of clarity.
0: No clear separation or poorly structured response.

Part 2: Final Answer Accuracy (<answer>) – Max 10 Points
The final answer will be compared against the ground truth optimal visiting order.
Scoring Rule:
The model response is compared to the ground-truth order, with a penalty applied for each incorrect location in the order. If the correct location is not in the correct position, it receives a penalty of 2 points per misplaced location.

Important Notes:
The maximum total score is 30 points:
- Up to 20 points from reasoning (<reason>)
- Up to 10 points from final answer accuracy (<answer>)
The minimum total score is 0 points.

Figure 14: The unique scoring prompt for the Planning task.

Scoring Rubric -- Total: 30 Points
Problem Type: Prediction
Objective: Evaluate an LLM's geospatial reasoning ability by assessing how well it can predict the most likely final destination based on partial trajectory.

Part 1: Reasoning Process (<reason>) -- Max 20 Points
  Trajectory Analysis (5 points)
    5: Correctly maps the trajectory points and identifies the overall direction pattern
    2~3: Minor errors in understanding the trajectory sequence or direction
    0: Fails to properly analyze the trajectory or direction of travel
  Coordinate Processing (4 points)
    4: Accurately processes geographic coordinates to determine spatial relationships
    1~2: Minor errors in coordinate interpretation or spatial calculations
    0: Significant errors in coordinate processing or spatial analysis
  Candidate Destination Evaluation (5 points)
    5: Thoroughly analyzes each candidate destination relative to the trajectory
    2~3: Incomplete analysis of some candidate destinations
    0: Fails to properly evaluate candidates against the trajectory pattern
  Distance Calculation (3 points)
    3: Correctly assesses distances from last position to each candidate
    1: Minor errors in distance assessment
    0: Major errors or omissions in distance calculations
  Logical Reasoning (3 points)
    3: Employs sound geospatial logic to reach a reasoned conclusion
    1~2: Minor flaws in geospatial reasoning
    0: Significant gaps in logical reasoning process

Part 2: Final Answer Accuracy (<answer>) -- Max 10 Points
  Correctness of Prediction (7 points)
    10: Selects the destination that matches the ground truth answer
    1~2: Selects an incorrect destination but very close to the ground truth answer
    0: Selects an incorrect destination that is not close to the ground truth answer

Important Notes:
  The maximum total score is 30 points:
    - Up to 20 points from reasoning (<reason>)
    - Up to 10 points from final answer accuracy (<answer>)
  The minimum total score is 0 points.

Figure 15: The unique scoring prompt for the Prediction task.

Scoring Rubric -- Total: 30 Points
Problem Type: Navigation
Objective: Evaluate an LLM's geospatial reasoning ability by assessing how well it can determine the optimal (shortest or fastest) route between two locations given a specific road network.

Part 1: Reasoning Process (<reason>) -- Max 20 Points
  Network Comprehension (5 points)
    5: Correctly identifies all road segments and how they connect to form the complete road network
    2~3: Minor errors in understanding how segments connect
    0: Major misunderstanding of the road network structure
  Route Identification (5 points)
    5: Correctly identifies all possible routes from start to destination
    2~3: Misses some possible routes or includes impossible connections
    0: Fails to identify valid routes or creates non-existent connections
  Distance Calculation (6 points)
    6: Correctly calculates distances for necessary route segments using coordinate data
    2~4: Minor calculation errors or missing important segements
    0: Major calculation errors or inappropriate method for distance determination
  Optimal Route Selection (2 points)
    2: Correctly identifies the optimal route based on calculated distances
    1: Selects a non-optimal but valid route
    0: Selects an invalid or impossible route
  Clarity & Structure (2 points)
    2: Reasoning is clearly structured and logically flows
    1: Minor organizational issues or unclear explanations
    0: Disorganized or confusing reasoning

Part 2: Final Answer (<answer>) -- Max 10 Points
  The final answer will be compared against the ground truth answer.
  Accuracy (5 points)
    5: Route is completely correct and optimal
    2: Route is valid but not optimal
    0: Route is invalid or impossible
  Completeness (4 points)
    4: Includes all necessary segments in correct order
    1: Minor omissions or ordering issues
    0: Major omissions or incorrect ordering
  Clarity (1 points)
    1: Directions are clear, concise, and human-readable
    0: Directions are confusing or impossible to follow

Figure 16: The unique scoring prompt for the Navigation task.

Scoring Rubric -- Total: 30 Points
Problem Type: Recommendation
Objective: Evaluate an LLM's geospatial reasoning ability by assessing how well it can recommend the optimal Point of Interest that satisfies both geospatial constraints and user preferences..

Part 1: Reasoning Process (<reason>) -- Max 20 Points
  Location Understanding (4 points)
    4: Correctly identifies the user's location and understands spatial context
    2: Minor errors in interpreting the starting location
    0: Fails to comprehend the spatial nature of the query
  Proximity Analysis (5 points)
    5: Accurately calculates distances between user location and each POI
    2~3: Minor errors in distance calculation
    0: Incorrect proximity analysis or missing evaluation
  Constraint Evaluation (6 points)
    6: Correctly applies distance constraints (e.g., "within X meters") and thoroughly evaluates user preference constraints (ratings, features, etc.)
    2~4: Missing or incorrect application of distance constraints or user preferences
    0~1: Incomplete or incorrect analysis of user preferences
  Decision Framework (3 points)
    3: Creates a logical framework to weigh trade-offs between proximity and preferences
    0~1: Incomplete decision framework or illogical weighting of factors
  Clarity & Structure (2 points)
    2: Response is clearly organized into logical steps with appropriate reasoning
    1: Slight structural confusion or lack of clarity
    0: Disorganized or unclear reasoning process

Part 2: Final Answer Accuracy (<answer>) -- Max 10 Points
  10: Selects the optimal POI that matches the ground truth answer
  3~5: Selects a reasonable POI that satisfies most constraints but is not optimal
  0: Selects an inappropriate POI that violates key constraints

Important Notes:
  The maximum total score is 30 points:
    - Up to 20 points from reasoning (<reason>)
    - Up to 10 points from final answer accuracy (<answer>)
  The minimum total score is 0 points.

Figure 17: The unique scoring prompt for the Recommendation task.

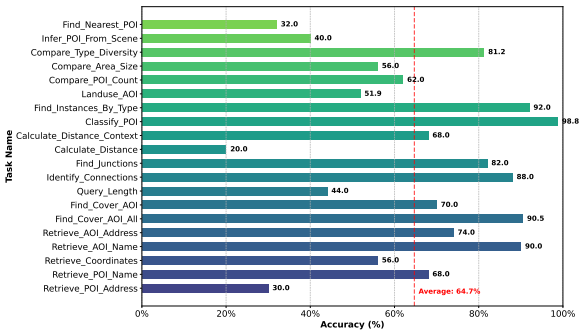

Figure 18: GPT-4o's Knowledge Module accuracy distribution across question templates on San Francisco.

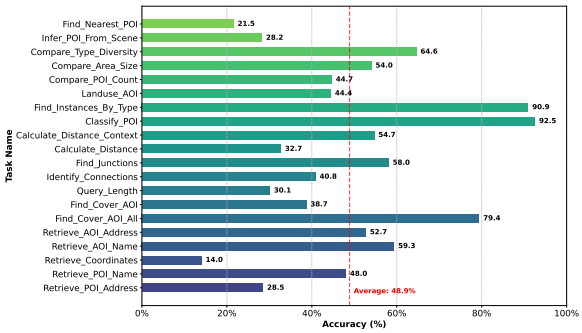

Figure 19: Llama-3.1-70B-Instruct's Knowledge Module accuracy distribution across question templates on San Francisco.

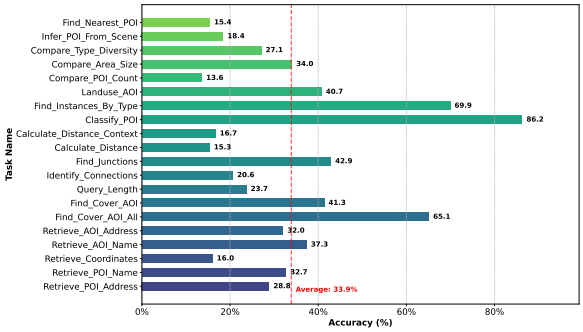

Figure 20: Llama-3.1-8B-Instruct's Knowledge Module accuracy distribution across question templates on San Francisco.

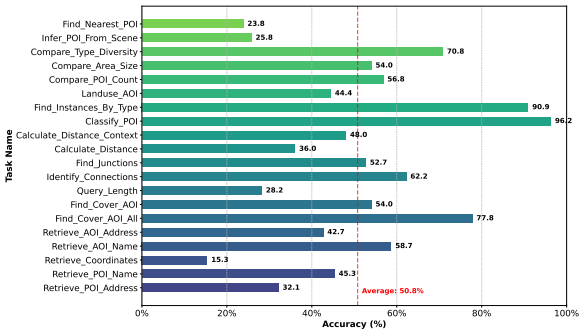

Figure 21: Qwen2.5-72B-Instruct's Knowledge Module accuracy distribution across question templates on San Francisco.

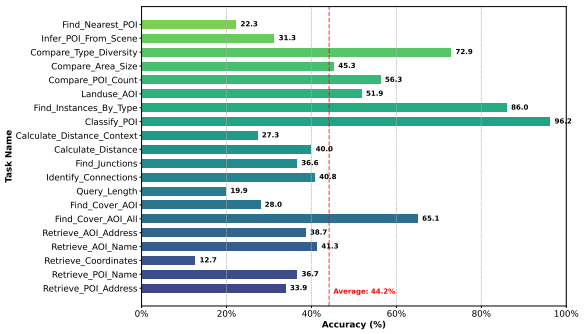

Figure 22: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on San Francisco.

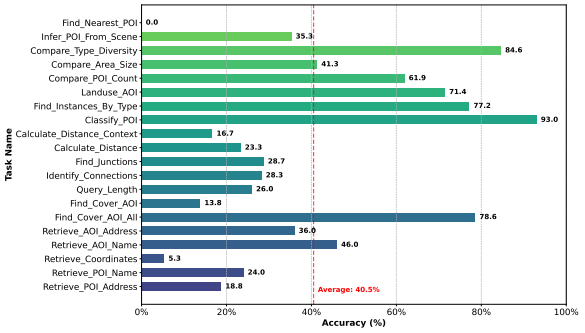

Figure 23: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on New Delhi.

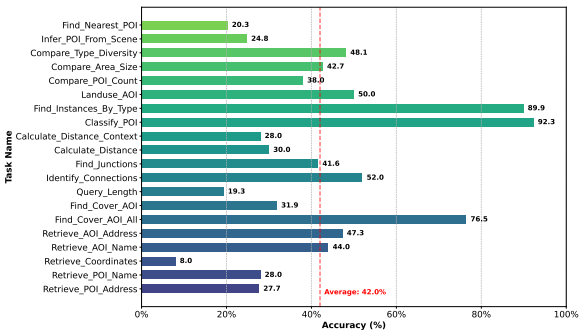

Figure 24: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on New York.

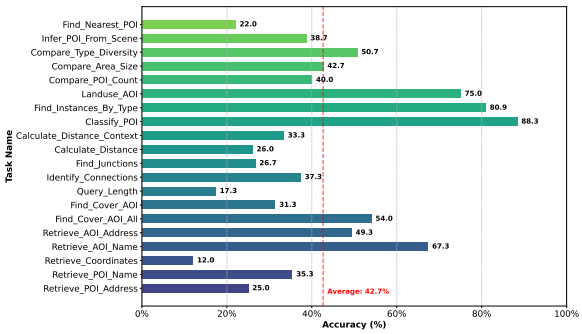

Figure 25: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on Tokyo.

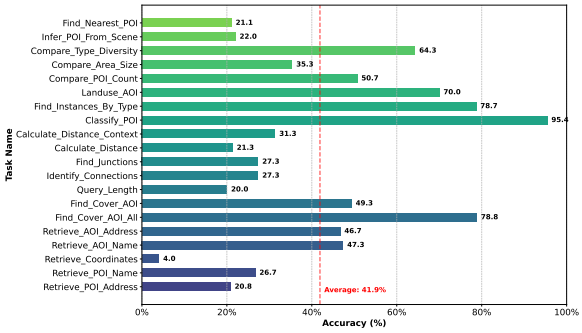

Figure 26: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on Berlin.

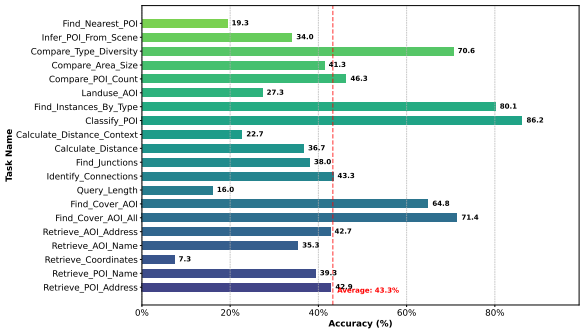

Figure 27: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on Paris.

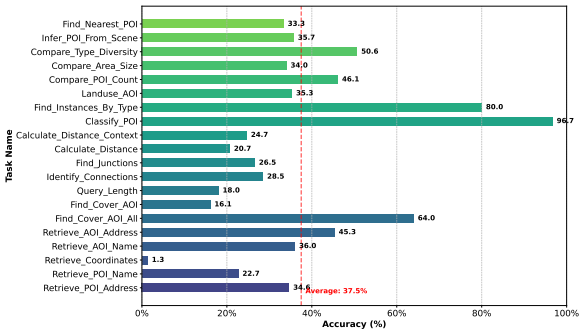

Figure 28: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on Dubai.

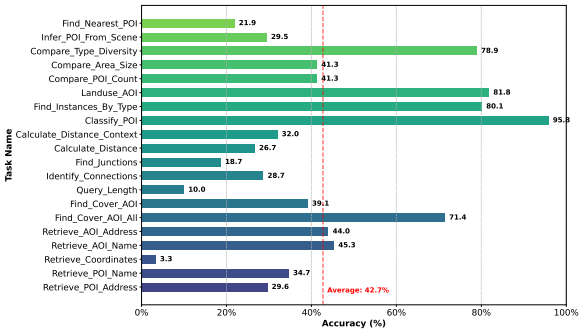

Figure 29: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on Rome.

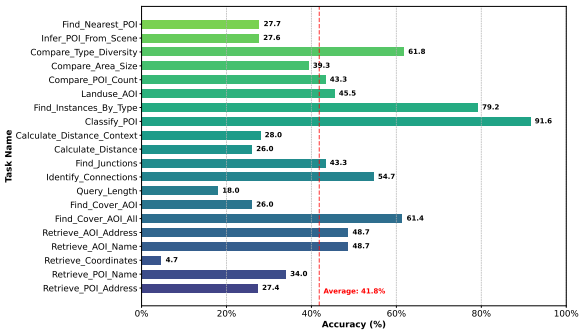

Figure 30: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on London.

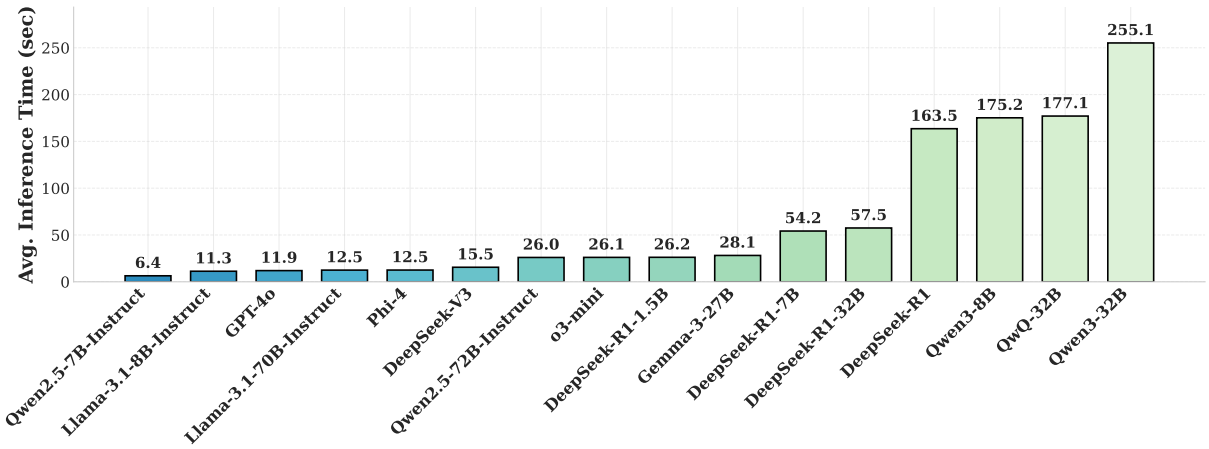

Figure 31: Comparison of average inference time per query across different models on Reasoning Module.

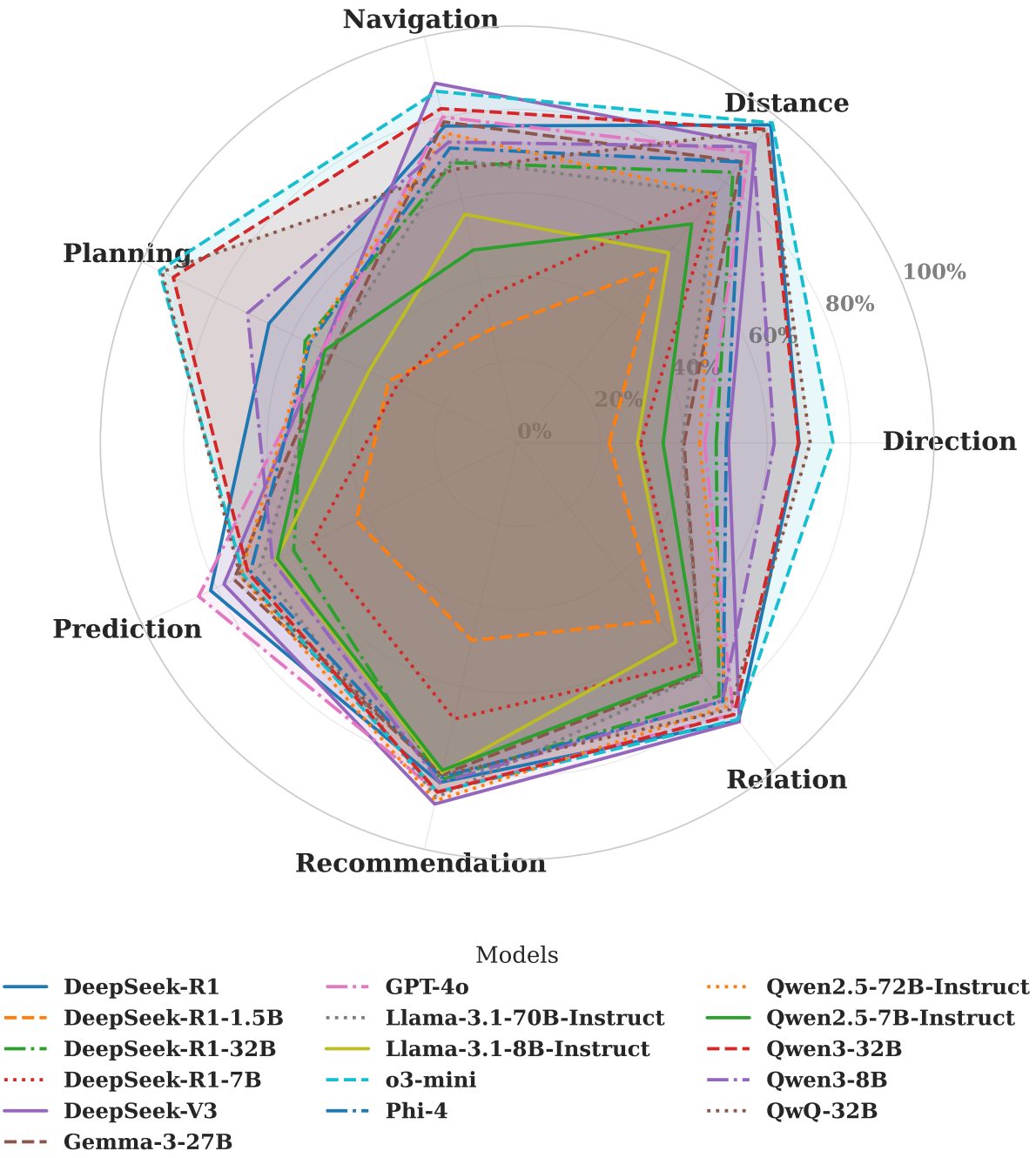

Figure 32: Radar chart of urban geospatial reasoning capabilities on all evaluated models.

Distance

Case:

Case Prompt: You are an expert urban geospatial reasoning assistant. Your task is to answer questions with adherence to the following structure and rules. Your response MUST consist of ONLY two sections formatted as follows:

Section 1: Reasoning Process
Provide a detailed, step-by-step explanation of your reasoning process enclosed within <reason> and </reason> tags. This section should capture all relevant considerations and intermediate steps, ensure that your reasoning is comprehensive yet focused to the question.

Format your reasoning process as outlined below:
Step 1: Extract Coordinates
- Clearly state the latitude and longitude of both locations in decimal degrees.
Step 2: Convert Degrees to Radians
Step 3: Calculate Differences
Step 4: Apply the Haversine Formula
Step 5: Computation

Section 2: Final Answer
Present your concise and definitive final answer enclosed within <answer> and </answer> tags.
Format strictly as: [Value] [Unit] (e.g., <answer>2.34 km</answer>).

Question:
Based on the latitude and longitude of The White House (38°53'51"N, 77°2'11"W) and The Pentagon (38°52'18"N, 77°3'22"W), what is the direct distance in kilometers between these two landmarks?

Model Response: <reason>
To calculate the direct distance between two points given their latitude and longitude, we can use the Haversine formula. This formula calculates the great-circle distance between two points on a sphere given their longitudes and latitudes.
Step 1: Convert the latitude and longitude from degrees to radians.
- Latitude of The White House: 38.8975° N = $38.8975 \times \pi/180 \approx 0.6790$ radians
- Longitude of The White House: 77.0364° W = $-77.0364 \times \pi/180 \approx -1.3444$ radians
- Latitude of The Pentagon: 38.8717° N = $38.8717 \times \pi/180 \approx 0.6785$ radians
- Longitude of The Pentagon: 77.0561° W = $-77.0561 \times \pi/180 \approx -1.3449$ radians
Step 2: Calculate the differences in latitude and longitude.
- $\Delta\text{lat} = 0.6790 - 0.6785 = 0.0005$ radians
- $\Delta\text{lon} = -1.3444 - (-1.3449) = 0.0005$ radians
Step 3: Apply the Haversine formula.
- $a = \sin^2(\Delta\text{lat}/2) + \cos(\text{lat1}) \times \cos(\text{lat2}) \times \sin^2(\Delta\text{lon}/2)$
- $c = 2 \times \text{atan2}(\sqrt{a}, \sqrt{1-a})$
- $d = R \times c$
Where R is the radius of the Earth (mean radius = 6,371 km).
Step 4: Calculate the intermediate values.
- $a = \sin^2(0.0005/2) + \cos(0.6790) \times \cos(0.6785) \times \sin^2(0.0005/2)$
- $a = \sin^2(0.00025) + \cos(0.6790) \times \cos(0.6785) \times \sin^2(0.00025)$
- $a \approx 0.0000000625 + 0.7915 \times 0.7917 \times 0.0000000625$
- $a \approx 0.0000000625 + 0.0000000486$
- $a \approx 0.0000001111$
- $c = 2 \times \text{atan2}(\sqrt{0.0000001111}, \sqrt{1 - 0.0000001111})$
- $c \approx 2 \times \text{atan2}(0.001054, 0.9999999444)$
- $c \approx 2 \times 0.001054$
- $c \approx 0.002108$ radians
Step 5: Calculate the distance.
- $d = 6371 \times 0.002108 \approx 13.43$ km
</reason>
<answer>
The direct distance between The White House and The Pentagon is approximately 13.43 kilometers.
</answer>

Evaluation:

Part 1: Reasoning Process
Coordinate Extraction (2 points):
The model correctly identifies the coordinates for both The White House and The Pentagon.
The White House: $38^\circ 53' 51''$N, $77^\circ 2' 11''$W is correctly represented as 38.8975° N, 77.0364° W.
The Pentagon: $38^\circ 52' 18''$N, $77^\circ 3' 22''$W is correctly represented as 38.8717° N, 77.0561° W.
The coordinates are given in decimal degrees, and no errors are found here. Score: 2/2
Conversion to Radians (3 points):
The model correctly converts all decimal degree values to radians using the formula: radian = degree $\times \pi/180$.
Latitude and longitude for both locations are converted correctly.
The results of the conversion are also shown with appropriate precision. Score: 3/3
Difference Calculation (3 points):
The model correctly calculates the differences in latitude and longitude in radians:
$\Delta\text{lat} = 0.6790 - 0.6785 = 0.0005$ radians.
$\Delta\text{lon} = -1.3444 - (-1.3449) = 0.0005$ radians.
Both differences are accurate. Score: 3/3
Haversine Formula Application (5 points):
The model correctly applies the Haversine formula:
The intermediate value a is computed as $a = \sin^2(\Delta\text{lat}/2) + \cos(\text{lat1}) \times \cos(\text{lat2}) \times \sin^2(\Delta\text{lon}/2)$.
The intermediate value c is calculated using the formula $c = 2 \times \text{atan2}(\sqrt{a}, \sqrt{1-a})$.
The model uses the correct Earth radius of 6,371 km.
The calculations are shown step by step, and the intermediate values for a and c are computed correctly, despite small rounding errors. Score: 5/5
Clarity & Structure (2 points):
The response is well-structured and easy to follow, with clear separation between the reasoning () and the final answer ().
No significant issues in organization or clarity. Score: 2/2
Total for Reasoning: 15/15
Part 2: Final Answer Accuracy
The model's final answer is 13.43 kilometers.
The ground truth is 3.34 kilometers.
The model's answer has a significant deviation from the correct value.
The error is $|13.43 - 3.34| = 10.09$ kilometers.
Using the scoring formula: Score = $\max(0, 10 - 10.09)$, the result is 0. Score: 0/10
Final Evaluation:
The reasoning process is accurate and clearly explained, earning the full 15 points.
However, the final answer is significantly incorrect, leading to a score of 0 for the answer accuracy.

Reason: [[15]] **Answer:** [[0]] **Total:** [[15]]

Figure 33: Case of Distance task.

Direction

Case:

Case Prompt: You are an expert urban geospatial reasoning assistant. Your task is to answer questions with adherence to the following structure and rules. Your response MUST consist of ONLY two sections formatted as follows:

Section 1: Reasoning Process
Provide a detailed, step-by-step explanation of your reasoning process enclosed within <reason> and </reason> tags. This section should capture all relevant considerations and intermediate steps, ensure that your reasoning is comprehensive yet focused to the question.

Format your reasoning process as outlined below:
Step 1: Extract Coordinates
- Clearly state the latitude and longitude of both locations in decimal degrees.
Step 2: Convert Degrees to Radians
Step 3: Calculate the Difference in Longitude
Step 4: Apply the Bearing Formula
Step 5: Convert the Bearing to Degrees
- Convert the calculated bearing from radians to degrees.
Step 6: Determine Descriptive Direction
- Based on the calculated bearing, provide the cardinal or intercardinal direction (e.g., North, Northeast, East-Southeast, etc.)

Section 2: Final Answer
Present your concise and definitive final answer enclosed within <answer> and </answer> tags.
Format strictly as: Bearing: [Value] degrees, Cardinal direction: [Direction] (e.g., <answer>Bearing: 123.96 degrees, Cardinal direction: Southeast</answer>)

Question:
Given two points of interest (POIs) in London:
1. The British Museum (Coordinates: 51.5196, -0.1270)
2. Tower of London (Coordinates: 51.5082, -0.0760)
Calculate the bearing from POI 1 to POI 2, measured clockwise from true north. Additionally, provide the direction in descriptive terms.

Model Response: <reason>
Let me calculate the bearing from the British Museum to the Tower of London step by step.
Step 1: Extract Coordinates
British Museum (POI 1): Latitude = 51.5196° N, Longitude = 0.1270° W
Converting to decimal degrees format: Latitude = 51.5196, Longitude = -0.1270
Tower of London (POI 2): Latitude = 51.5082° N, Longitude = 0.0760° W
Converting to decimal degrees format: Latitude = 51.5082, Longitude = -0.0760
Step 2: Convert Degrees to Radians
To convert degrees to radians, I'll multiply by $\pi/180$:
POI 1 (British Museum):
Lat1 (in radians) = $51.5196 \times (\pi/180) = 0.8992$ radians
Lon1 (in radians) = $-0.1270 \times (\pi/180) = -0.0022$ radians
POI 2 (Tower of London):
Lat2 (in radians) = $51.5082 \times (\pi/180) = 0.8990$ radians
Lon2 (in radians) = $-0.0760 \times (\pi/180) = -0.0013$ radians
Step 3: Calculate the Difference in Longitude
$\Delta\text{Lon} = \text{Lon2} - \text{Lon1} = -0.0013 - (-0.0022) = 0.0009$ radians
Step 4: Apply the Bearing Formula
The formula to calculate bearing θ in radians is:
$\theta = \text{atan2}(\sin(\Delta\text{Lon}) \times \cos(\text{Lat2}), \cos(\text{Lat1}) \times \sin(\text{Lat2}) - \sin(\text{Lat1}) \times \cos(\text{Lat2}) \times \cos(\Delta\text{Lon}))$
Calculating each component:
$\sin(\Delta\text{Lon}) = \sin(0.0009) = 0.0009$ (for small angles, $\sin(x) \approx x$)
$\cos(\text{Lat2}) = \cos(0.8990) = 0.6230$
$\sin(\Delta\text{Lon}) \times \cos(\text{Lat2}) = 0.0009 \times 0.6230 = 0.0006$
$\cos(\text{Lat1}) = \cos(0.8992) = 0.6228$
$\sin(\text{Lat2}) = \sin(0.8990) = 0.7821$
$\cos(\text{Lat1}) \times \sin(\text{Lat2}) = 0.6228 \times 0.7821 = 0.4871$
$\sin(\text{Lat1}) = \sin(0.8992) = 0.7823$
$\cos(\Delta\text{Lon}) = \cos(0.0009) = 0.9999$ (for small angles, $\cos(x) \approx 1$)
$\sin(\text{Lat1}) \times \cos(\text{Lat2}) \times \cos(\Delta\text{Lon}) = 0.7823 \times 0.6230 \times 0.9999 = 0.4873$
Numerator = $\sin(\Delta\text{Lon}) \times \cos(\text{Lat2}) = 0.0006$
Denominator = $\cos(\text{Lat1}) \times \sin(\text{Lat2}) - \sin(\text{Lat1}) \times \cos(\text{Lat2}) \times \cos(\Delta\text{Lon}) = 0.4871 - 0.4873 = -0.0002$
$\theta = \text{atan2}(0.0006, -0.0002) = 1.89$ radians
Step 5: Convert the Bearing to Degrees
Bearing (in degrees) = $\theta \times (180/\pi) = 1.89 \times (180/\pi) = 108.3$ degrees
The bearing is 108.3° from North, but we need to ensure it's in the $0 - 360^\circ$ range.
108.3° is already in the proper range, so our final bearing is 108.3°.
Step 6: Determine Descriptive Direction
A bearing of 108.3° corresponds to the East-Southeast (ESE) direction because:
East is at 90°
Southeast is at 135°
108.3° falls between East and Southeast, closer to East, making it East-Southeast (ESE)
</reason>
<answer>
The bearing from the British Museum to the Tower of London is 108.3° from true north, which corresponds to the East-Southeast (ESE) direction.
</answer>

Evaluation:

Evaluation of (Reasoning Process):
Coordinate Extraction (2/2):
The coordinates are clearly extracted and converted correctly into decimal degrees:
British Museum: $(51.5196, -0.1270)$
Tower of London: $(51.5082, -0.0760)$
Conversion to Radians (3/3):
The model correctly uses the formula (degrees $\times \pi/180$) and provides reasonable approximations:
Lat1 = 0.8992, Lon1 = -0.0022
Lat2 = 0.8990, Lon2 = -0.0013
Difference Calculation (3/3):
ΔLon is correctly computed:
$\Delta\text{Lon} = -0.0013 - (-0.0022) = 0.0009$ radians
Bearing Formula Application (4/5):
The bearing formula is applied correctly. However, there is a slight numerical discrepancy in the intermediate trigonometric values:
$\sin(\Delta\text{Lon})$ is approximated as 0.0009, which is acceptable for small angles, but using the exact sine would be more accurate.
Final $\theta = \text{atan2}(0.0006, -0.0002) = 1.89$ radians $\rightarrow 108.3^\circ$ This is off by ≈ 1.44 degrees from the ground truth (109.74°). Likely due to rounding and using approximations like $\sin(x) \approx x$. The structure and logic are solid, but some loss of precision results in a small error.
Clarity & Structure (2/2):
The reasoning is well-organized with clear, logical steps. Each calculation step is explained with the proper flow.
Reasoning Subtotal: 14/15

Evaluation of (Final Answer):
Reported Bearing: 108.3°
Ground Truth: 109.74°
Absolute Error = $|109.74 - 108.3| = 1.44$
Score = $\max(0, 10 - (1.44 \times 0.4)) = 10 - 0.576 = 9.42 \rightarrow$ Rounded to 9.42
Descriptive Direction: Both model and ground truth say East-Southeast (ESE) — no penalty.
Answer Subtotal: 9.42/10

Final Assessment: The model demonstrates a solid grasp of the bearing calculation process, with excellent clarity and structure. Slight imprecision in the trigonometric approximations caused a minor bearing error but within acceptable bounds. The final answer is nearly accurate, and the descriptive term matches correctly.

Reason: [[14]] **Answer:** [[9.42]] **Total:** [[23.42]]

Figure 34: Case of Direction task.

Relation (road-road)

Case:

Case Prompt: You are an expert urban geospatial reasoning assistant. Your task is to answer questions with adherence to the following structure and rules. Your response MUST consist of ONLY two sections formatted as follows:

Section 1: Reasoning Process
Provide a detailed, step-by-step explanation of your reasoning process enclosed within <reason> and </reason> tags. This section should capture all relevant considerations and intermediate steps, ensure that your reasoning is comprehensive yet focused to the question.

Section 2: Final Answer
Provide your final answer as a single spatial relation term enclosed in <answer> tags. Example: <answer>Intersecting</answer>

Question:
In New York City, consider the following two roads:
1. Columbus Ave (sampled coordinates: [(40.76986, -73.98428), (40.772545, -73.98234), (40.77470, -73.98075), (40.77674, -73.97928), (40.77923, -73.97750)])
2. Broadway (sampled coordinates: [(40.77909, -73.98199), (40.77702, -73.98206), (40.77470, -73.98215), (40.77180, -73.98217), (40.76865, -73.98207)])
Based on their coordinates, determine the spatial relationship between them. Are they:
- Intersecting (they cross each other),
- Parallel (they run alongside each other without crossing), or
- Disconnected (they do not intersect or run parallel)?

Model Response: <reason>
To determine the spatial relationship between Columbus Ave and Broadway based on the given coordinates, I will follow these steps:
1. Plot the coordinates on a map to visualize the roads.
2. Check if any of the coordinates from one road match or are very close to any of the coordinates from the other road. If there's a match or near-match, the roads intersect.
3. If there's no intersection, calculate the slopes of the linear segments between consecutive coordinates for each road.
4. Compare the slopes of the roads. If the slopes are approximately equal, the roads are parallel. If the slopes are different and there's no intersection, the roads are disconnected.
Step 1: Plotting the coordinates on a map would require specialized software or tools, which I cannot perform in this text-based environment. However, I can proceed with the next steps using the given coordinates.
Step 2: Checking for intersection by comparing coordinates:
- Columbus Ave: [(40.76986, -73.98428), (40.772545, -73.98234), (40.77470, -73.98075), (40.77674, -73.97928), (40.77923, -73.97750)]
- Broadway: [(40.77909, -73.98199), (40.77702, -73.98206), (40.77470, -73.98215), (40.77180, -73.98217), (40.76865, -73.98207)]
There's no exact coordinate match, but we need to check if any points are close enough to indicate an intersection.
Step 3: Calculate slopes for each road segment:
- Slope formula: $m = (y2 - y1)/(x2 - x1)$
- For Columbus Ave segments:
- Segment 1: $m1 = (-73.98234 + 73.98428)/(40.772545 - 40.76986)$
- Segment 2: $m2 = (-73.98075 + 73.98234)/(40.77470 - 40.772545)$
- Segment 3: $m3 = (-73.97928 + 73.98075)/(40.77674 - 40.77470)$
- Segment 4: $m4 = (-73.97750 + 73.97928)/(40.77923 - 40.77674)$
- For Broadway segments:
- Segment 1: $m5 = (-73.98206 + 73.98199)/(40.77702 - 40.77909)$
- Segment 2: $m6 = (-73.98215 + 73.98206)/(40.77470 - 40.77702)$
- Segment 3: $m7 = (-73.98217 + 73.98215)/(40.77180 - 40.77470)$
- Segment 4: $m8 = (-73.98207 + 73.98217)/(40.76865 - 40.77180)$
Step 4: Compare slopes:
- If any slope from Columbus Ave matches or is close to any slope from Broadway, the roads could be parallel.
- If none of the slopes match or are close, and there's no intersection, the roads are disconnected.
I will now compare the slopes calculated in Step 3 to determine if the roads are parallel.
</reason>
<answer>
The slopes of the segments of Columbus Ave and Broadway are not equal, indicating that the roads are not parallel. Since there is no exact coordinate match, and the slopes do not suggest parallelism, the roads are likely disconnected. However, without precise calculations of the slopes and checking for near-intersections, it's difficult to definitively conclude the relationship without further analysis. Given the available information and the absence of exact intersection points or matching slopes, the most likely spatial relationship is Disconnected.
</answer>

Evaluation:

Evaluation of the Model's Response
Part 1: Reasoning Process
1. Coordinate Extraction (1 point)
The model correctly restates the coordinates for both Columbus Ave and Broadway.
Score: 1/1
2. Intersection/Parallel/Disconnected Criteria (3 points)
The model identifies intersection by checking for exact or near coordinate matches.
It defines parallelism by comparing slopes of road segments.
It considers roads as disconnected if slopes differ and there is no intersection.
While the slope comparison approach is conceptually valid, it oversimplifies the geometry of roads and does not account for the fact that roads can intersect without sharing exact coordinates or having equal slopes.
Minor misunderstanding in how to rigorously detect intersections (e.g., using line segment intersection algorithms).
Score: 2/3
3. Analysis of Spatial Relationship (8 points)
The model makes an effort to analyze slopes between segments, using an approximate slope formula.
However, it makes several mathematical and conceptual errors:
It uses longitude as y and latitude as x, which inverts the usual convention for slope on a map (latitude is typically y).
It does not actually compute or compare slope values numerically—just sets up the formulas.
The analysis misses a key spatial observation: both roads share a coordinate with the same latitude (40.77470) and very close longitudes (~ 0.0014 degrees apart), implying possible intersection or very close proximity, which should be explored more thoroughly.
It fails to assess whether line segments between points might intersect geometrically.
This leads to a weak and ultimately incorrect analysis.
Score: 3/8
4. Clarity & Structure (3 points)
The response is clearly structured into and .
Steps are numbered and organized logically.
Score: 3/3
Subtotal for Reasoning: 9/15

Part 2: Final Answer Accuracy
The final answer is "Disconnected", which is incorrect.
The correct relationship is "Intersecting", as both roads cross near (40.77470 latitude), based on real-world NYC map data.
The model itself acknowledges the possibility of a close intersection but fails to pursue or conclude it.
Score: 0/10
Final Evaluation Summary
The model demonstrates a partially sound approach in setting up a framework to determine spatial relationships, but the actual execution and depth of analysis fall short.
The final answer is incorrect due to failure to analyze intersection proximity effectively.

Reason: [[9]] **Answer:** [[0]] **Total:** [[9]]

Figure 35: Case of Relation (road-road) task.

Relation (road-area)

Case:

Case Prompt: You are an expert urban geospatial reasoning assistant. Your task is to answer questions with adherence to the following structure and rules. Your response MUST consist of ONLY two sections formatted as follows:

Section 1: Reasoning Process
Provide a detailed, step-by-step explanation of your reasoning process enclosed within <reason> and </reason> tags. This section should capture all relevant considerations and intermediate steps, ensure that your reasoning is comprehensive yet focused to the question.

Section 2: Final Answer
Provide your final answer as a single spatial relation term enclosed in <answer> tags. Example: <answer>Intersecting</answer>

Question:
In New York City, consider the following road and area:
1. Road: Central Park West (sampled coordinates: [(40.77802, -73.97471), (40.78042, -73.97290), (40.78336, -73.97073), (40.78580, -73.96881), (40.78800, -73.96736)])
2. Area: Central Park (sampled boundary coordinates: [(40.78529, -73.96935), (40.78343, -73.96430), (40.78024, -73.96150), (40.77752, -73.96347), (40.77881, -73.96834), (40.78202, -73.97172)])
Based on their coordinates, determine the spatial relationship between them. Does the road:
Pass through the area (the road lies entirely or partially within the area),
Touch the area (the road intersects the boundary of the area but does not pass through it), or
Lie outside the area (the road does not intersect or touch the area)?

Model Response: <reason>
To determine the spatial relationship between Central Park West and Central Park, I will analyze the coordinates of both entities.
First, I will examine the coordinates of Central Park West, which appear to form a linear sequence. I will plot these coordinates to visualize the road's path. The coordinates suggest that the road runs in a generally north-south direction.
Next, I will analyze the coordinates of Central Park's boundary. These coordinates form a polygon that encloses the park. I will also plot these coordinates to visualize the park's boundary.
To determine the spatial relationship, I will check for intersections between the road and the park's boundary. I will use a spatial intersection algorithm to identify any points where the road's path intersects with the park's boundary.
Upon examining the coordinates, I notice that the road's path appears to intersect with the park's boundary at multiple points. Specifically, the road's coordinates $(40.78336, -73.97073)$ and $(40.78580, -73.96881)$ lie within the park's boundary.
However, not all of the road's coordinates are within the park's boundary. The road's coordinates $(40.77802, -73.97471)$ and $(40.78042, -73.97290)$ lie outside the park's boundary.
This observation suggests that the road does not lie entirely within the park's boundary, but it does intersect with the boundary at certain points.
</reason>
<answer>
Passes through the area
</answer>

Evaluation:

Reasoning Process (<reason>) – 15 Points
1. Coordinate Extraction (1 Point):
The response correctly states the coordinates for both the road and the area.
Score: 1

2. Boundary/Path Analysis (3 Points):
The model identifies the need to determine the relationship using intersection criteria. However, the reasoning contains a key misunderstanding: it notes that the road "intersects with the park's boundary at multiple points" and observes that some coordinates lie outside and some within the boundary, yet it incorrectly concludes that this means the road "passes through" the area rather than touching it. Although the method generally aligns with checking for intersections, this misapplication of criteria results in partial credit.
Score: 2

3. Analysis of Spatial Relationship (8 Points):
The reasoning describes a method for analyzing the relationship, including observing intersection points and comparing coordinates to the boundary, resembling basic spatial analysis. However, the analysis ultimately fails because it misinterprets the situation: the model identifies intersection points but does not distinguish between the definitions of "passes through" and "touches." This leads to an incorrect conclusion.
Score: 5

4. Clarity & Structure (3 Points):
The response is cleanly separated into <reason> and <answer> sections, and the structure is logical and easy to follow.
Score: 3

Reasoning Score Total: 1 + 2 + 5 + 3 = 11

Final Answer Accuracy (<answer>) – 10 Points
The final answer is "Passes through the area," but the ground truth is "Touch the area." The answer is incorrect, and no partial credit is awarded for this section.
Score: 0

Total Score
Reason: **11**, Answer: **0**, Total: **11**

Reason: [[11]] **Answer:** [[0]] **Total:** [[11]]

Figure 36: Case of Relation (road-area) task.

Relation (area-area)

Case:

Case Prompt: You are an expert urban geospatial reasoning assistant. Your task is to answer questions with adherence to the following structure and rules. Your response MUST consist of ONLY two sections formatted as follows:

Section 1: Reasoning Process
Provide a detailed, step-by-step explanation of your reasoning process enclosed within <reason> and </reason> tags. This section should capture all relevant considerations and intermediate steps, ensure that your reasoning is comprehensive yet focused to the question.

Section 2: Final Answer
Provide your final answer as a single spatial relation term enclosed in <answer> tags. Example: <answer>Intersecting</answer>

Question:
In New York City, consider the following two areas:
1. Theater District: (sampled boundary coordinates: [(40.76408, -73.98480), (40.76081, -73.98720), (40.75602, -73.99067), (40.75474, -73.98787), (40.75357, -73.98504), (40.75801, -73.98185), (40.76180, -73.97916), (40.76297, -73.98191), (40.76408, -73.98480)])
2. One Astor Plaza: (sampled boundary coordinates: [(40.75848, -73.98674), (40.75785, -73.98715), (40.75740, -73.98608), (40.75802, -73.98564)])
Based on their coordinates, determine the spatial relationship between them. Are they:
- Overlapping (they share common space),
- Adjacent (they share a boundary but do not overlap), or
- Separate (they do not share a boundary or overlap)?

Model Response: <reason>
To determine the spatial relationship between the Theater District and One Astor Plaza based on their boundary coordinates, we need to analyze whether these areas overlap, are adjacent, or are separate.
Step-by-Step Analysis:
1. Theater District Coordinates:
- $(40.76408, -73.98480)$
- $(40.76081, -73.98720)$
- $(40.75602, -73.99067)$
- $(40.75474, -73.98787)$
- $(40.75357, -73.98504)$
- $(40.75801, -73.98185)$
- $(40.76180, -73.97916)$
- $(40.76297, -73.98191)$
- $(40.76408, -73.98480)$

2. One Astor Plaza Coordinates:
- $(40.75848, -73.98674)$
- $(40.75785, -73.98715)$
- $(40.75740, -73.98608)$
- $(40.75802, -73.98564)$

Visualization and Geometric Analysis:
1. Plot the Coordinates:
- Plot the coordinates of both areas on a map to visualize their relative positions.

2. Check for Overlap:
- If any point from One Astor Plaza lies within the boundary of the Theater District, or vice versa, they are overlapping.
- From the coordinates, it appears that One Astor Plaza is entirely within the boundary of the Theater District. For example, the coordinates of One Astor Plaza are all within the range of the Theater District's coordinates.

3. Check for Adjacency:
- If the boundaries of the two areas touch at one or more points but do not overlap, they are adjacent.
- Given that One Astor Plaza is within the Theater District, this condition does not apply.

4. Check for Separation:
- If there is no shared boundary or overlap, they are separate.
- Since One Astor Plaza is within the Theater District, this condition does not apply.

Conclusion:
Based on the analysis, the spatial relationship between the Theater District and One Astor Plaza is:
- Overlapping: One Astor Plaza is entirely within the boundary of the Theater District.
</reason>
<answer>Overlapping.</answer>

Evaluation:

Part 1: Reasoning Process (<reason>) — Max 15 Points

1. Coordinate Extraction (1 Point):
The model correctly extracts and lists the coordinates for both the Theater District and One Astor Plaza.
Score: 1

2. Overlap/Adjacency/Separation Criteria (3 Points):
The response clearly defines the criteria for overlap, adjacency, and separation:
- Overlap: If one area's points fall within the boundary of the other.
- Adjacency: If their boundaries touch but do not overlap.
- Separation: If there is no intersection or shared boundary.

The criteria are correctly applied.
Score: 3

3. Analysis of Spatial Relationship (8 Points):
The reasoning demonstrates a correct analysis:
- The coordinates of One Astor Plaza were compared to see if they fall within the boundary of the Theater District.
- The conclusion correctly identifies that One Astor Plaza lies entirely within the Theater District, and thus the areas overlap.

Spatial reasoning is consistent and accurate, even without detailed computational steps.
Score: 8

4. Clarity & Structure (3 Points):
The response is well-structured, with a clear separation between the reasoning process and the final answer. The explanation is easy to follow and logically presented.
Score: 3

Reasoning Score Total: 1 + 3 + 8 + 3 = 15

Part 2: Final Answer Accuracy (<answer>) — Max 10 Points
The final answer, "Overlapping," matches the ground truth, indicating the correct identification of the spatial relationship.
Score: 10

Total Score
Reasoning: 15
Answer: 10
Total: 25

Reason: [[15]] **Answer:** [[10]] **Total:** [[25]]

Figure 37: Case of Relation (area-area) task.

Planning

Case:

Case Prompt: You are an expert urban geospatial reasoning assistant. Your task is to answer questions with adherence to the following structure and rules. Your response MUST consist of ONLY two sections formatted as follows:

Section 1: Reasoning Process
Provide a detailed, step-by-step explanation of your reasoning process enclosed within <reason> and </reason> tags. This section should capture all relevant considerations and intermediate steps, ensure that your reasoning is comprehensive yet focused to the question.

Steps to follow:
1. Understand the Problem
2. Gather the Given Data
3. Identify the Farthest Location
4. Plot the Locations
5. Find the Optimal Path

Section 2: Final Answer
Present your concise and definitive final answer enclosed within <answer> and </answer> tags.
Format strictly as: [Location 1]->[Location 2]->[Location 3]->[Location 4]->[Location 5] (e.g., <answer>Central Park->The Met->Times Square->Statue of Liberty->Brooklyn Bridge</answer>).

Question:

Model Response: <reason>
1. **Understand the Problem**: We need to find the optimal visiting order for the given locations in London, starting from The Northall, to minimize the total straight-line travel distance.
2. **Gather the Given Data**: We have five locations: The Northall, Trafalgar Square, London Bridge, London Eye, and Big Ben. We also have the bearing and distance between each consecutive pair of locations.
3. **Identify the Farthest Location**: Calculate the distances from The Northall to each of the other locations using the given bearings and distances. The farthest location from The Northall is London Bridge (approximately 3.11 km).
4. **Plot the Locations**: Plot the locations on a 2D coordinate system, considering the bearings and distances. This will help us visualize the relative positions of the locations.
5. **Find the Optimal Path**: Since we have identified the farthest location (London Bridge), we can start building the optimal path from there. We need to find the nearest unvisited location to London Bridge, which is London Eye (2.26 km). From London Eye, the nearest unvisited location is Big Ben (0.45 km). From Big Ben, the nearest unvisited location is Trafalgar Square (approximately 1.43 km, calculated using the bearing and distance from London Eye to Big Ben and Big Ben to Trafalgar Square). The last unvisited location is The Northall, which is the starting point.
6. **Calculate the Total Distance**: Calculate the total distance by summing up the distances between each pair of locations in the optimal order: London Bridge -> London Eye -> Big Ben -> Trafalgar Square -> The Northall.
</reason>
<answer>The optimal visiting order is: The Northall -> Trafalgar Square -> London Bridge -> London Eye -> Big Ben, with a total distance of approximately 6.97 kilometers.
</answer>

Evaluation:

Reasoning Evaluation
Coordinate Extraction (2 points)
The model correctly mentions the locations and the provided bearings and distances. However, there is no direct extraction or conversion of the bearing and distance data to coordinates for further analysis. This could result in an incomplete or less accurate method for finding the optimal path.
Score: 1 (since it missed doing the actual conversion of the distances and bearings into coordinates)
Farthest Location Identification (3 points)
The model claims that London Bridge is the farthest location, with a distance of 3.11 km. However, the actual farthest location, according to the provided data, is London Bridge with a distance of 2.79 km. This discrepancy suggests that the reasoning for identifying the farthest location is incorrect.
Score: 1 (incorrect identification of the farthest location)
Coordinate Conversion and Calculation (5 points)
The model does not explicitly perform the coordinate conversions or use them in the calculations. Without the coordinates, it is difficult to validate the proposed path and total distance.
Score: 1 (since it skips over the necessary coordinate conversion)
Path Evaluation (8 points)
The model attempts to construct a path based on the nearest unvisited location approach, which is a reasonable strategy. However, the model's calculation for the optimal path appears to be incorrect due to the misidentification of distances (like the wrong assumption that London Bridge is the farthest).
Furthermore, the total distance calculation of approximately 6.97 km seems to be based on the wrong distances. The actual optimal path has a total distance of 3.87 km.
Score: 3 (some good reasoning on the path, but overall flawed due to incorrect assumptions and calculations)
Clarity & Structure (2 points)
The response is logically structured, with clear sections for each part of the reasoning. However, the explanation could have benefited from more explicit steps for converting bearings to coordinates and calculating the optimal path.
Score: 2 (clearly separated into and sections with a logical structure)
Total for Reasoning : 8/20

Final Answer Evaluation
The final order proposed by the model is: The Northall -> Trafalgar Square -> London Bridge -> London Eye -> Big Ben.
This order is incorrect compared to the ground truth, which is: The Northall -> Trafalgar Square -> Big Ben -> London Eye -> London Bridge.
The distance stated by the model (6.97 km) is also incorrect, as the correct distance is 3.87 km.
Score: 3/10 (The order is incorrect, and the total distance is significantly different)

Reason: [[8]] **Answer:** [[3]] **Total:** [[11]]

Figure 38: Case of Planning task.

Prediction

Case:

Case Prompt: You are an expert urban geospatial reasoning assistant. Your task is to answer questions with adherence to the following structure and rules. Your response MUST consist of ONLY two sections formatted as follows:

Section 1: Reasoning Process
Provide a detailed, step-by-step explanation of your reasoning process enclosed within <reason> and </reason> tags. This section should capture all relevant considerations and intermediate steps, ensure that your reasoning is comprehensive yet focused to the question.
Steps to follow:
1. Analyze the sequence of locations in the partial trajectory
2. Determine the overall direction of travel
3. Calculate distances between the last recorded position and each candidate destination
4. Consider whether each candidate destination aligns with the established travel pattern
5. Predict the most likely final destination based on directional consistency and geospatial logic

Section 2: Final Answer
Provide only a single location name as your final answer, strictly enclosed within <answer> and </answer> tags. (e.g., <answer>Central Park</answer>)

Question:
Given a user's partial trajectory in a city/area, analyze the direction of travel and the distance to each candidate destination to predict the most likely final destination.
Partial Trajectory in Shanghai:
1. Shanghai Hongqiao Railway Station (31.1941, 121.3183) to Zhongchun Road (31.1146, 121.3634)
2. Zhongchun Road (31.1146, 121.3634) to Minhang Martyrs Cemetery (31.0966, 121.3751)
3. Minhang Martyrs Cemetery (31.0966, 121.3751) to Gao'an Road No.1 Primary School (Huazhan Campus) (31.1174, 121.4631)
4. Gao'an Road No.1 Primary School (Huazhan Campus) (31.1174, 121.4631) to Kangqiao East Metro Station (31.1525, 121.6217)
5. Kangqiao East Metro Station (31.1525, 121.6217) to Pudong Canal Bridge (31.2419, 121.6980)
6. Pudong Canal Bridge (31.2419, 121.6980) to Yingbin Interchange Bridge (31.1891, 121.7657)
Candidate Destinations:
A. Jing'an Kerry Centre (31.2299, 121.4562)
B. Shanghai Disneyland (31.1468, 121.6677)
C. HBC Cinema (Darongcheng Branch) (31.2862, 121.4338)
D. Shanghai Pudong International Airport (31.1493, 121.8149)
E. Shanghai Peace Park (31.2764, 121.5096)

Model Response: <reason>
I need to analyze the user's trajectory in Shanghai and determine the most likely final destination from the given candidates. Let me break this down step by step:
First, I'll map out the trajectory to understand the general direction of travel:
Starting at Shanghai Hongqiao Station $(31.1941, 121.3183)$ in the west
Moving to locations progressively eastward, ending at Yingbin Interchange $(31.1891, 121.7657)$ in the east
Let me visualize the overall direction:
The trajectory starts in western Shanghai (Hongqiao)
Moves through central areas (through Minhang District)
Then continues to eastern Shanghai (Pudong area)
The general movement shows a clear west-to-east pattern
The last recorded position $(31.1891, 121.7657)$ is in eastern Shanghai
Let me analyze each candidate destination:
A. Jing'an Kerry Centre $(31.2299, 121.4562)$: Central Shanghai, west of the last position, B. Shanghai Disneyland $(31.1468, 121.6677)$: Eastern Shanghai, but southwest of the last position, C. Cinema in Darong City $(31.2862, 121.4338)$: North-central Shanghai, west of the last position, D. Shanghai Pudong International Airport $(31.1493, 121.8149)$: Far eastern Shanghai, southeast of the last position, E. Shanghai Peace Park $(31.2764, 121.5096)$: North-central Shanghai, west of the last position
Consider the distance from the last position (Yingbin Interchange) to each candidate:
To A (Jing'an Kerry Centre): Moving backward (westward) significantly
To B (Disneyland): Moving backward (slightly westward and southward)
To C (Cinema): Moving significantly backward (westward)
To D (Pudong Airport): Continuing the eastward trajectory
To E (Peace Park): Moving backward (westward)
Analyzing the consistency with the trajectory pattern:
The trajectory shows consistent eastward movement
Only destination D (Pudong Airport) continues this eastward pattern
All other destinations would require reversing direction
Additionally, the overall trajectory makes sense for someone traveling from Hongqiao Station (west) to Pudong Airport (east) across the city.
Based on this analysis, the most likely destination is D. Shanghai Pudong International Airport, as it's the only option that:
Continues the established eastward movement pattern
Is a logical endpoint for someone crossing Shanghai from west to east
Makes sense as a destination given the starting point was another transportation hub (Hongqiao Station)
</reason>
<answer>
Based on the trajectory analysis, the most likely final destination is D. Shanghai Pudong International Airport $(31.1493, 121.8149)$. The user has been moving consistently eastward from Hongqiao Station across Shanghai, with their last recorded position at Yingbin Interchange being very close to and directly en route to Pudong Airport. This destination is the only candidate that maintains the established directional pattern without requiring a significant change in course.
</answer>

Evaluation:

Part 1: Reasoning Process (Max 20 points)

Trajectory Analysis (5 points)
The model correctly identifies the sequence of locations and outlines the direction of travel from west to east. The analysis clearly maps out the progression from Shanghai Hongqiao Station to the last recorded position at Yingbin Interchange in eastern Shanghai.
Score: 5/5 - The trajectory analysis is thorough and accurate, providing a clear understanding of the travel pattern.

Coordinate Processing (4 points)
The model appropriately identifies the relative position of the candidate destinations in relation to the last recorded position. It correctly points out that most of the destinations are located west of the last position, except for Shanghai Pudong International Airport, which aligns with the trajectory.
Score: 4/4 - The model correctly processes the coordinates and assesses spatial relationships between the last position and the candidate destinations.

Candidate Destination Evaluation (5 points)
The model evaluates each candidate destination based on their direction relative to the trajectory. It rightly dismisses candidates that are located to the west of the last position, as these destinations would require reversing the direction. It appropriately highlights Shanghai Pudong International Airport as the logical continuation of the eastward movement.
Score: 5/5 - The evaluation of each candidate is comprehensive and logically sound.

Distance Calculation (3 points)
Although distances are not explicitly calculated, the model's analysis of relative positions is adequate for determining the most likely destination. It accurately identifies that only Shanghai Pudong International Airport continues in the same eastward direction without requiring a change in course.
Score: 3/3 - The model effectively handles the spatial analysis without needing precise distance calculations.

Logical Reasoning (3 points)
The reasoning is consistent with the observed directional pattern. The model correctly identifies that Shanghai Pudong International Airport aligns with the trajectory, considering both the eastward travel and the nature of the starting point (Hongqiao Station, a major transport hub).
Score: 3/3 - The reasoning is logically sound and follows the directional consistency throughout the trajectory.

Part 2: Final Answer Accuracy (Max 10 points)

Correctness of Prediction (7 points)
The model correctly predicts Shanghai Pudong International Airport (D), which is the correct destination based on the trajectory and geospatial analysis.
Score: 7/7 - The prediction is correct.

Justification Quality (3 points)
The justification is clear and well-articulated. The model explains why Shanghai Pudong International Airport is the most likely destination, referencing the consistent eastward travel and the fact that it makes logical sense as an endpoint for someone traveling across Shanghai from Hongqiao Station.
Score: 3/3 - The justification is concise and well-reasoned, supporting the correct prediction.

Final Score:
Reasoning: [20/20]
The reasoning process is thorough, accurate, and logically sound.
Answer: [10/10]
The prediction is correct, and the justification is clear and well-argued.
Total: [30/30]

Reason: [[20]] **Answer:** [[10]] **Total:** [[30]]

Figure 39: Case of Prediction task.

Navigation

Case:

Case Prompt: You are an expert urban geospatial reasoning assistant. Your task is to answer questions with adherence to the following structure and rules. Your response MUST consist of ONLY two sections formatted as follows:

Section 1: Reasoning Process
Provide a detailed, step-by-step explanation of your reasoning process enclosed within <reason> and </reason> tags. This section should capture all relevant considerations and intermediate steps, ensure that your reasoning is comprehensive yet focused to the question.

Format your reasoning process as outlined below:
Step 1: Network Comprehension
Step 2: Route Identification
Step 3: Optimal Route Selection
Step 4: Generate Precise Driving Directions

Section 2: Final Answer
Present your concise and definitive final answer enclosed within <answer> and </answer> tags.

Question:
I'm driving from the Seattle Art Museum (47.6076, -122.3380) to Madison Park (47.6354, -122.2782). The possible road network between these two locations includes the following segments (consider them as straight lines):
I-5 Express: (47.6076, -122.3380) → (47.6420, -122.3215)
E Lake Washington Blvd: (47.6438, -122.2991) → (47.6260, -122.2925)
Portage Bay Viaduct: (47.6420, -122.3215) → (47.6438, -122.2991)
E Madison St: (47.6260, -122.2925) → (47.6354, -122.2782)
Boyer Ave E: (47.6420, -122.3215) → (47.6260, -122.2925)

Could you provide me with step-by-step directions for the shortest route to my destination?

Model Response: <reason>
To find the shortest route, I will analyze the given road network and identify the most direct path between the Seattle Art Museum and Madison Park. First, I will break down the road network into individual segments and their corresponding endpoints:
1. I-5 Express: $(47.6076, -122.3380) \rightarrow (47.6420, -122.3215)$
2. E Lake Washington Blvd: $(47.6438, -122.2991) \rightarrow (47.6260, -122.2925)$
3. Portage Bay Viaduct: $(47.6420, -122.3215) \rightarrow (47.6438, -122.2991)$
4. E Madison St: $(47.6260, -122.2925) \rightarrow (47.6354, -122.2782)$
5. Boyer Ave E: $(47.6420, -122.3215) \rightarrow (47.6260, -122.2925)$
Next, I will identify the starting point $(47.6076, -122.3380)$ and the destination $(47.6354, -122.2782)$. The first step is to find the closest road segment to the starting point. Segment 1 (I-5 Express) connects the starting point to $(47.6420, -122.3215)$. From $(47.6420, -122.3215)$, I will explore possible routes to the destination. Segment 3 (Portage Bay Viaduct) connects $(47.6420, -122.3215)$ to $(47.6438, -122.2991)$, and Segment 2 (E Lake Washington Blvd) connects $(47.6438, -122.2991)$ to $(47.6260, -122.2925)$. Finally, Segment 4 (E Madison St) connects $(47.6260, -122.2925)$ to the destination $(47.6354, -122.2782)$. By combining these segments, I can construct the shortest route.
</reason>
<answer>
The shortest route from the Seattle Art Museum to Madison Park is:
1. Take I-5 Express from (47.6076, -122.3380) to (47.6420, -122.3215).
2. Take Portage Bay Viaduct from (47.6420, -122.3215) to $(47.6438, -122.2991)$.
3. Take E Lake Washington Blvd from (47.6438, -122.2991) to (47.6260, -122.2925).
4. Take E Madison St from (47.6260, -122.2925) to (47.6354, -122.2782).
</answer>

Evaluation:

Part 1: Reasoning Process (Max 20 points)
Network Comprehension (5 points):
The model demonstrates a good understanding of the road network. It correctly identifies each segment, their endpoints, and the overall structure of the network. The connections are accurately stated, and all provided segments are considered.
Score: 5
Route Identification (5 points):
The model correctly identifies a possible route from the Seattle Art Museum to Madison Park, starting from the museum and following a series of road segments to the destination. However, the model does not evaluate all possible routes explicitly, such as the one involving Boyer Ave E. It assumes that the route via Portage Bay Viaduct and E Lake Washington Blvd is the shortest but doesn't clearly compare all routes.
Score: 3
Distance Calculation (6 points):
The model does not explicitly perform any distance calculations, even though it outlines the steps for constructing a path. While it mentions that it will find the shortest route, the reasoning process lacks any explicit distance computation or comparison.
Score: 2
Optimal Route Selection (2 points):
The model presents the route involving I-5 Express, Portage Bay Viaduct, E Lake Washington Blvd, and E Madison St. However, this is not the optimal route. The optimal route is via Boyer Ave E (I-5 Express → Boyer Ave E → E Madison St), which is shorter, but the model does not identify this route or provide reasoning for its selection.
Score: 1
Clarity & Structure (2 points):
The reasoning is generally clear and structured. The model breaks down the process logically and presents a well-organized approach. However, the explanation would benefit from a clearer comparison of the routes to explain why one is optimal.
Score: 2
Part 1 Total: 13 points

Part 2: Final Answer (Max 10 points)
Accuracy (5 points):
The route provided in the final answer is not the optimal one. The optimal route is the one involving I-5 Express → Boyer Ave E → E Madison St, but the model mistakenly suggests the route involving Portage Bay Viaduct and E Lake Washington Blvd.
Score: 3
Completeness (4 points):
The answer provides the correct segments for the route it suggests, but this route is not the optimal one. There is no mention of the Boyer Ave E option, which should be part of the optimal route.
Score: 2
Clarity (1 points):
The directions in the final answer are clear, concise, and easy to follow. The model provides logical and sequential instructions, making it easy to understand.
Score: 1
Part 2 Total: 6 points

Final Evaluation:
Reasoning Process: The model does a good job outlining the steps for identifying a route and breaking down the network, but it fails to compare all routes and lacks explicit distance calculations. This leads to the selection of a non-optimal route.
Final Answer: The final answer is clear and easy to follow, but it suggests a non-optimal route. The failure to identify the shortest route is the main issue here.

Reason: [[13]] **Answer:** [[6]] **Total:** [[19]]

Figure 40: Case of Navigation task.

Recommendation

Case:

Case Prompt: You are an expert urban geospatial reasoning assistant. Your task is to answer questions with adherence to the following structure and rules. Your response MUST consist of ONLY two sections formatted as follows:

Section 1: Reasoning Process
Provide a detailed, step-by-step explanation of your reasoning process enclosed within <reason> and </reason> tags. This section should capture all relevant considerations and intermediate steps, ensure that your reasoning is comprehensive yet focused to the question.

Section 2: Final Answer
Present your concise and definitive final answer enclosed within <answer> and </answer> tags.

Question:
I will be walking from $(40.6785, -73.9685)$ to $(40.6810, -73.9647)$ along this planned route: $[(40.6785, -73.9685)$, $(40.6794, -73.9682)$, $(40.6802, -73.9679)$, $(40.6810, -73.9676)$, $(40.6816, -73.9674)$, $(40.6810, -73.9647)]$. I'm looking for a lunch spot that's not far off the route.

Options include:
Zaytoons: $(40.6787, -73.9687)$, info: ['Open hour': '11:30AM-10PM', 'Review': 'I really like the place. The food is super fresh and authentic Palestinian cuisine. The service is excellent. We sat in here on a very cold night and the downside was that it was quite cold in there. The appetizers were exceptional.']
The Nuaa Table: $(40.6793, -73.9685)$, info: ['Open hour': '5-10PM', 'Review': 'This Thai restaurant is a must try! The food presentation is beautiful and contemporary, yet the taste is still authentically Thai.']
Chuko: $(40.6808, -73.9675)$, info: ['Open hour': '12-3PM, 5:30-10PM', 'Review': 'It's hard to not kick yourself after spending $25 on a bowl of ramen, but I'm not mad. The noodles were fresh and perfectly chewy, the broth was delicious, and it was a perfect spicy bowl of comfort on a cold day. Me and my friend we shared a bottle of sake, which completed the entire experience. Recommended dishes: Kimchi Ramen']
Socialites Lounge: $(40.6815, -73.9651)$, info: ['Open hour': '4PM-1AM', 'Review': 'Staff pretty chill. Food was delightful *Homemade. Rum Punch on point (added fruit) would of been nice. D.j. was jamming...Wish it was bigger']
Udom Thai: $(40.6774, -73.9635)$, info: ['Open hour': '12-11:30PM', 'Review': 'Craving authentic Thai flavors? Look no further than Udom Thai in Prospect Heights, Brooklyn From the cozy atmosphere to the mouthwatering spicy dishes made with the freshest ingredients, Udom Thai delivers an unforgettable dining experience.']

Which location offers the best combination of convenience and a comfortable ambiance, while also catering to my preference for spicy and distinctive Asian cuisine?

Model Response: <reason>
Step 1: Evaluate Proximity to Route
Zaytoons: $(40.6787, -73.9687)$ is very close to the starting point.
The Nuaa Table: $(40.6793, -73.9685)$ is near the first waypoint.
Chuko: $(40.6808, -73.9675)$ is near the middle of the route.
Socialites Lounge: $(40.6815, -73.9651)$ is slightly off the route but close to the endpoint.
Udom Thai: $(40.6774, -73.9635)$ is far from the route.
Step 2: Check Open Hours
Zaytoons: Open at 11:30AM, suitable for lunch.
The Nuaa Table: Opens at 5PM, not suitable for lunch.
Chuko: Open from 12-3PM, perfect for lunch.
Socialites Lounge: Opens at 4PM, not suitable for lunch.
Udom Thai: Open at 12PM, suitable for lunch.
Step 3: Assess Cuisine Preference
Zaytoons: Palestinian cuisine, not explicitly spicy or Asian.
The Nuaa Table: Thai cuisine, but not open for lunch.
Chuko: Japanese ramen with spicy options like Kimchi Ramen.
Socialites Lounge: General food, more of a lounge setting with homemade food.
Udom Thai: Authentic Thai with spicy dishes, but far from the route.
Step 4: Consider Ambiance and Reviews
Zaytoons: Positive reviews, but ambiance was noted as cold.
The Nuaa Table: Not applicable due to timing.
Chuko: Excellent reviews, cozy ambiance, and a comfortable dining experience.
Socialites Lounge: Not applicable due to timing.
Udom Thai: Great reviews, but location is inconvenient.
Step 5: Final Decision
Chuko is the best choice as it is directly on the route, open for lunch, offers spicy Asian cuisine, and has a comfortable ambiance with excellent reviews.
</reason>
<answer>Chuko</answer>

Evaluation:

Reasoning Process Evaluation (Max 20 points):

1. Route Understanding (4 points): The model successfully identifies the spatial relationship between the planned route and each lunch spot. It clearly maps out the proximity of each restaurant to the route, using qualitative descriptions like "very close," "near," and "far." The model does not present exact numerical distances but effectively conveys the general proximity. Score: 4/4

2. Proximity Analysis (5 points): The model correctly identifies and ranks the proximity of each restaurant to the route. It appropriately eliminates Udom Thai for being far off the route, and Chuko is identified as the best match in terms of location. This evaluation is accurate, even if the model does not calculate precise distances. Score: 5/5

3. Constraint Evaluation (6 points):
○ Operational Constraints: The model does a good job of checking the open hours for each restaurant, correctly eliminating options that aren't open for lunch, such as The Nuaa Table and Socialites Lounge. This is a critical step, and the model handles it well.
○ User Preferences: The model identifies Chuko's spicy ramen as a perfect match for the user's preference for spicy Asian cuisine. The analysis of cuisine is well-handled, and the choice is justified. Score: 6/6

4. Decision Framework (3 points): The decision-making process is logical and thorough. The model uses a structured approach to prioritize proximity, hours, cuisine preferences, and ambiance. Each step is clearly laid out, and Chuko emerges as the best option. Score: 3/3

5. Clarity & Structure (2 points): The model's reasoning is presented in a well-organized manner. It follows a step-by-step approach, clearly breaking down each criterion (proximity, hours, cuisine, ambiance) and presenting the results in a logical order. Score: 2/2
Final Answer Evaluation (Max 10 points):

The model selects Chuko, which is the correct and optimal choice based on the user's criteria. It takes into account the proximity to the route, open hours, cuisine preference for spicy Asian food, and the comfortable ambiance. The answer aligns perfectly with the ground truth. Score: 10/10

Reason: [[20]] **Answer:** [[10]] **Total:** [[30]]

Figure 41: Case of Recommendation task.

Recommendation

Case:

Case Prompt: You are an expert urban geospatial reasoning assistant. Your task is to answer questions with adherence to the following structure and rules. Your response MUST consist of ONLY two sections formatted as follows:

Section 1: Reasoning Process
Provide a detailed, step-by-step explanation of your reasoning process enclosed within <reason> and </reason> tags. This section should capture all relevant considerations and intermediate steps, ensure that your reasoning is comprehensive yet focused to the question.

Section 2: Final Answer
Present your concise and definitive final answer enclosed within <answer> and </answer> tags.

Question:
I'm currently located at $(38.8962, -77.0122)$ and looking for a museum nearby.
Available options:
National Museum of Asian Art: $(38.8882, -77.0274)$, features: ['Overview': 'Gallery run by the Smithsonian displaying Asian artwork such as painting & pottery.', 'Rating': 4.7(2113)]
National Children's Museum: $(38.8943, -77.0298)$, features: ['Overview': 'Reimagined kids' museum with exhibits & hands-on program for science, technology, arts & more.', 'Rating': 4.2(1301)]
Smithsonian's National Postal Museum: $(38.8983, -77.0082)$, features: ['Overview': 'Museum exploring the history of the U.S. postal service, from the Pony Express to stamp collections.', 'Rating': 4.7(3108)]
National Guard Memorial Museum: $(38.8980, -77.0094)$, features: ['Overview': 'This museum is at the headquarters if the National Guard Association of the United States. It is a great stop that helps visitors understand how the National Guard was the country's original defense force.', 'Rating': 4.6(122)]
National Portrait Gallery: $(38.8979, -77.0230)$, features: ['Overview': 'Major, free-admission museum with art works portraying influential Americans throughout history.', 'Rating': 4.8(12684)]

My daughter and I would like to visit the most popular museum that offers educational experiences, must be within 600 meters of our current location.

Model Response: <reason>
To find the most suitable museum for the query, we need to consider multiple factors:

1. Distance: Calculate the distance from the current location $(38.8962, -77.0122)$ to each museum option. We will consider only museums within 600 meters.
2. Popularity: Assess the popularity of each museum based on the number of ratings and the overall rating.
3. Educational experiences: Identify museums that offer interactive or hands-on experiences suitable for children.

Using the Haversine formula to calculate distances:
- National Museum of Asian Art: approximately 404 meters away
- National Children's Museum: approximately 375 meters away
- Smithsonian's National Postal Museum: approximately 143 meters away
- National Guard Memorial Museum: approximately 161 meters away
- National Portrait Gallery: approximately 749 meters away

Considering distance, only National Museum of Asian Art, National Children's Museum, Smithsonian's National Postal Museum, and National Guard Memorial Museum are within 600 meters.

Analyzing popularity:
- National Museum of Asian Art: 4.7(2113)
- National Children's Museum: 4.2(1301)
- Smithsonian's National Postal Museum: 4.7(3108)
- National Guard Memorial Museum: 4.6(122)

Smithsonian's National Postal Museum is the most popular among the options within 600 meters.

Assessing educational experiences:
- National Children's Museum: 'hands-on program for science, technology, arts & more'
- Smithsonian's National Postal Museum: 'exploring the history of the U.S. postal service, from the Pony Express to stamp collections' (less hands-on)

The National Children's Museum offers more interactive experiences suitable for children.

However, it's essential to note that while National Children's Museum offers educational experiences, it has lower overall rating compared to Smithsonian's National Postal Museum.

Since both museums meet the distance and educational requirements, and prioritizing popularity, I would recommend the more popular option with suitable educational content.
</reason>
<answer>
We will visit the **National Children's Museum** at $(38.8943, -77.0298)$, as it offers the most suitable balance of educational experiences and is within 600 meters of our current location. Although it's less popular than Smithsonian's National Postal Museum based on the number of ratings, it has more hands-on activities suitable for a visit with a child.
</answer>

Evaluation:

Part 1: Reasoning Process (<reason>)—Max 20 points

1. Location Understanding (4 points)
The model correctly identifies the user's current location $(38.8962, -77.0122)$ and applies spatial context by considering the distances between this location and the museums.
Score: 4/4

2. Proximity Analysis (5 points)
Using the Haversine formula, the model accurately calculates distances and selects only the options within the 600-meter constraint:
- National Portrait Gallery (749m) is appropriately excluded.
- The distances for the other museums are reasonable approximations within acceptable error margins.
Score: 5/5

3. Constraint Evaluation (6 points)
- The model evaluates the user's priority constraints: distance (within 600m), popularity (rating and number of reviews), and educational experiences (interactive and suitable for a child).
- The educational experiences of each museum are analyzed correctly, with the National Children's Museum identified as the most hands-on. However, the final weighting of popularity vs. education is flawed. The Smithsonian's National Postal Museum, which is more popular AND meets the constraints of distance and educational focus (e.g., history-related content), should have been chosen. The model prioritizes the more hands-on nature of the Children's Museum but skews its reasoning by downplaying the significance of its lower popularity.
Weakness: The Smithsonian's National Postal Museum better satisfies the user's overall preferences (proximity, popularity, and education). The model's prioritization of "hands-on" over the other more prominent constraints is inconsistent with the framework it established.
Score: 4/6

4. Decision Framework (3 points)
The model uses a structured framework (distance -> popularity -> education) to evaluate the options and makes a decision based on user constraints. However, the final trade-off between factors (popularity and educational suitability) is misapplied. While it acknowledges a conflict between popularity and interactivity, the decision to prioritize the latter does not align well with the user's "most popular" requirement.
Score: 2/3

5. Clarity & Structure (2 points)
The response is well-structured and clearly organized. The reasoning is broken into distinct logical steps (distance, popularity, education), making it easy to follow.
Score: 2/2

Reason Total: 17/20

Part 2: Final Answer Accuracy (<answer>)—Max 10 points

1. Accuracy of the Final Answer (10 points)
The chosen answer, "National Children's Museum," does not reflect the optimal match for the user's preferences as defined in the ground truth. The ground truth (Smithsonian's National Postal Museum) better satisfies **all** constraints:
- It is the most popular museum (rating: 4.7, 3108 reviews).
- It is within 600 meters.
- While its educational focus is not as hands-on as the Children's Museum, it still provides a relevant and engaging educational experience (e.g., historic exhibits about the U.S. postal system).

Though the National Children's Museum is a reasonable choice (as it satisfies proximity and education constraints), it fails to meet the "most popular" criterion emphasized in the question, making it suboptimal.
Score: 5/10

Final Evaluation

The model demonstrates strong reasoning in calculating proximity, evaluating constraints, and structuring its response. However, it incorrectly prioritizes interactivity over popularity, leading to the selection of a suboptimal answer. While the reasoning process is well-constructed and insightful, the final conclusion diverges from the ground truth, reducing its accuracy.

Rating:
Reason: 17/20
Answer: 5/10
Total: 22/30

Reason: [[17]] **Answer:** [[5]] **Total:** [[22]]

Figure 42: Another case of Recommendation task.