

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

Mutian Bao^{1*} Qiuyi Qi^{1*} Tian Liang¹ Jinjian Zhang²
Wei Zhou² Ming Kong^{1,3†} Linjian Mo^{2†} Qiang Zhu^{1†}

¹Zhejiang University ²Ant Group

³Zhejiang Key Laboratory of Geographic Information Science
{bmt, zhuq}@zju.edu.cn

Abstract

Current evaluations of geospatial reasoning in LLMs are frequently impeded by the entanglement of factual recall and spatial logic, which often obscures the models' true capabilities in complex city-scale environments. To address this, we introduce UrbanGeoEval, a comprehensive benchmark featuring a dual-module framework designed to disentangle these competencies. The Knowledge Module assesses urban memory via scalable map-based queries, while the Reasoning Module isolates pure logical inference across 3,148 realistic tasks by providing necessary geospatial context. Unlike prior benchmarks that hand the model pre-computed spatial text, UrbanGeoEval provides raw geometry and forces the model to act as a spatial computing engine. Our evaluation methodology introduces a reliable hybrid pipeline that merges deterministic programmatic checks with an LLM-as-a-Judge, achieving expert-level evaluation accuracy. Extensive experiments on 18 widely used LLMs uncover critical insights: (1) models exhibit severe geographic biases and resolution gaps; (2) failures in complex multi-hop tasks often stem from brittle foundational spatial skills rather than high-level logic deficits. UrbanGeoEval provides a precise diagnostic tool for advancing urban geospatial intelligence in LLMs.

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 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

*Equal contribution.

†Corresponding authors.

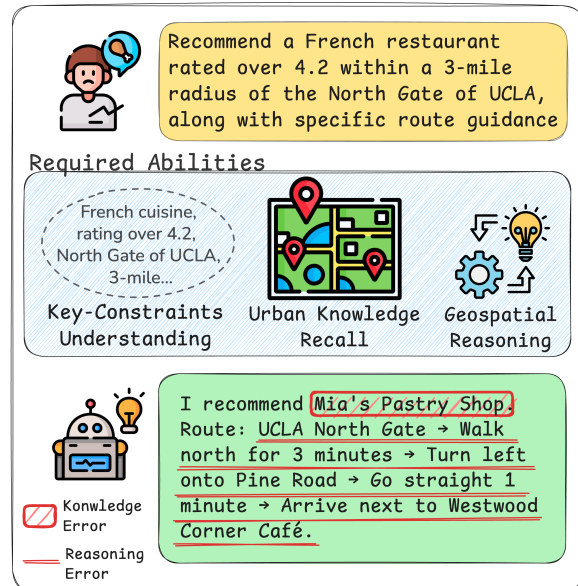


Figure 1: A typical LLM failure on a city-scale geospatial query, where a knowledge error (hallucinating a POI) leads to a subsequent reasoning error (generating an invalid path). This confounding of error types motivates our decoupled benchmark design.

(POIs), road networks, and heterogeneous Areas of Interest (AOIs), along with fine-grained spatial relations and diverse constraints (Wang et al., 2020). 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.,

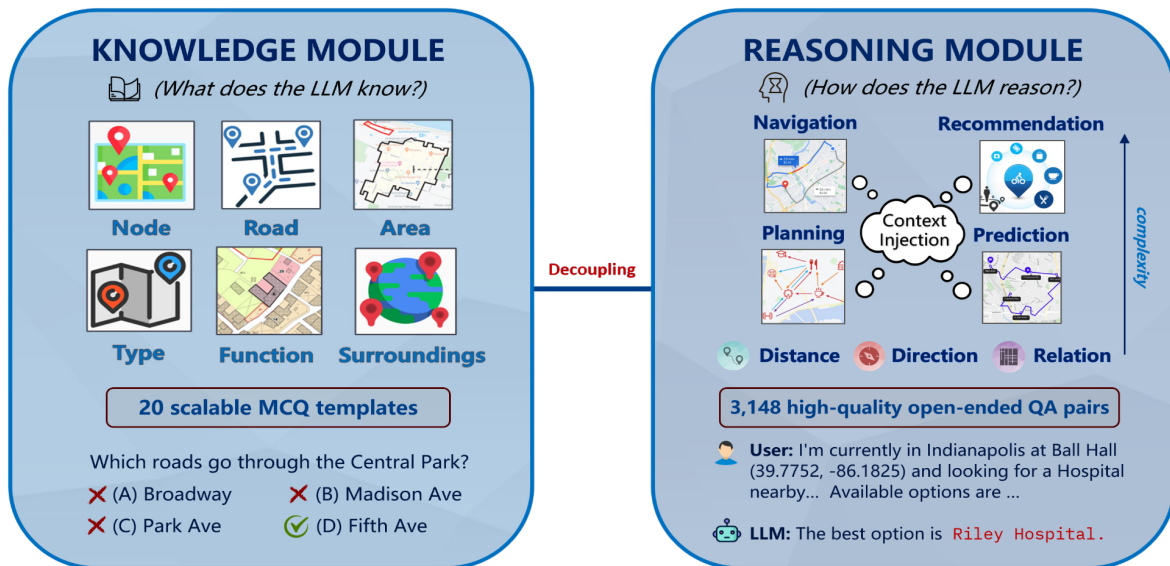


Figure 2: The overall architecture of the UrbanGeoEval benchmark. The Knowledge Module (left) uses an automated pipeline to generate multiple-choice questions from OSM data. The Reasoning Module (right) uses human-annotated scenarios to create open-ended questions that test pure reasoning with in-context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

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 locations (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. Equally important, existing text-centric GeoQA benchmarks (Dihan et al., 2025; Feng et al., 2025) provide models with heavily pre-processed spatial descriptions, so that routing or trip-planning is essentially turned into a reading-comprehension task plus straightforward scalar arithmetic. As a result, they fail to evaluate whether an LLM can function as an end-to-end spatial computing engine operating directly on raw geometric data.

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 dimen-

sions 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 that range 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 reasoning 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

¹Code and data are available at <https://github.com/mtskyyy/UrbanGeoEval>.

- 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 introduce a hybrid, process-aware evaluation framework that pairs deterministic programmatic checkers with an LLM-as-Judge, guided by fine-grained rubrics to evaluate not only the final answer, but also the complete reasoning process.
- We conduct a large-scale analysis of 18 popular LLMs spanning 1.5B–671B parameters across 18 cities on six continents, providing a detailed diagnosis 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 (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 Geospatial Evaluation in the Era of Reasoning LLMs

Existing geospatial benchmarks face limitations that are amplified by the emergence of reasoning-optimized LLMs like DeepSeek-R1 (DeepSeek-AI et al., 2025). While these models leverage Reinforcement Learning Fine-Tuning to excel in symbolic STEM (Hendrycks et al., 2021; Rein et al., 2023) domains, their transferability to noisy, real-world geospatial environments remains underexplored. Current text-based benchmarks like CityEval (Feng et al., 2025), STBench (Li et al., 2024), TourismQA (Contractor et al., 2021), and MapEval-Textual (Dihan et al., 2025) typically lack the complexity to challenge these advanced capabilities, while multimodal benchmarks such as MapIQ (Srivastava et al., 2025), MapWise (Mukhopadhyay

Statistic	Value
Knowledge Dimensions	6
Question Templates	20
Questions per City	~3,000
Total Questions	~54,000
Evaluated Cities (18 Total, 6 Continents)	
Auckland, Bangkok, Berlin, Buenos Aires, Cairo, Dubai, London, Moscow, Nairobi, New Delhi, New York, Paris, Rio de Janeiro, Rome, San Francisco, Sydney, Taipei, Tokyo	

Table 1: Data statistics for the Knowledge Module.

et al., 2025) are hindered by data scarcity. Crucially, the closest text-based predecessors operate at a different cognitive level: MapEval-Textual hands the model verbalized routes such as “Travel time from A to B is 11 mins (0.6 km)”, so even with the multiple-choice options removed, the task primarily tests text extraction and 1D scalar arithmetic; CityEval’s spatial-reasoning split couples reading comprehension with single-line distance addition, its mobility-prediction split evaluates the semantic understanding of human temporal routines, and its navigation split is decomposed into localized step-by-step MCQs. In contrast, UrbanGeoEval supplies raw WGS84 coordinates for points, polylines, and AOI polygons; topological relations and pre-computed metrics are deliberately withheld, forcing the LLM to perform intrinsic 2D trigonometry, line/polygon intersection analysis, end-to-end routing over a raw spatial graph, and constraint integration over geometric POI fields. A detailed comparison is provided in Appendix G.

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 envisioned as a sequence of “knowledge probes” that carefully chart the LLM’s internal, fixed model of the urban world, comparable to making queries to a geographic in-

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.

formation system (GIS).

3.1.1 Data Collection and Generation

Our approach starts with an automated pipeline that extracts and organizes geospatial data (POIs, AOIs, and road networks) from the comprehensive, open-source OpenStreetMap² (OSM) database. To connect this structured data with natural language, we employ the Nominatim³ reverse geocoding engine, converting raw coordinates into readable postal addresses. This curated knowledge base forms the ground truth for an automated question-generation pipeline that creates a large-scale, extensible collection of multiple-choice questions. For our experiments, we built a dataset covering 18 international cities spanning six continents to capture geographic diversity, with summary statistics reported in Table 1.

3.1.2 Taxonomy of Knowledge Tasks

These dimensions cover the full vector geometry stack, including 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 rigorously evaluate the models and reduce reliance on superficial pattern matching, we developed four methods for generating plausible but incorrect distractor options. Each method targets a distinct aspect of knowledge resolution: spatial proximity to probe geographic precision, semantic similarity to assess disambiguation, numerical perturbation to test quantitative accuracy, and attribute substitution to examine entity-property associations. Further details on the distractor generation logic are provided in Appendix A.1.

3.2 Reasoning Module

The Reasoning Module evaluates process-level geospatial reasoning given all necessary facts in context, so the model is not required to retrieve world knowledge from pretraining. Its main aim is to examine the full Chain-of-Thought, assessing not only whether the answer is correct but also the reasoning steps behind it, thereby separating true inference from lucky guesses.

²<https://openstreetmap.org/>

³<https://nominatim.org/>

3.2.1 Background Context Schema

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

- **POI:** Encompasses its categorical and spatial attributes, including the entity name, WGS84 coordinates, human-readable address, category type, and metadata such as user ratings and opening hours.
- **Road Segment:** Consists of the road name and a polyline sequence of WGS84 coordinates sampled along its centerline to represent its spatial topology.
- **AOI:** Includes its name, land-use type, and a spatial boundary defined by a closed ring of 4–10 uniformly sampled WGS84 coordinates.

This structured information is serialized into prompts as either compact key-value blocks or JSON-like fragments to ensure the model can effectively parse the provided geospatial context.

3.2.2 Taxonomy of Reasoning Tasks

We developed seven realistic scenarios spanning two complexity tiers: *Foundational Spatial Reasoning* (e.g., Distance, Direction, Relation) and *Advanced Composite Reasoning* (e.g., Itinerary Planning, Trajectory Prediction, Navigation, POI Recommendation). Refer to Appendix E for specific examples.

Distance (DS) Asking model to compute the great-circle 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. 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, road-area interaction, and area-area relation. 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, 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

To build our Reasoning Module, 15 trained annotators extracted geospatial scenarios from global web sources, such as travel Q&A sites, tourism forums, and online mapping services. We then converted these into question-answer pairs through an automated pipeline (see Table 3 for distribution). Our quality control follows a dual-pass protocol: we first validate geometric correctness and constraint completeness, then review for linguistic clarity. Any cases with ambiguous or non-deterministic solutions were strictly revised or removed to ensure answer uniqueness. Detailed recruitment and the validation mechanism are described in Appendix A.2 and A.3.

4 Evaluation

For the **Knowledge Module**, we employ *Accuracy* as the primary metric to measure the models’ ability to recall static geospatial facts across six core

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.

dimensions.

Evaluating the **Reasoning Module** is more complex, as a correct answer can hide a flawed logical process. To ensure the authenticity and rigor of the evaluation, we implement a hybrid assessment framework that disentangles logic from outcome. While the reasoning process for all tasks is assessed by the LLM-as-a-Judge, the final answer is evaluated according to the specific type of task:

- **Programmatic Evaluation:** For tasks with deterministic ground truth, we use automated scripts to guarantee complete accuracy. Specifically, for quantitative tasks (e.g., Distance, Direction), we apply a formula-based error penalty whose unified definition is provided in Appendix C. For topological and entity matching tasks (e.g., Relation, Prediction), we use *Exact Match* to verify correctness. For the Planning task, we adopt *Hamming Distance* to measure the alignment between the predicted and optimal sequences.
- **LLM-based Evaluation:** For tasks where the outcome is formulated in natural language (e.g., Navigation and Recommendation), we utilize GPT-4o judge. Following detailed, task-specific **Scoring Rubrics** and checklists in Appendix C, the judge performs a fine-grained assessment of both the reasoning chain and the final answer.

Based on this hybrid framework, each response is assigned a *Reason Score* and an *Answer Score*. To enable fair cross-task comparisons, the total score is normalized to a 0–100% scale.

5 Experiments

We conduct a comprehensive evaluation on Urban-GeoEval to assess the urban geospatial capabilities of different LLMs. Details on experimental setup and implementations can be found in Appendix B.

5.1 Main Results

Knowledge Module Performance As detailed in Table 4, GPT-4o leads with the highest average accuracy (54.59%) among evaluated models,

likely due to its superior parameter scale and high-quality training. While larger models consistently outperform their smaller counterparts, these gains exhibit diminishing returns. For instance, a ten-fold increase in model scale (e.g., Qwen2.5-7B to 72B) yields only a 7-point accuracy improvement. Notably, the Qwen series consistently surpasses Llama series among open-source models, potentially owing to broader global coverage within its pre-training corpora.

Reasoning Module Performance The evaluation of the Reasoning Module uncovers more nuanced patterns (see Figure 3 and Table 5). At the top of the leaderboard are models purpose-built for reasoning. Among them, the Qwen3.5-397B-A17B leads with the highest overall score, closely followed by o3-mini, Qwen3-32B, MiniMax-M2.5, QwQ-32B and DeepSeek-R1. Remarkably, the leading dense 32B-class reasoning models (e.g., QwQ-32B) surpass even proprietary instruction-tuned models such as GPT-4o, and Qwen3-32B outperforms the far larger DeepSeek-R1 (671B). These results indicate that geospatial intelligence is driven less by raw parameter scale and more by the quality of post-training: well-designed RL methodologies, especially those that encourage logical, low-hallucination Chains-of-Thought that correctly incorporate geospatial constraints, appear to be the dominant factor. From a practical standpoint, distilling strong geospatial reasoning into efficient, deployable dense models is therefore a critical direction for future urban-AI deployment.

5.2 In-depth Analysis

Geospatial Biases and Granularity Bottlenecks Evaluation on Knowledge Module uncovers two systemic limitations in current LLMs: regional disparity and resolution gaps. As shown in Figure 5, GPT-4o exhibits a large performance variance across cities, with notable accuracy drops in Asian cities such as New Delhi and Tokyo, due to the skew in pre-training corpora. Detailed results for other LLMs are provided in Appendix D. 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 tail geo-entities. More extensive results are provided in Appendix D.

Model	Node	Area	Road	Type	Function	Surroundings	Overall
GPT-4o	38.83	68.18	59.59	76.15	53.29	28.89	54.59
Llama-3.1-70B-Instruct	30.05	56.02	38.24	77.63	49.09	22.55	45.71
Llama-3.1-8B-Instruct	22.96	45.48	28.03	66.69	32.04	20.43	35.89
Qwen2.5-72B-Instruct	30.94	59.25	36.54	78.86	53.56	27.11	47.64
Qwen2.5-7B-Instruct	23.65	46.97	28.57	74.28	48.60	25.46	40.57
Average	29.29	55.18	38.19	74.72	47.32	24.89	44.88

Table 4: Overall accuracy (%) on the six dimensions of the Knowledge Module, averaged across all 18 cities.

Model	Foundational Reasoning						Advanced Reasoning							
	DS		DR		RE		PN		PR		NV		RC	
	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	9.94	13.80	8.54	13.95	7.27	18.88	9.50	17.35	4.29	17.41	8.55	18.65	7.12
Qwen3.5-397B-A17B	15.00	9.99	14.36	9.74	13.36	8.04	15.75	8.47	18.32	5.84	17.39	8.11	19.04	7.09
MiniMax-M2.5	14.12	9.84	12.42	8.73	12.69	7.52	18.28	9.73	17.06	4.77	14.19	6.08	18.78	7.43
Qwen3-32B	14.27	9.81	12.51	6.89	14.05	6.80	17.90	9.39	17.91	3.42	16.63	8.05	19.04	6.74
Qwen3-8B	13.53	9.09	11.43	6.14	13.77	6.00	13.86	6.93	16.56	1.52	15.28	6.93	18.68	6.40
QwQ-32B	14.32	9.61	12.97	7.58	13.97	6.36	18.68	9.42	17.88	3.89	14.05	6.15	17.71	6.82
DeepSeek-R1	14.65	9.76	12.77	6.53	14.15	7.16	13.08	4.35	18.41	5.69	15.96	7.44	18.26	6.81
DeepSeek-V3	13.53	9.45	11.39	3.24	14.13	7.25	10.41	4.48	18.23	4.35	17.78	8.77	19.27	7.38
DeepSeek-R1-32B	12.33	8.34	9.87	4.00	13.12	6.30	10.66	5.73	14.73	2.85	14.03	6.66	18.17	6.71
Llama3.1-70B-Ins.	11.13	7.96	9.01	2.40	12.92	4.84	9.85	4.25	16.98	2.89	14.45	6.48	18.92	7.24
Llama3.1-8B-Ins.	8.42	7.13	7.92	0.56	12.00	3.30	7.73	2.83	16.60	2.00	12.82	4.06	18.19	6.15
Qwen2.5-72B-Ins.	11.89	7.35	10.85	1.91	13.86	6.34	11.43	4.12	18.05	2.73	15.78	7.08	19.01	7.41
Qwen2.5-7B-Ins.	11.66	4.47	9.01	0.80	13.14	4.41	10.29	3.90	16.68	1.44	11.23	3.01	17.97	6.20
Gemma-3-27B	13.90	7.33	10.28	1.08	13.33	4.40	9.88	3.53	18.30	3.08	16.38	7.31	18.07	6.44
Phi-4	13.58	7.84	11.50	1.87	13.76	6.06	11.39	4.04	17.63	2.71	14.89	6.88	18.44	6.26
DeepSeek-R1-7B	11.42	7.21	7.09	0.52	12.08	4.62	5.70	1.49	14.03	1.82	8.45	2.24	15.28	5.10
DeepSeek-R1-1.5B	8.96	3.57	5.49	0.19	9.73	2.75	6.78	0.49	10.82	1.56	7.45	0.94	11.26	3.32
Task Avg.	12.84	8.18	10.80	4.03	13.22	5.90	12.30	5.41	16.91	3.42	14.48	6.24	17.99	6.52

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).

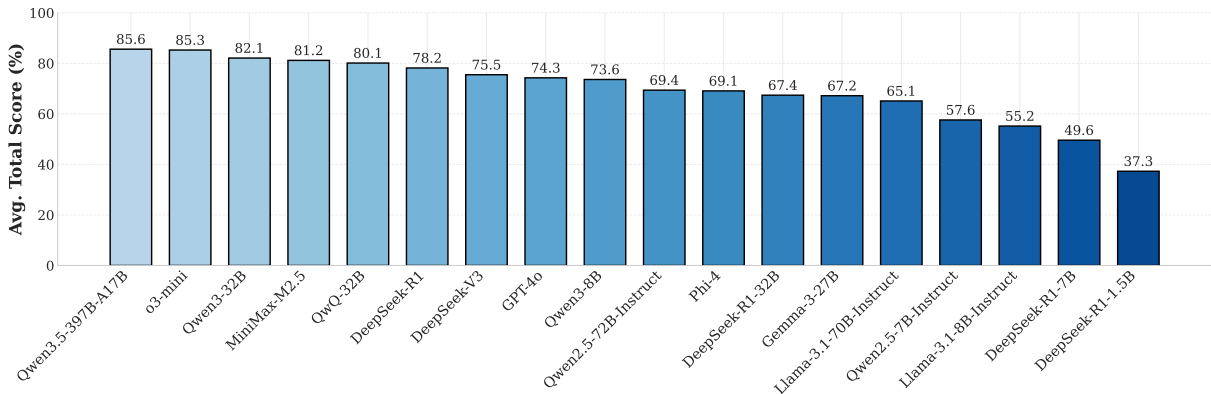


Figure 3: Overall performance of all 18 evaluated models on the Reasoning Module.

Brittle Reasoning and Failure Modes As shown in Figure 4, reasoning performance varies substantially across tasks. Models tend to perform reasonably well on semantic tasks such as *Recommendation*, but they face significant difficulties with more algorithmic tasks, including *Planning* and *Navigation*. Through an in-depth case analy-

sis of reasoning trajectories, we uncover three distinct categories of failure: (1) **Arithmetic Drift**, in which smaller models frequently arrive at incorrect answers despite following valid reasoning steps, due to limited numerical precision; (2) **Topological Blindness**, where models are unable to mentally reconstruct geometric structures from raw coordi-

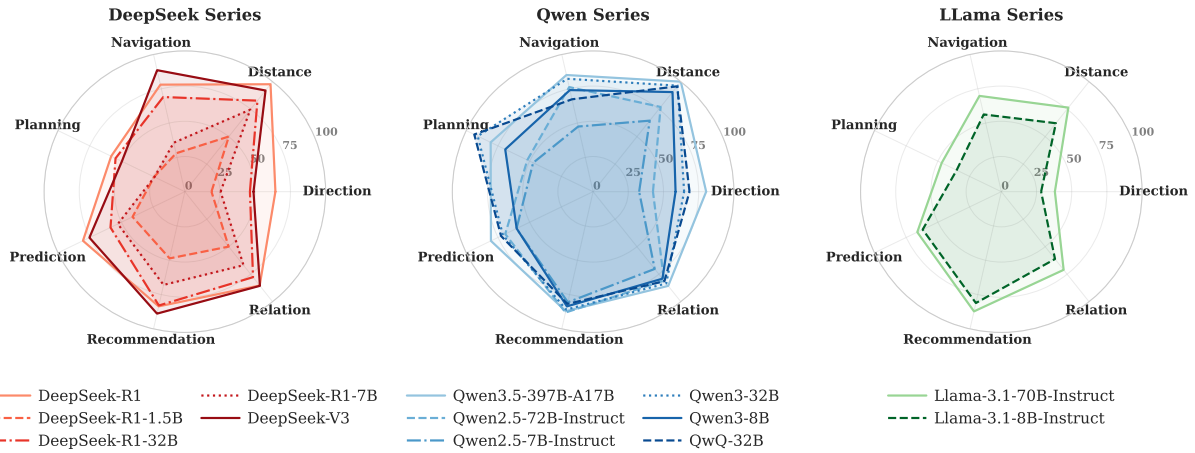


Figure 4: Performance breakdown across seven reasoning tasks for the DeepSeek, Qwen, and Llama model series.

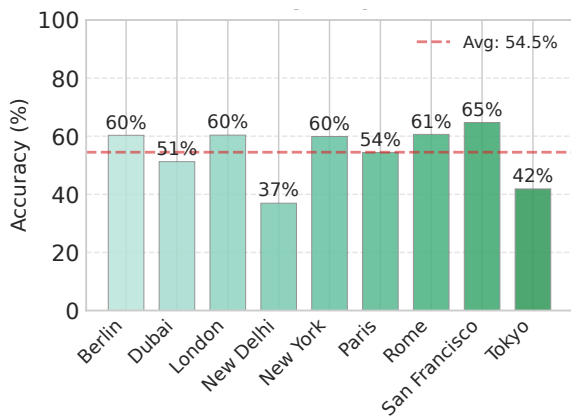


Figure 5: GPT-4o Knowledge Module accuracy by city.

nates, resulting in mistakes when inferring spatial relationships; and (3) **Connectivity Failure**, especially in *Navigation* tasks, where models hallucinate non-existent connections between coordinate nodes. Taken together, these failure modes highlight the vulnerability of current geospatial reasoning capabilities. Appendix E presents example model traces for each failure mode, along with the corresponding rubric-based evaluations produced by our judge.

Semantic Complexity of Reasoning Trajectories

Our analysis of model outputs shows that geospatial reasoning exhibits complex, high-dimensional semantic patterns that inflexible, programmatic verification cannot fully capture: (1) **Trajectory Diversity**, geospatial tasks often admit multiple equally valid reasoning paths (e.g., tackling a *Relation* problem by explicitly computing slopes or by using more visual-topological reasoning). Deterministic parsers systematically penalize this legitimate semantic variability. (2) **Heuristic Validity**,

in NP-hard settings like *Planning*, effective reasoning relies primarily on strategic heuristics, rather than by exhaustive search over all permutations. We observe that LLM-Judge can still recognize correct reasoning when program-based metrics would reject it due to minor numerical discrepancies. (3) **Structural Fluidity**, semantic evaluation decouples reasoning quality from output format, emphasizing the underlying rationale rather than surface syntax and thus avoiding false negatives for models that reason correctly but deviate from rigid templates.

Foundational Skills as Building Blocks

Our analysis confirms that sophisticated geospatial reasoning is fundamentally grounded in basic geospatial abilities. The Pearson correlation matrix (Figure 6) shows strong positive associations between foundational and advanced tasks, consistent with intuitive expectations about spatial cognition. In particular, *Direction* emerges as crucial for *Planning*, since accurate orientation judgments govern the selection of optimal routes. Likewise, *Distance* and *Relation* function as prerequisites for *Prediction* and *Navigation*: the former involves determining whether a target lies along the extrapolated path within a plausible range, while the latter requires grasping road connectivity to identify shortest paths. Taken together, these interdependencies indicate that a curriculum learning strategy that emphasizes early mastery of foundational skills is key to training robust geo-aware models.

5.3 Validation of the LLM Judge

Human-LLM Consistency Three domain experts evaluated 90 stratified samples. Our GPT-

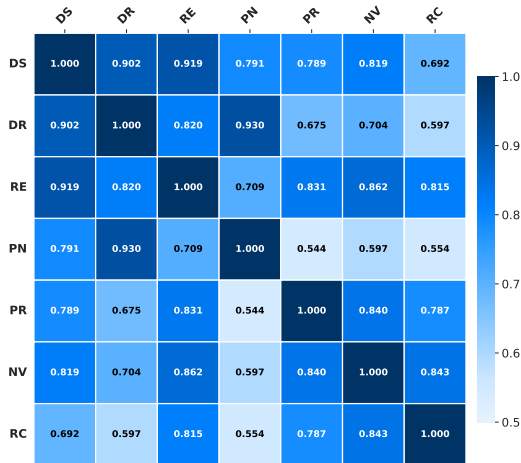


Figure 6: Pearson correlation matrix of model average total scores across the seven reasoning tasks.

4o-based judge exhibited a high Pearson correlation with the aggregated human scores ($r = 0.86$, $p < 10^{-25}$), exceeding the inter-annotator correlation among the experts ($r = 0.81$).

Judge Robustness To mitigate potential evaluator bias, we re-run our evaluation with Qwen3-32B as an alternative judge across models spanning all series and scale tiers. The resulting rankings were nearly identical to the original GPT-4o-based ordering (Spearman $\rho = 0.98$), indicating that our evaluation protocol is largely model-agnostic. Moreover, a pure LLM-based evaluation strongly correlates with our hybrid programmatic framework, confirming that advanced judges can reliably approximate strict verification scripts (see Appendix F).

6 Conclusion

This paper introduces a benchmark that decouples urban geospatial knowledge from reasoning through a detailed task taxonomy. Our framework highlights recurring weaknesses in current LLMs, such as geospatial biases, coarse-grained knowledge representations, and uneven performance on complex spatial reasoning tasks. Comprehensive analysis further demonstrates that our benchmark and evaluation protocol can reliably assess LLMs’ capabilities in urban geospatial contexts.

Limitations

The current benchmark is text-only, presenting opportunities to incorporate real-time multimodal inputs, and to further expand the geographic scope to rural and semi-urban environments. Furthermore, our scalable evaluation framework, which relies on

OpenStreetMap, could be further enhanced by integrating varied data sources. 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 annotators based on publicly available information from Internet sources. The annotation team, composed of a diverse group of well-educated individuals aged 18–30, was recruited through a professional data annotation platform. We strictly adhered to fair labor standards by compensating all participants, including both annotators and quality assurance personnel, at a rate of \$7.16 USD per hour. Given the flexibility of this remote task and the economic context of the participants’ residential regions, this rate exceeds the local average hourly wage, ensuring a competitive and generous remuneration. 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. An ethical contribution of this study is its explicit uncovering of geospatial biases in existing LLMs, we caution that deploying current LLMs in location-based services without bias-aware adaptation may amplify regional inequities; our benchmark is intended as a diagnostic tool rather than a deployment-readiness score.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 42394060 and 42394064, Ant Group Research Fund, and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

References

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.
- Danish Contractor, Krunal Shah, Aditi Partap, Parag Singla, and Mausam Mausam. 2021. [Answering poi-recommendation questions using tourism reviews](#). In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21*, page 281–291, New York, NY, USA. Association for Computing Machinery.
- 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, Wenye 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, Jiayi 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 Dairani, Julian Michael, and Samuel R. Bowman. 2023. [Gpqa: A graduate-level google-proof qa benchmark](#). *Preprint*, arXiv:2311.12022.
- Jonathan Roberts, Timo Lüddecke, 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. [Spatialrag: 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 Data Construction Details

A.1 Distractor Generation Strategies

To make sure that Knowledge Module rigorously evaluates LLMs’ underlying urban knowledge rather than superficial pattern recognition, we employ four principled strategies for generating plausible but incorrect distractor options. Each strategy is designed to probe a distinct facet of spatial knowledge representation. Table 6 summarizes which strategies apply to each task template.

A.1.1 Spatial Proximity

This strategy exploits geographic proximity to generate confounding options. Distractors are sampled from entities that are spatially close to the correct answer, testing whether models can distinguish between nearby but distinct locations.

Retrieve_POI_Address Distractors are addresses of other POIs located within a 500-meter radius of the target POI. By selecting geographically proximate alternatives, we ensure that simple heuristics based on neighborhood co-occurrence cannot reliably identify the correct answer.

Retrieve_Coordinates Given the correct coordinates (lng, lat) , we generate four distractors by applying fixed offsets of approximately 1 km in each cardinal direction:

$$\{(lng \pm \delta, lat \pm \delta)\}, \quad \delta = 0.01 \quad (1)$$

This tests fine-grained coordinate resolution capabilities.

Infer_POI_From_Scene & Find_Nearest_POI

For scene inference tasks, distractors are POIs sampled from locations at least 500–1000 meters away from the target, specifically from four corner regions surrounding the query location. This spatial separation ensures that distractors represent plausible but spatially inconsistent alternatives.

A.1.2 Semantic Similarity

This strategy generates distractors that are semantically related but factually incorrect, testing the model’s ability to disambiguate between similar entities.

Retrieve_POI_Name & Retrieve_AOI_Name

Distractors are randomly sampled names from the global pool of POIs or AOIs, excluding duplicates. We enforce uniqueness constraints to prevent trivial elimination:

- All distractor names must be distinct from each other and from the correct answer.
- If duplicate names exist in the candidate pool, we resample until uniqueness is achieved.

Retrieve_AOI_Address Similar to name retrieval, we sample addresses from other AOIs while ensuring no address appears more than once among the options.

Find_Cover_AOI & Find_Cover_AOI_All

Distractors are AOIs that do not contain the queried POI(s). For `Find_Cover_AOI`, we select POIs from outside the target AOI. For `Find_Cover_AOI_All`, we sample AOIs that fail to cover all specified POIs, with iterative resampling to guarantee name uniqueness.

Classify_POI & Find_Instances_By_Type For category classification, distractors are drawn from the set of similar POI categories excluding the correct one. For instance retrieval, distractors are POIs belonging to similar categories different from the target type.

Landuse_AOI We define 13 canonical land-use types (e.g., *Residential*, *Commercial*, *Education*, *Park&GreenLand*). Distractors are randomly sampled from this predefined taxonomy, excluding the ground-truth type.

A.1.3 Numerical Variation

For tasks requiring quantitative reasoning, distractors are generated by systematically perturbing numerical values within controlled bounds.

Calculate_Distance Given the true geodesic distance d^* , distractors are sampled uniformly from $[0.5, 25]$ km subject to:

$$0.8 < |d - d^*| \leq 8 \quad (\text{in km}) \quad (2)$$

This constraint ensures distractors are neither trivially distinguishable (too far from d^*) nor ambiguously close.

Query_Length Road length distractors are generated as multiplicative variants of the true length L^* :

$$\{\max(0.1L^*, 100), 2L^*, 3L^*\} \quad (3)$$

The lower bound of 100 meters prevents implausibly short options.

Compare_POI_Count, Compare_Area_Size & Compare_Type_Diversity For comparative tasks, we employ an iterative sampling procedure (up to 1000 attempts) to ensure:

1. All four candidate AOIs have unique names.
2. Exactly one AOI achieves the maximum value for the queried metric (no ties).

This eliminates ambiguous scenarios where multiple options could be considered correct.

A.1.4 Attribute Substitution

This strategy generates distractors by substituting related but incorrect attributes, testing whether models correctly associate entities with their properties.

Find_Junctions Distractors are fabricated junction names constructed by combining the target road with randomly selected other road names. We verify that generated junction names do not accidentally match real junctions along the target road.

Identify_Connections The distractor generation depends on the connectivity of the target road.

A.2 Human Annotation Process

Recruitment and Ethics We recruited 15 professional annotators through a qualified data annotation platform. The team consisted of well-educated individuals (undergraduate degree or higher) aged 18–30. All annotators were compensated fairly for their labor, adhering to ethical research standards. Importantly, the data collection utilized public internet sources (e.g., OpenStreetMap, Google Maps) and strictly avoided any private or personally identifiable information (PII).

Real-World Scenario Sourcing To distinguish UrbanGeoEval from datasets generated via synthetic templates, we prioritized scenarios rooted in genuine user needs.

- **Web-Based Query Extraction:** Beyond formulating intents based on personal experience, annotators were explicitly instructed to consult external real-world sources. They extracted geospatial scenarios and common user inquiries from global travel Q&A websites, tourism forums, and online travel guides. This ensures that the benchmark questions reflect actual information-seeking behaviors found in urban environments rather than artificial constructs.

- **Real-World Map Verification:** Annotators utilized real-world map interfaces to validate these scenarios. They manually extracted the necessary geospatial context (coordinates, road networks, opening hours) to ensure the problem was solvable using only the provided information.

A.3 Quality Control Protocols

We implemented a quality control workflow to guarantee the correctness of both the geospatial data and the reasoning logic.

Visual-Aided Verification To reduce spatial hallucinations during annotation, we implemented a visual verification step. For each annotated record, annotators were instructed to attach a screenshot of the ground-truth map (for instance, the actual route from a mapping service or the precise POI location). This visual evidence enabled validators to quickly compare the text-based coordinates and reasoning process against the real-world reference.

Dual-Layer Validation Mechanism Our validation process followed a strict hierarchy:

1. **100% Full Inspection:** The professional annotation platform employed a dedicated Quality Assurance (QA) team. Every single record underwent a preliminary check to verify formatting and basic logical consistency.
2. **15% Acceptance Sampling:** In the final acceptance phase, we performed a random sampling inspection of 15% of each submitted batch. We established a strict accuracy threshold of 90%. If the accuracy of the sampled subset fell below this threshold, the *entire* batch was rejected and returned to the annotation team for a complete rework.

Reasoning Logic Check Beyond data accuracy, we specifically verified the validity of the reasoning chains. For objective tasks (e.g., Distance, Planning), scripts were used to verify geometric correctness. For subjective tasks, reviewers ensured that the provided solution was not only correct but followed a logical human-like thought process given the constraints.

B Details on Experimental Setup

B.1 Evaluated Models

Our evaluation includes closed-source models (GPT-4o [OpenAI et al., 2024](#), o3-mini [OpenAI](#),

Dimension	Template ID	Spatial	Semantic	Numerical	Attribute
Node	Retrieve_POI_Address	✓			
	Retrieve_POI_Name		✓		
	Retrieve_Coordinates	✓			
	Calculate_Distance			✓	
	Calculate_Distance_Context			✓	
Area	Retrieve_AOI_Address		✓		
	Retrieve_AOI_Name		✓		
	Find_Cover_AOI		✓		
	Find_Cover_AOI_All		✓		
Road	Query_Length			✓	
	Find_Junctions				✓
	Identify_Connections				✓
Type	Classify_POI		✓		
	Landuse_AOI		✓		
	Find_Instances_By_Type		✓		
Function	Compare_POI_Count			✓	
	Compare_Area_Size			✓	
	Compare_Type_Diversity			✓	
Surroundings	Infer_POI_From_Scene	✓			
	Find_Nearest_POI	✓			

Table 6: Mapping of distractor generation strategies to task templates. Each task employs one primary strategy designed to probe specific aspects of geographic knowledge. **Spatial**: distractors based on geographic proximity; **Semantic**: distractors based on entity similarity; **Numerical**: distractors based on value perturbation; **Attribute**: distractors based on property substitution.

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-R1⁴ (DeepSeek-AI et al., 2025) and Qwen3 (Yang et al., 2025a) series. All details about the evaluated LLMs are presented in Table 7.

B.2 Implementation Details

We accessed close-source LLMs via their official APIs. Open-source models were deployed locally using the vLLM engine (Kwon et al., 2023) on servers 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 Reasoning Module, we used model-specific recommended configurations for each LLMs and a temperature of 0.6 for our LLM-as-a-Judge, GPT-4o.

C Evaluation of Reasoning Module Tasks

This section provides details on the prompts used to guide model outputs, the scoring rubrics used

⁴Smaller 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).

by the Judge-LLM, and the unified programmatic answer-scoring formulas used by our Hybrid evaluation framework for the geospatial reasoning tasks.

C.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.

C.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. We collaborated with an enterprise specializing in geospatial daily life services, and their experts worked with us to curate these scoring rubrics to capture acceptable heuristic logic. 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.

C.3 Programmatic Answer-Scoring Formulas

For tasks with deterministic ground truth, the final-answer score $s_{\text{ans}} \in [0, 10]$ is computed by scripts. Each formula is bounded below by 0.

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
Llama	DeepSeek-R1-32B	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B
	Llama-3.1-8B-Instruct	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
Qwen	Llama-3.1-70B-Instruct	https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct
	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
MiniMax	MiniMax-M2.5	https://huggingface.co/MiniMaxAI/MiniMax-M2.5
Google	Gemma-3-27B	https://huggingface.co/google/Gemma-3-27B-it
Microsoft	Phi-4	https://huggingface.co/microsoft/Phi-4

Table 7: Detailed information of the evaluated models.

Distance. Given a model-predicted distance d_{pred} and the ground-truth great-circle distance d_{gt} (both in km), the score uses an absolute-error penalty:

$$s_{\text{ans}}^{\text{DS}} = \max(0, 10 - |d_{\text{pred}} - d_{\text{gt}}|), \quad (4)$$

rounded to two decimals. No scaling factor is applied because at city scale a 1 km error is already a meaningful loss and a 10 km error is operationally useless.

Direction. Given a model-predicted bearing θ_{pred} and the ground-truth bearing θ_{gt} (both in degrees, measured clockwise from true north), let $\Delta\theta$ denote the circular bearing error:

$$\Delta\theta = \min(|\theta_{\text{pred}} - \theta_{\text{gt}}|, 360 - |\theta_{\text{pred}} - \theta_{\text{gt}}|). \quad (5)$$

Let $c_{\text{pred}}, c_{\text{gt}}$ denote the predicted and ground-truth descriptive cardinal directions (e.g., *Northeast*, *East-Southeast*). The bearing score is then

$$s_{\text{bearing}} = 10 - 0.25 \cdot \Delta\theta, \quad (6)$$

and the cardinal-direction penalty is

$$p_{\text{card}} = \begin{cases} -2, & c_{\text{pred}} \neq c_{\text{gt}}, \\ 0, & c_{\text{pred}} = c_{\text{gt}}. \end{cases} \quad (7)$$

The final answer score is the floor-clipped sum,

$$s_{\text{ans}}^{\text{DR}} = \max(0, s_{\text{bearing}} + p_{\text{card}}), \quad (8)$$

again rounded to two decimals. The coefficient 0.25 is calibrated to a 16-point compass system (N, NNE, NE, ENE, . . .) in which each sector spans 22.5° : an error within the same sector ($\Delta\theta < 10^\circ$) yields high credit ($> 7.5/10$), an error of one full sector ($\sim 22.5^\circ$) yields moderate credit ($\sim 4.4/10$), and an error of $\sim 40^\circ$ (close to two sectors) drives the bearing score to zero, reflecting that two-sector deviations are categorically misleading for human navigation. The cardinal-direction penalty is applied *before* the global $\max(0, \cdot)$ floor, so the final answer score is always non-negative and bounded above by 10.

Relation, Prediction. For categorical and entity-matching tasks, we use Exact Match: $s_{\text{ans}}^{\text{RE/PR}} = 10$ if the model’s output matches the ground-truth label exactly, otherwise 0.

Planning. Let σ_{gt} be the ground-truth optimal visiting order and σ_{pred} the model’s predicted order, both as sequences of normalized location names. The answer score uses a per-position Hamming-style penalty,

$$s_{\text{ans}}^{\text{PN}} = \max(0, 10 - 2 \cdot d_{\text{H}}(\sigma_{\text{pred}}, \sigma_{\text{gt}})), \quad (9)$$

where $d_{\text{H}}(\cdot, \cdot)$ is the number of misaligned positions.

D 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.

D.1 Detailed Performance on the Knowledge Module

D.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`.

D.1.2 Cross-City Performance

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

D.1.3 Full Model-by-City Accuracy Matrix

To complement the single-model view in Figures 22–30 and the GPT-4o bias chart in the main paper (Figure 5), Table 8 reports the Knowledge Module accuracy of all five evaluated models across the full set of 18 cities. The geographic disparity revealed here is both substantial and systematic: cities in the Global South and South Asia (Cairo, Rio de Janeiro, Buenos Aires, New Delhi) rank among the weakest for *every* evaluated model, while North American and Western European cities (San Francisco, New York, London, Paris) consistently achieve the highest scores. GPT-4o’s accuracy alone spans 36.94% (New Delhi) to 64.72%

(San Francisco)—a spread of nearly 28 points—and comparable ranges are observed across all other models. The cross-provider consistency of this pattern points to a systemic bias in mainstream pre-training corpora.

D.2 Detailed Performance on the Reasoning Module

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

D.2.1 Overall Reasoning Capability Landscape

Figure 32 provides a consolidated radar chart of the total scores for all 18 evaluated models across the seven reasoning tasks.

D.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 trade-off between reasoning performance and computational efficiency.

E 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.

F Details on Judge-LLM Robustness

To ensure the reliability of our evaluation framework, we conducted extensive ablation studies focusing on the scoring mechanism and the choice of the judge model.

F.1 Hybrid vs. Pure LLM-as-a-Judge Evaluation

Our primary evaluation employs a Hybrid Framework, where the *Reasoning Process* is scored by an LLM (GPT-4o) and the *Final Answer* is scored programmatically (using scripts for deterministic tasks like `Distance` and `Relation`). We compared this against a Pure LLM-as-a-Judge approach, where GPT-4o scores both the reasoning and the answer end-to-end.

As shown in Table 9, the Pure LLM Judge tends to be slightly more lenient, resulting in consistently

City	GPT-4o	Llama-3.1-70B	Llama-3.1-8B	Qwen2.5-72B	Qwen2.5-7B	City Avg.
Auckland	59.15	50.01	40.82	51.10	45.03	49.22
Bangkok	57.65	45.51	38.42	49.24	40.14	46.19
Berlin	60.32	47.92	36.73	47.73	41.92	46.92
Buenos Aires	53.03	43.14	35.17	44.29	37.92	42.71
Cairo	49.47	39.03	38.88	42.35	34.62	40.87
Dubai	51.24	42.14	30.17	44.19	37.50	41.05
London	60.37	47.82	34.68	50.37	41.84	47.02
Moscow	52.95	42.83	33.69	44.13	36.46	42.01
Nairobi	55.64	45.37	36.32	45.99	40.81	44.83
New Delhi	36.94	44.28	36.38	48.54	40.51	41.33
New York	59.89	47.92	32.95	50.22	42.02	46.60
Paris	54.41	50.02	33.91	51.28	43.28	46.58
Rio de Janeiro	49.92	39.14	33.42	44.43	36.60	40.70
Rome	60.62	48.24	35.71	48.23	42.72	47.10
San Francisco	64.72	48.88	33.89	50.79	44.17	48.49
Sydney	60.74	49.23	40.48	48.84	42.86	48.43
Taipei	53.66	46.38	37.32	46.65	39.20	44.64
Tokyo	41.85	44.96	37.15	49.15	42.66	43.15
Model Avg.	54.59	45.71	35.89	47.64	40.57	44.88

Table 8: Knowledge Module accuracy (%) for each of the five evaluated models on all 18 cities. Each cell is the unweighted mean across the 20 question templates for that city. The bottom row reproduces the Overall column of Table 4; the rightmost column reports the cross-model mean per city.

higher scores (an average inflation of ~ 1.5 points). The Hybrid mechanism imposes rigorous programmatic checks on answers, which means it penalizes responses that only appear correct in format or are close approximations, which an LLM might otherwise pass over as acceptable. The strong alignment between the two methods ($r = 0.9982$) confirms that the Hybrid approach provides a stricter, more rigorous lower bound for model performance without altering the relative landscape.

Model	Pure LLM Judge	Hybrid Framework	Diff
DeepSeek-R1	81.99	79.90	-2.09
DeepSeek-R1-1.5B	41.53	37.61	-3.92
DeepSeek-R1-32B	68.24	67.48	-0.76
DeepSeek-R1-7B	52.39	49.51	-2.88
DeepSeek-V3	76.61	75.51	-1.09
Gemma-3-27B	68.50	66.48	-2.02
GPT-4o	74.53	74.53	+0.00
Llama-3.1-70B-Inst.	64.99	63.86	-1.14
Llama-3.1-8B-Inst.	53.83	53.30	-0.53
o3-mini	87.96	87.50	-0.46
Phi-4	71.56	69.31	-2.25
Qwen2.5-72B-Inst.	70.24	68.85	-1.39
Qwen2.5-7B-Inst.	58.91	56.87	-2.04
Qwen3-32B	84.45	83.76	-0.70
Qwen3-8B	76.38	74.55	-1.83
QwQ-32B	83.40	82.58	-0.82

Table 9: Comparison of total scores (normalized to 100) between the Pure LLM Judge and our proposed Hybrid Framework. The Hybrid approach systematically reduces scores by enforcing strict answer verification.

F.2 Judge Model Consistency

To investigate potential bias in using GPT-4o as the primary judge, we re-ran the evaluation with Qwen3-32B as an alternative judge on models spanning all series and scale tiers.

The two judges exhibited a near-perfect Spearman rank correlation of $\rho = 0.98$. Table 10 details the ranking changes. Only minor local swaps occurred, while the overall tiers remained stable. This stability confirms that our evaluation criteria are objective and not dependent on the specific preferences of a single proprietary model.

G Comparison with Related Benchmarks

Our benchmark fills a critical gap: evaluating intrinsic, city-scale spatial logic without relying on external tools or reading comprehension. Furthermore, unlike prevalent MCQ datasets, our benchmark prioritizes Open-ended Generation and enforces rigorous Quality Control. Uniquely, UrbanGeoEval implements Decoupled Evaluation, explicitly separating memory recall from spatial reasoning to avoid confounding errors in model diagnosis. A detailed comparison with existing works is provided in Table 11.

Model	GPT-4o Rank	Qwen3 Rank	Diff	Model	GPT-4o Rank	Qwen3 Rank	Diff
o3-mini	1	1	–	Qwen2.5-72B	9	8	↑ 1
Qwen3-32B	2	4	↓ 2	Gemma-3-27B	10	10	–
QwQ-32B	3	3	–	Llama-3.1-70B	11	11	–
DeepSeek-R1	4	2	↑ 2	Qwen2.5-7B	12	12	–
DeepSeek-V3	5	5	–	Llama-3.1-8B	13	13	–
Qwen3-8B	6	6	–	DS-R1-7B	14	14	–
GPT-4o	7	7	–	DS-R1-1.5B	15	15	–
Phi-4	8	9	↓ 1				

Table 10: Rank stability comparison between GPT-4o Judge and Qwen3-32B Judge across models spanning all series and scale tiers.

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 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 latitude ($\Delta\phi$) and longitude ($\Delta\lambda$) in radians.

1~2: Minor mistake in one of the difference values or no calculation 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 calculation 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 = $\max(0, 10 - |\text{model_answer} - \text{ground_truth}|)$, 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 (Δ long) 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 = $\max(0, 10 - (|\text{model answer} - \text{ground truth bearing}|) * 0.25)$, 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 (10 points)

10: Selects the destination that matches the ground truth answer

0: Selects an incorrect destination that is not 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 segments

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.

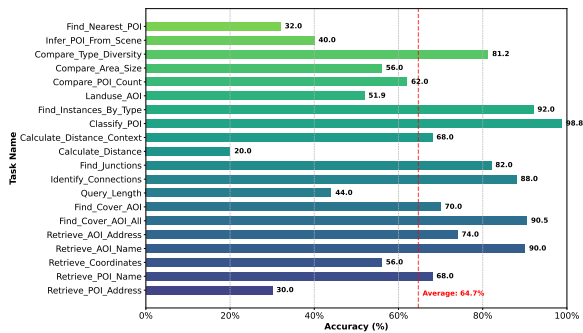


Figure 18: GPT-4o's Knowledge Module accuracy distribution across question templates on San Francisco.

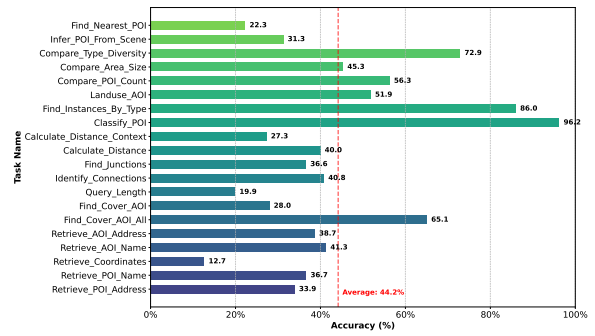


Figure 22: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on San Francisco.

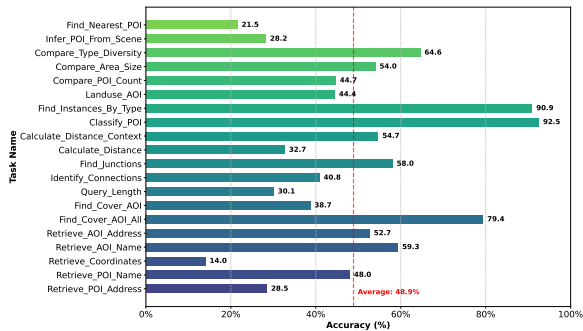


Figure 19: Llama-3.1-70B-Instruct's Knowledge Module accuracy distribution across question templates on San Francisco.

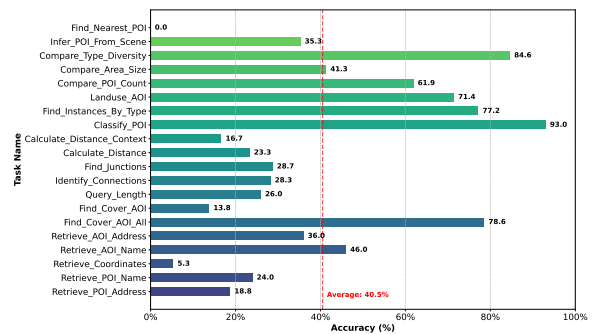


Figure 23: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on New Delhi.

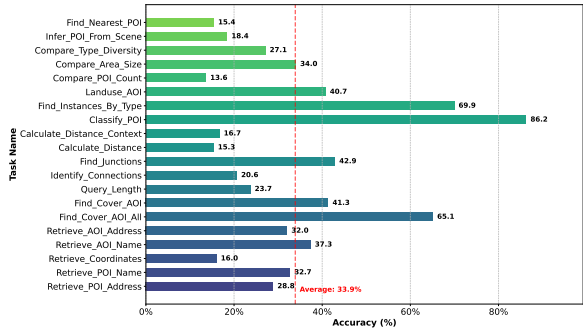


Figure 20: Llama-3.1-8B-Instruct's Knowledge Module accuracy distribution across question templates on San Francisco.

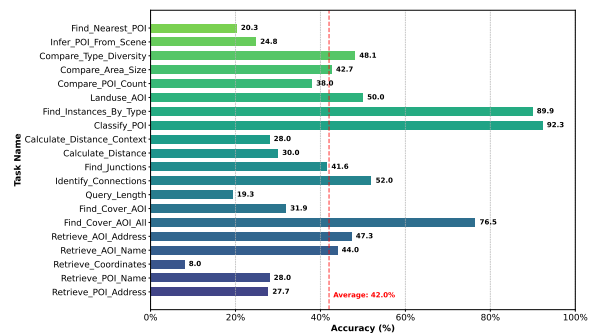


Figure 24: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on New York.

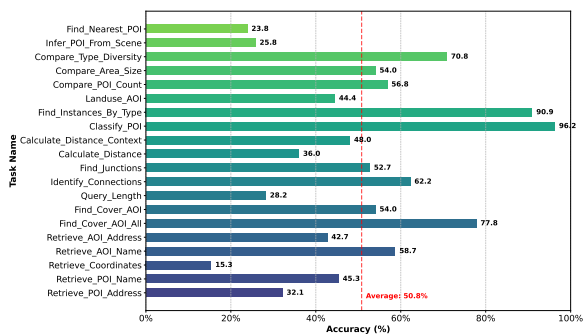


Figure 21: Qwen2.5-72B-Instruct's Knowledge Module accuracy distribution across question templates on San Francisco.

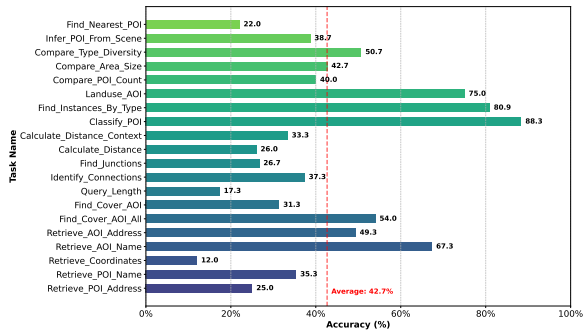


Figure 25: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on Tokyo.

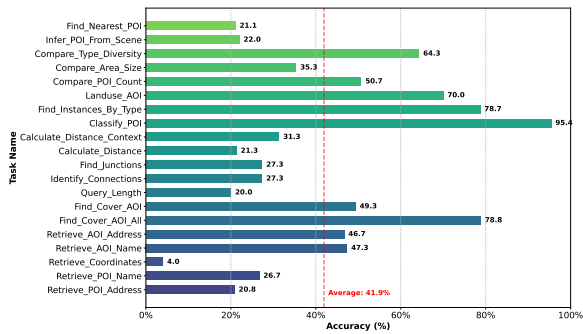


Figure 26: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on Berlin.

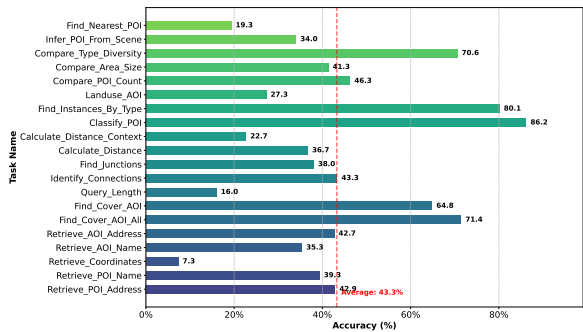


Figure 27: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on Paris.

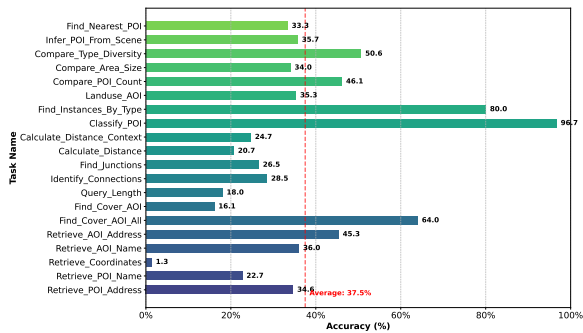


Figure 28: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on Dubai.

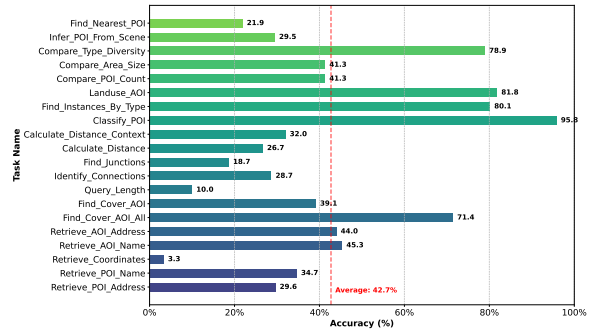


Figure 29: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on Rome.

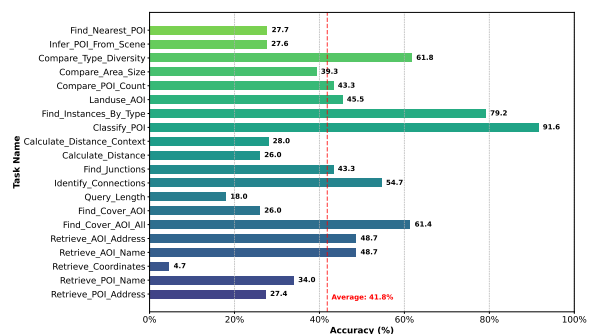


Figure 30: Qwen2.5-7B-Instruct's Knowledge Module accuracy distribution across question templates on London.

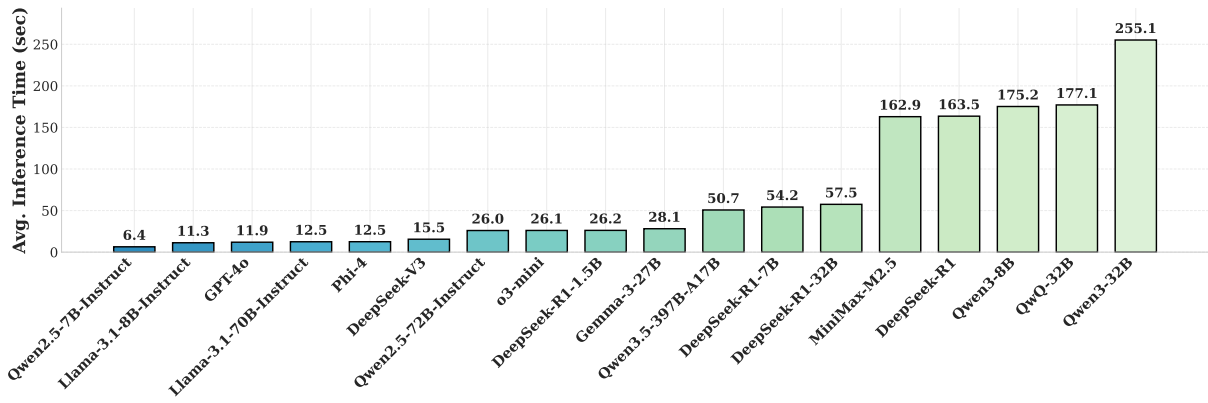


Figure 31: Comparison of average inference time per query across different models on Reasoning Module.

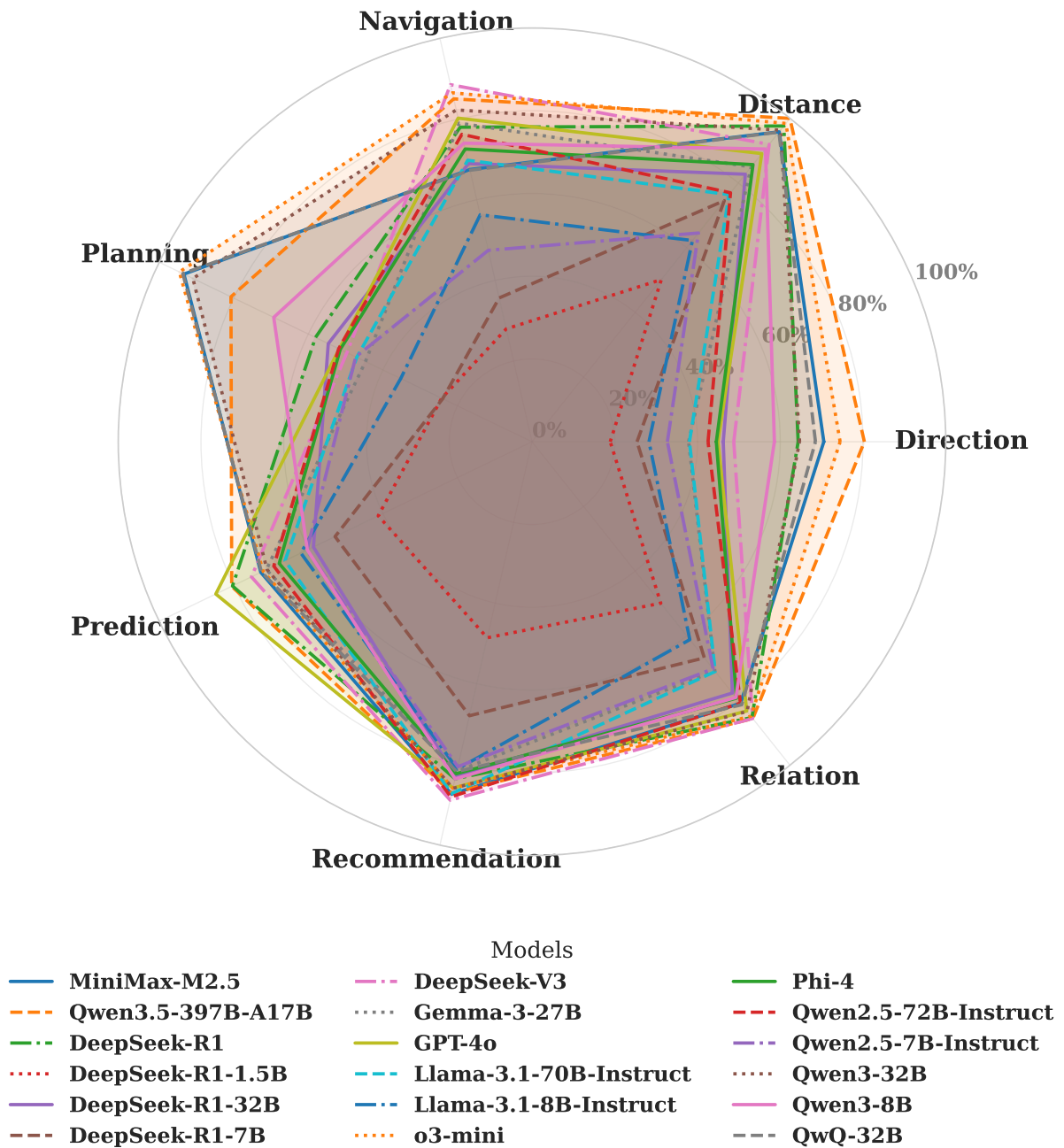


Figure 32: Radar chart of urban geospatial reasoning capabilities on all evaluated models.

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: $38.8975^\circ \text{ N} = 38.8975 \times \pi/180 \approx 0.6790$ radians

- Longitude of The White House: $77.0364^\circ \text{ W} = -77.0364 \times \pi/180 \approx -1.3444$ radians

- Latitude of The Pentagon: $38.8717^\circ \text{ N} = 38.8717 \times \pi/180 \approx 0.6785$ radians

- Longitude of The Pentagon: $77.0561^\circ \text{ W} = -77.0561 \times \pi/180 \approx -1.3449$ 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.

- $a = \sin^2(\Delta\text{lat}/2) + \cos(\text{lat}_1) \times \cos(\text{lat}_2) \times \sin^2(\Delta\text{lon}/2)$

- $c = 2 \times \text{atan2}(\sqrt{a}, \sqrt{1-a})$

- $d = R \times c$

Where R is the radius of the Earth (mean radius = 6,371 km).

Step 4: Calculate the intermediate values.

- $a = \sin^2(0.0005/2) + \cos(0.6790) \times \cos(0.6785) \times \sin^2(0.0005/2)$

- $a = \sin^2(0.00025) + \cos(0.6790) \times \cos(0.6785) \times \sin^2(0.00025)$

- $a \approx 0.0000000625 + 0.7915 \times 0.7917 \times 0.0000000625$

- $a \approx 0.0000000625 + 0.0000000486$

- $a \approx 0.000001111$

- $c = 2 \times \text{atan2}(\sqrt{0.000001111}, \sqrt{1-0.000001111})$

- $c \approx 2 \times \text{atan2}(0.001054, 0.9999999444)$

- $c \approx 2 \times 0.001054$

- $c \approx 0.002108$ 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: $38^\circ 53' 51'' \text{ N}, 77^\circ 2' 11'' \text{ W}$ is correctly represented as $38.8975^\circ \text{ N}, 77.0364^\circ \text{ W}$.

The Pentagon: $38^\circ 52' 18'' \text{ N}, 77^\circ 3' 22'' \text{ W}$ is correctly represented as $38.8717^\circ \text{ N}, 77.0561^\circ \text{ W}$.

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: $\text{radian} = \text{degree} \times \pi/180$.

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 based on the model's rounded radian values. Score: 3/3

Haversine Formula Application (5 points):

The model correctly states the Haversine formula and attempts to apply it.

However, there is a significant calculation error in Step 4.

The model calculates $a \approx 0.000001111$. The next step requires \sqrt{a} .

$\sqrt{0.000001111} \approx 0.000333$, but the model calculates the intermediate term as 0.001054 (implied by the resulting c).

This arithmetic error propagates to the final result ($c \approx 0.002108$ instead of ≈ 0.00052), causing the distance to be approximately 4 times larger than it should be.

Score: 2/5

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: 12/15

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: $\text{Score} = \max(0, 10 - 10.09)$, the result is 0. Score: 0/10

Final Evaluation:

The reasoning process is structured correctly, but it contains a critical arithmetic error during the execution of the Haversine formula (Step 4), which leads to an incorrect final answer. Consequently, points are deducted from the reasoning section, and the answer accuracy score is 0.

Reason: [[12]] **Answer:** [[0]] **Total:** [[12]]

Figure 33: Case of Distance task.

Benchmark	Knowledge Probing	Spatial Reasoning	Raw-geometry Computation	Itinerary Planning	Open-ended Generation	Quality Control	Decoupled Evaluation
GeoLLM	✓	✗	✗	✗	✗	✗	✗
TourismQA	✗	✓	✗	✗	✗	✓	✗
GeoSQA	✓	✓	✗	✗	✗	✓	✗
STBench	✓	✓	✓	✗	✗	✗	✗
MapEval	✓	✓	✗	✓	✗	✓	✗
CityEval	✓	✓	✗	✓	✗	✗	✗
UrbanGeoEval (Ours)	✓	✓	✓	✓	✓	✓	✓

Table 11: Comparison of UrbanGeoEval with existing GeoQA benchmarks.

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° 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 = $\max(0, 10 - (1.44 \times 0.25)) = 10 - 0.36 = 9.64 \rightarrow$ Rounded to 9.64

Descriptive Direction: Both model and ground truth say East-Southeast (ESE) — no penalty.

Answer Subtotal: 9.64/10

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:** [[9.64]] **Total:** [[23.64]]

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: $m = (y_2 - y_1) / (x_2 - x_1)$
- For Columbus Ave segments:
 - Segment 1: $m_1 = (-73.98234 + 73.98428) / (40.772545 - 40.76986)$
 - Segment 2: $m_2 = (-73.98075 + 73.98234) / (40.77470 - 40.772545)$
 - Segment 3: $m_3 = (-73.97928 + 73.98075) / (40.77674 - 40.77470)$
 - Segment 4: $m_4 = (-73.97750 + 73.97928) / (40.77923 - 40.77674)$
- For Broadway segments:
 - Segment 1: $m_5 = (-73.98206 + 73.98199) / (40.77702 - 40.77909)$
 - Segment 2: $m_6 = (-73.98215 + 73.98206) / (40.77470 - 40.77702)$
 - Segment 3: $m_7 = (-73.98217 + 73.98215) / (40.77180 - 40.77470)$
 - Segment 4: $m_8 = (-73.98207 + 73.98217) / (40.76865 - 40.77180)$

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:

Here are five locations in London: The Northall, Trafalgar Square, London Bridge, London Eye, Big Ben

The Northall->Trafalgar Square:

Bearing: 298.34 degrees

Direction: West-Northwest

Distance: 0.32 kilometers

Trafalgar Square->London Bridge:

Bearing: 90.29 degrees

Direction: East

Distance: 2.79 kilometers

London Bridge->London Eye:

Bearing: 256.65 degrees

Direction: West-Southwest

Distance: 2.26 kilometers

London Eye->Big Ben:

Bearing: 232.00 degrees

Direction: Southwest

Distance: 0.45 kilometers

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: 6/10 (The order is incorrect, and the total distance is significantly different)

Reason: [[8]] **Answer:** [[6]] **Total:** [[14]]

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)

The model's final prediction is D. Shanghai Pudong International Airport, which exactly matches the ground-truth destination, the answer score is 10/10.

Score: 10/10

Final Score:

Reasoning: [20/20]

Answer: [10/10]

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) → (47.6420, -122.3215)

2. E Lake Washington Blvd: (47.6438, -122.2991) → (47.6260, -122.2925)

3. Portage Bay Viaduct: (47.6420, -122.3215) → (47.6438, -122.2991)

4. E Madison St: (47.6260, -122.2925) → (47.6354, -122.2782)

5. Boyer Ave E: (47.6420, -122.3215) → (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):

o 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.

o 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 of 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.