TurnBack: A Geospatial Route Cognition Benchmark for Large Language Models through Reverse Route

Hongyi Luo[1,2] **Qing Cheng**[1,2] **Daniel Matos**[1] **Hari Krishna Gadi**[1]
Yanfeng Zhang[1] **Lu Liu**[1] **Yongliang Wang**[1] **Niclas Zeller**[3]
Daniel Cremers[2,4] **Liqiu Meng**[2]
[1]Huawei Riemann Lab* [2]TU München [3]Hochschule Karlsruhe [4]MCML
{hongyi.luo, qing.cheng1, daniel.matos, hari.krishna.gadi1}@huawei.com
{zhangyanfeng8, luliu1, wangyongliang775}@huawei.com
niclas.zeller@h-ka.de, {cremers, liqiu.meng}@tum.de

Abstract

Humans can interpret geospatial information through natural language, while the geospatial cognition capabilities of Large Language Models (LLMs) remain underexplored. Prior research in this domain has been constrained by non-quantifiable metrics, limited evaluation datasets and unclear research hierarchies. Therefore, we propose a large-scale benchmark and conduct a comprehensive evaluation of the geospatial route cognition of LLMs. We create a large-scale evaluation dataset comprised of 36000 routes from 12 metropolises worldwide. Then, we introduce PathBuilder, a novel tool for converting natural language instructions into navigation routes, and vice versa, bridging the gap between geospatial information and natural language. Finally, we propose a new evaluation framework and metrics to rigorously assess 11 state-of-the-art (SOTA) LLMs on the task of route reversal. The benchmark reveals that LLMs exhibit limitation to reverse routes: most reverse routes neither return to the starting point nor are similar to the optimal route. Additionally, LLMs face challenges such as low robustness in route generation and high confidence for their incorrect answers. Code & Data available here: TurnBack.

1 Introduction

Geospatial cognition is crucial for enabling LLMs to perform advanced route navigation and urban planning, such as "Take me home, pass by any supermarket, and find a mailbox within 500m of it." Humans naturally complete such tasks by relying on their innate geospatial cognitive abilities, which enable them to reason about geospatial relationships based solely on linguistic cues. Equipping LLMs with sophisticated geospatial cognition can significantly enable real-life applications. Recent research indicates that LLMs are able to encode geospatial knowledge (OpenAI et al., 2024; Fu et al., 2024; Hong et al., 2023; Liu et al., 2025). However, despite these domain-specific advances, a unified hierarchical framework for evaluating geospatial cognition in LLMs remains absent, making the realization of advanced geospatial reasoning an open research challenge.

The geospatial cognitive hierarchy proposed by (Werner et al., 1997), widely accepted across geoinformatics and cognitive science for decades (Yang et al., 2025), provides an ideal structure for such evaluation. This framework delineates three hierarchical levels as Figure 1:

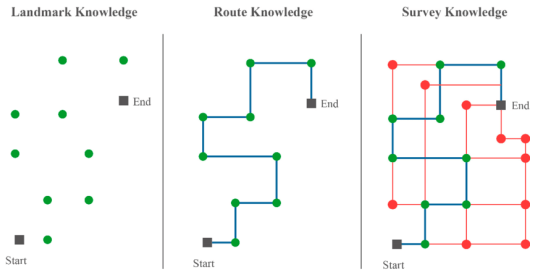


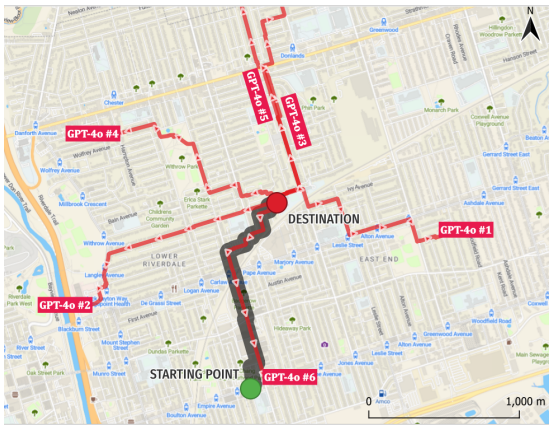
Figure 1: Geospatial cognition hierarchy from Points → Routes → Networks (Quesnot and Roche, 2014). **Landmark level**—knowledge of static landmarks (e.g. a building's address); **Route level**—understanding of connections between landmarks, including the length of a route; **Survey level**—comprehensive geospatial knowledge that enables identification of any landmark and planning routes between them (note that route planning belongs to the *survey*, not the *route*).

Current research disproportionately focuses on Landmark-level geospatial cognition, likely because such knowledge is easier to textualise, yet the same body of work indicates that LLMs face significant cognitive challenges with Route-level knowledge (Momennejad et al., 2023; Mooney et al., 2023a; Feng et al., 2023). Through our experiments and literature review, we found that current LLMs lack map-like survey knowledge. However, they can interpret simpler route information to a certain degree. Therefore, we set up this benchmark with

*This research was led by the Huawei Riemann Lab.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 8722–8741
November 4-9, 2025

(a) Route reversal of LLMs



(b) Robustness of route reversal by GPT-4o

Figure 2: **a)**: In an "Easy" route reversal task, LLMs exhibited low accuracy and consistency. None of the reverse routes return to the starting point. Gemini achieved the highest similarity score of 73.4, whereas Llama attained a similarity score of 22.6. **b)**: After six iterations of the same "Hard" route reversal task, GPT-4o exhibited significant dissimilarity. The robustness score for the reversed routes is 23.6. (© MapTiler & OpenStreetMap)

a focus on the Route Knowledge.

Route Knowledge task emphasize sequence-based navigation, landmark-dependent instructions, and procedural paths that do not require global geospatial cognition. Among them, route reversal has been recognized as a representative task for geospatial cognition evaluation within cognitive science and control systems research (Furgale and Barfoot, 2010; Karimpur et al., 2016). It plays a crucial role in evaluating human geospatial cognition (Allison and Head, 2017) and has contributed to the development of animal-inspired navigation algorithms for robots (Kumar et al., 2018). The route reversal task involves two distinct conceptual subtasks: (1) spatially determining the endpoint relative to the starting location from a given forward route, and (2) effectively navigating back to the original starting point (Mallot, 2024). Unlike landmark knowledge, which can be easily conveyed through textual description, route reversal inherently demands geospatial reasoning without requiring the comprehensive survey knowledge typically associated with route planning. Consequently, route reversal currently presents the most compelling and targeted scenario for evaluating route-level geospatial cognition for LLMs.

In this paper, we introduce TurnBack, a benchmark explicitly designed to evaluate the route-level geospatial cognition of LLMs on the route reversal task. To achieve this, we develop an algorithm leveraging OpenStreetMap and OpenRouteService (GIScience, 2024) to generate extensive, realistic route datasets complete with navigational instructions. LLMs are then tasked with generating reversed navigational instructions, to guide the construction of reversed routes using our novel tool, PathBuilder (PB). Furthermore, we propose a comprehensive evaluation framework to systematically measure the performance of SOTA LLMs on this task and analyse the disorders. The complete workflow for our proposed approach is illustrated in Figure 3. Our contributions are as follows:

- **TurnBack Benchmark**: We release the first large-scale route-reversal dataset including 36 000 pedestrian routes across 12 global cities at three difficulty levels. We also provide a comprehensive evaluation schema to reveal the performance of LLMs. It offers a reproducible probe of route-level geospatial cognition in LLMs.
- **PathBuilder**: a novel language-to-geometry converter. It bridges the gap between the formal language of route geometry and the natural language processing capabilities of LLMs.
- **Comprehensive Disorders Study**: We benchmark nine SOTA LLMs and expose four recurring geospatial cognition disorders. We point out that LLM currently suffers from architectural weaknesses in geospatial cognition.

2 Related Work

LLMs' Geospatial Cognition: LLMs perform well on **Landmark** cognition tasks like answering geographic questions (Bhandari et al., 2023),

but struggle with **Survey** cognition tasks such as route planning and navigation (Mansourian and Oucheikh, 2024; Yan and Lee, 2024; Gupta et al., 2024). While some work shows LLMs can internalize spatial representations like latitude and longitude (Gurnee and Tegmark, 2023), their cognition is still in an early transition from Landmark to Route knowledge, facing significant challenges.

Route Reversal Benchmark: Existing benchmarks for LLM geospatial cognition have three common limitations. First, **Landmark Overfocus**: Many studies rely on repetitive landmark questions that play to the LLMs' language strengths (Manvi et al., 2024; Mooney et al., 2023b). Second, **Premature Survey Inquiry**: Some works test survey knowledge without first ensuring the model has robust route knowledge (Ding et al., 2024). Third, **Vision Confusion**: The use of vision makes it hard to attribute performance to either perception or internal reasoning (Feng et al., 2024). In contrast, route reversal is a long-standing metric in geospatial research that predates LLMs (Mallot, 2024; Allison and Head, 2017; Donald Heth et al., 2002; Coutrot et al., 2022). It effectively isolates route-based cognition from both vision and landmark knowledge. See Appendix A for more details.

3 Benchmark

3.1 TurnBack Dataset

The dataset in this paper covers all continents (except Antarctica), with two representative metropolitan cities in each continent: Toronto, Denver, Mexico City, São Paulo, London, Munich, Tokyo, Singapore, Sydney, Auckland, Cairo and Cape Town. 3,000 routes were extracted from each city and equally divided into 3 difficulty levels, resulting in 36,000 routes in total, with our proposed algorithm 1. Routes selected for this study range between 500 and 2,500 meters, suitable for pedestrian navigation with OpenStreetMap and OpenRouteService (GIScience, 2024).

3.1.1 Data Generation

The dataset creation involves five steps: (1) generating starting points S_i following a Gaussian distribution within a city, with a particular latitude and longitude as its center; (2) selecting endpoints D_j randomly within a specified radius (r_{min}, r_{max}) from each S_i; (3) computing routes and extracting navigation instructions for each valid pair (S_i, D_j); (4) standardizing instructions through natural language processing; and (5) compiling formatted instructions with corresponding route geometry data. Theoretically, this dataset can be scaled infinitely as long as computational resources allow.

Difficulty	Samples	Avg. Length	Avg. Turns
Easy	12000	925	4.6
Medium	12000	1598	7.8
Hard	12000	2032	13.2

Table 1: Dataset characteristics across difficulty levels. Each level is defined by equally dividing the 0-100% range, with 5% buffer zones at transitions.

3.1.2 Data Split

Human geospatial cognition performance can be influenced by different urban road network patterns (Coutrot et al., 2022). In order to reveal the fine-grained performance with regard to the road pattern, we need to classify routes into different difficulty levels. Thus, we propose a simple method to measure the complexity of a route using two fundamental metrics: length, and number of turns, as shown in Eq. 1.

$$C = \frac{d_{\text{max}} - \frac{n_t}{l}}{d_{\text{max}} - d_{\text{min}}} \times 100 \quad (1)$$

where n_t denotes the number of turns in the route and l is the route length in meters. The parameters d_{min} and d_{max} are the minimum and maximum ratios of $\frac{n_t}{l}$ within the dataset. The complexity C is normalized to a range from 0 to 100.

We partitioned the dataset into three difficulty levels (easy, medium, and hard) by equally distributing them according to the complexity. This systematic approach ensures clear difficulty demarcation. As shown in Table 1, the dataset characteristics vary significantly across difficulty levels. In addition, interesting geographic heterogeneity does exist across cities, see Table 9.

3.1.3 PathBuilder

This section presents a novel tool that facilitates the construction of path based on natural language instructions or vice versa. Current routing engines, such as Google Maps or OpenRouteService, exhibit varied navigation instruction styles but generally adhere to common formatting principles. Frequent use of verbs like "turn", "go", "keep", and "continue" enables these instructions to be translated into sample commands for path construction. Geometrically, a route is represented by a sequence of

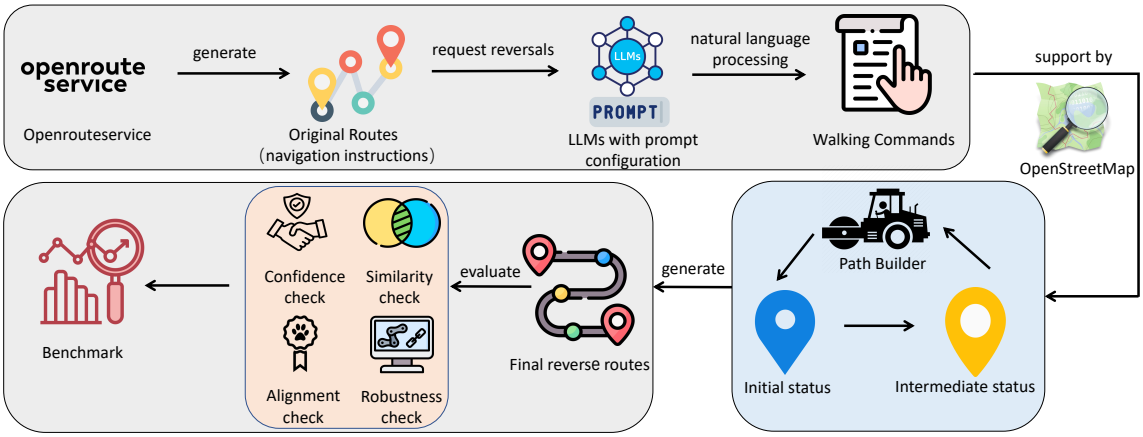


Figure 3: A three-stage workflow: (1) **Data Generation**: A route engine generates random routes and collects navigation instructions of route reversal via a prompt-based LLM; (2) **Routes Construction**: A Path-Builder constructs reversed routes with geometric support from the OpenStreetMap; (3) **Evaluation & Benchmarking**: Following a multimetric evaluation, a comparison with the geospatial route cognition capabilities of SOTA LLMs is performed.. (© Icons by Flaticon)

Algorithm 1: PathBuilder: Navigation Instruction to Geometric Path

Input: Navigation instructions I, OpenStreetMap data
Output: Geometric path P

Phase 1: Parse Instructions;
Convert natural language instructions to command sequence C;
Initialize starting position S and direction θ (North: $0°$);
$Q \leftarrow \emptyset$; `// Init coord. queue`
Phase 2: Process Commands;
foreach *command* $c_i \in C$ **do**
  if c_i *is turn command* **then**
    Update θ based on turn angle;
  else if c_i *is move command* **then**
    Update position S based on θ and distance;
    $Q \leftarrow Q \cup \{S\}$; `// Store new pos.`
Phase 3: Generate Path;
Connect points in Q using routing engine;
return P

connected points, with each navigation command guiding the selection of subsequent points based on specific rules. Thus, replicating a route involves reversing the original translation from geometry to natural language. As outlined in Algorithm 1, the path-building process comprises three main phases.

We evaluated the PathBuilder's performance using similarity scores across diverse urban environments, specifically in Tokyo, Munich, and Toronto, see Appendix D. The results show that it is a powerful tool for generating geometry from navigation instruction.

3.2 Evaluation Metrics

The metrics employed comprise two aspects: (1) geometric performance, and (2) LLMs' generation performance. They evaluate the route level knowledge of LLMs in terms of both the quality of the returned geometry as well as their thinking process.

3.2.1 Geometric Performance

There is an important assumption: since our original route was generated using a route-planning engine, the optimal reverse route should be backtracked. As Figure 5(b) shows, any reverse route that differs from the original one and does not return to the start point is considered a failure to some extent. Thus, we have two ground truth references: the starting point and the optimal reverse route.

Return Rate: percentage of reverse routes that return to the start point (tolerances of up to 20 meters are allowed).

Similarity Score: We define the similarity measure $\text{sim}(x_i, x_j)$ as a weighted sum of multiple geographical and mathematical metrics, where x_i and x_j denote two routes being compared. The metrics include Length Ratio (LR), Hausdorff Distance (HD), Fréchet Distance (FD), Edit Distance (ED), Jaccard Index (JI), Angle (A), and Sum of Coordinates Offsets (SCO). The similarity score is given by:

$$\text{Sim}(x_i, x_j) = w_{\text{LR}} \cdot \text{LR} + \cdots + w_{\text{SCO}} \cdot \text{SCO} \quad (2)$$

where w_k represent their respective weight.

In our evaluation, a similarity score above 80 indicates strong resemblance, scores over 90 suggest

near equivalence, while values below 50 denote dissimilarity, and scores under 30 indicate completely distinct routes. Further details on the metrics can be found in Table 8.

3.2.2 Generation Performance

Robustness: We evaluate the consistency of responses generated by LLMs for the same routing task using a robustness score. We do so by computing the standard deviation σ of pairwise similarities among responses:

$$\sigma = \sqrt{\frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (\text{sim}(x_i, x_j) - S)^2}, \quad (3)$$

where S is the average similarity across all unique response pairs. Finally, σ is normalized using min-max normalization over the full dataset to yield a robustness score $R \in [0, 100]$, where higher values indicate greater consistency and robustness.

Confidence: Inspired by Xu et al. (2024), we adapt their method to compute the confidence level of LLMs in the generated instructions.

$$Confidence_{set} = \frac{1}{N} \sum_{i=1}^{N} DirectionProb_i \quad (4)$$

where N is the total number of direction instructions, and $DirectionProb_i$ is the token probability of the direction word in the i-th instruction. This probability can be calculated from the log probabilities record of LLMs. For example, if the token "north" has an 80% likelihood among candidate tokens, we interpret the LLM as being 80% confident in "turning north." The detailed confidence calculation is provided in Appendix F.

Misalignment: As an important metric of LLM research, we decide to use misalignment to measure the performance of LLMs to output valid instructions. In our task, this metric is not about correctness but represents the LLMs' capacity for geospatial reasoning. Even after prompt engineering, they may still produce invalid navigation instructions or generate broken routes.

3.3 Prompt Design

Given that prompt engineering significantly affects the model's focus and returned content format (He et al., 2024), we use two types of prompts: the guide prompt, and the instruction prompt.

The guide prompt, shown in Figure 8, is meant to let the LLMs focus on the geographic information of the current experiment location. It attempts to ensure that the LLM does not use text-based semantic inversion methods because we found that LLMs tend to invert direction words to answer such as "turn left and walk 500 meters southeast" to "turn right and walk 500 meters northwest".

The instruction prompt, shown in Figure 9,is intended to align the return format of the LLM with the navigation instructions that the PB can execute. For each navigation instruction, the coordinates and nearby landmarks (if any) are given. This is to encourage LLMs to think with the geospatial information they have been trained with.

3.4 Landmark Preliminary Experiment

As an enhancement to the benchmark, we constructed 200 manually curated question sets for preliminary evaluation, illustrating that current LLMs exhibit geospatial cognition situated between advanced landmark knowledge and route knowledge. LLMs perform well on knowledge with abundant textual support but struggle with lesser-known landmarks. Notably, performance degrades sharply when tasked with directional reasoning or coordinate system calculation. Due to space limitations, details are provided in Appendix I.

4 Route Reversal Benchmark

We use **Reverse Route** denotes the answers returned by LLMs, **Original Route** to represent the sample route sent to LLMs for route reversal. In this paper we use Original Route as the ground truth for route reversal evaluation. Because all routes are optimized by the route planning engine in the generation process, the optimal reverse route is the original route backwards.

4.1 Target LLMs

We tested eleven SOTA LLMs:GPT-4o, GPT-o1, GPT-O3, Gemini1.5-pro, Gemini2.5-Pro, Llama3.3-70B, Deepseek(R1), Claude 3.5,Claude 3.7 and Grok. All open-source models were set up according to the HuggingFace tutorial and temperature was set 0. Temperature is proved not a determinant of performance, see Appendix H,G.

4.2 Results and Findings

Results: The main results of SOTA LLMs on the proposed benchmark are summarized in Table 2 and 3. As depicted in Figure 7, the similarity scores

Table 2: Benchmark of LLMs on route reversal task (geometric performance).

Model	Return Rate (%) ↑	Similarity ↑	Deviation Angle ↓ (°)	Hausdorff Distance ↓ (m)	Length Ratio (↔1)	Jaccard Index ↑
GPT-4o	6.34	41.06 (0.26)	32.18 (0.23)	169.47 (1.07)	0.88 (2.4e-3)	0.36 (1.6e-3)
GPT-o1	9.47	48.13 (0.24)	29.47 (0.22)	142.65 (0.89)	0.71 (2.9e-3)	0.43 (2.4e-3)
GPT-o3	12.25	53.99 (0.24)	23.22 (0.36)	123 (0.62)	0.71 (2.9e-3)	0.43 (2.4e-3)
Gemini1.5-Pro	11.93	61.71 (0.19)	36.63 (0.25)	136.23 (0.85)	1.12 (3.2e-3)	0.50 (3.1e-2)
Gemini2.5-Pro	**14.46**	**67.26** (0.15)	26.73 (0.32)	113.23 (0.55)	1.01 (2.2e-1)	**0.53** (2.7e-3)
Llama3.3-70B	4.06	42.78 (0.27)	**53.67** (0.35)	189.32 (1.19)	0.79 (3.4e-3)	0.37 (2.8e-3)
Deepseek	7.63	40.01 (0.16)	31.23 (0.23)	152.80 (0.94)	0.93 (4.2e-3)	0.34 (2.6e-3)
Deepseek R1	9.42	48.15 (0.19)	30.35 (0.20)	131.21 (0.78)	1.09 (3.1e-3)	0.41 (2.3e-3)
Claude 3.5	7.33	40.62 (0.17)	36.83 (0.25)	158.39 (1.03)	0.93 (2.2e-3)	0.35 (1.5e-3)
Claude 3.7	9.05	49.50 (0.20)	33.15 (0.20)	**128.56** (0.81)	**1.17** (5.5e-3)	0.44 (3.8e-3)
Grok	6.72	40.79 (0.29)	33.42 (0.23)	183.85 (1.16)	0.77 (3.1e-3)	0.35 (2.7e-3)
GPT-4o	2.93	36.19 (0.19)	84.17 (0.57)	340.28 (2.17)	1.06 (2.2e-3)	0.31 (1.2e-3)
GPT-o1	3.68	43.87 (0.19)	67.12 (0.47)	289.17 (1.81)	1.05 (3.6e-3)	0.39 (2.2e-3)
GPT-o3	5.16	47.16 (0.19)	63.85 (0.46)	283.90 (1.76)	1.05 (3.0e-3)	0.41 (2.2e-3)
Gemini1.5-Pro	5.12	47.34 (0.14)	68.20 (0.49)	287.90 (1.84)	1.08 (3.1e-3)	0.40 (2.1e-3)
Gemini2.5-Pro	**7.63**	**51.33** (0.14)	67.42 (0.47)	279.76 (1.72)	**1.22** (3.1e-3)	**0.42** (2.1e-3)
Llama3.3-70B	1.82	39.29 (0.13)	**104.25** (0.72)	**392.23** (2.47)	1.00 (3.6e-3)	0.34 (2.0e-3)
Deepseek	1.79	38.23 (0.12)	66.95 (0.48)	311.49 (1.93)	1.13 (6.2e-3)	0.32 (3.4e-3)
Deepseek R1	4.09	44.32 (0.14)	67.73 (0.52)	278.14 (1.73)	1.10 (5.1e-3)	0.38 (3.5e-3)
Claude 3.5	3.18	39.16 (0.13)	78.92 (0.53)	372.55 (2.33)	1.19 (7.2e-3)	0.33 (3.9e-3)
Claude 3.7	3.92	44.16 (0.14)	56.73 (0.42)	274.62 (1.70)	1.07 (4.4e-3)	0.38 (2.8e-3)
Grok	2.96	37.16 (0.23)	99.87 (0.67)	365.41 (2.23)	1.07 (4.1e-3)	0.32 (2.4e-3)
GPT-4o	0.21	24.13 (0.15)	132.15 (0.92)	789.23 (5.02)	1.19 (3.3e-3)	0.23 (1.2e-3)
GPT-o1	0.26	29.91 (0.16)	98.20 (0.75)	705.40 (4.60)	1.16 (5.0e-3)	0.26 (2.0e-3)
GPT-o3	0.57	33.81 (0.16)	91.87 (0.70)	682.57 (4.29)	**1.31** (5.4e-3)	0.28 (2.1e-3)
Gemini1.5-Pro	0.37	29.56 (0.10)	90.18 (0.67)	670.38 (4.20)	1.02 (4.0e-3)	0.27 (2.3e-3)
Gemini2.5-Pro	0.57	34.01 (0.10)	96.15(0.61)	550.38 (3.25)	0.91 (4.1e-3)	0.26 (2.4e-1)
Llama3.3-70B	0.49	24.87 (0.10)	**156.93** (1.07)	**810.45** (5.25)	1.25 (7.5e-3)	0.22 (2.5e-3)
Deepseek	**0.87**	26.86 (0.10)	102.82 (0.78)	733.61 (4.57)	0.98 (3.2e-3)	0.25 (1.6e-3)
Deepseek R1	0.18	**30.10** (0.10)	120.93 (0.87)	623.15 (3.86)	1.05 (4.1e-3)	**0.28** (2.2e-3)
Claude 3.5	0.83	26.10 (0.10)	120.78 (0.85)	747.82 (4.78)	1.18 (6.4e-3)	0.24 (2.4e-3)
Claude 3.7	0.61	28.18 (0.09)	86.47 (0.73)	672.41 (4.18)	1.09 (5.2e-3)	0.26 (2.4e-3)
Grok	0.14	23.98 (0.21)	144.28 (1.08)	747.56 (4.78)	0.85 (3.1e-3)	0.21 (1.5e-3)

Note: The values in parentheses are standard errors. The highest and lowest are shown in bold and underlined, respectively. Green, yellow, and red indicate easy, medium, and difficult, respectively (same as Table 3). The Hausdorff distance can be seen as the distance between two geometries. See Appendix C for more details.

are derived from the normalized cumulative distributions of similarity. Each route was tested 6 times.

Our analysis indicates that route reversals generated by LLMs generally exhibit low return rates, low similarity, and low robustness despite high confidence scores.

Findings I: Current SOTA LLMs struggle considerably with the route reversal task, as exemplified in Figure 2(a). Even in the easy dataset, involving routes typically shorter than 1 km with 3–5 turns, LLMs only achieve reture rates between 4.0% and 11.9%. For routes of medium complexity, most models fail entirely. This observed difficulty with basic route reversal tasks underscores a critical limitation in current LLMs' geospatial route cognition. Consequently, among routes that fail to return, the similarity between them ground truth is also quite low.

Findings II: LLMs suffer from **inconsistency** in geospatial route cognition. Routes generated by identical prompts of the same LLM exhibit notable inconsistencies, resulting in low robustness across all tested models. Gemini-2.5-pro achieved the highest robustness score at 69.15, while Llama-3.3-70B had the lowest at only 38.58. These results clearly indicate that current LLMs display significant stochasticity in their geospatial route cognition. Figure 2(b) shows an example of the inconsistency generation of GPT-4o. Theoretically, if LLMs have a certain level of knowledge about the question, there should be coherence among multiple responses, even if they are incorrect. However, we observe that LLMs' multiple responses are not convergent, which means that the route generated by their multiple responses have very little similarity to each other.

Table 3: Benchmark of LLMs on route reversal task (disorder performance).

Easy				Medium				Hard			
Model	Robustness	Confidence	Misalignment	Model	Robustness	Confidence	Misalignment	Model	Robustness	Confidence	Misalignment
GPT-4o	43.74	92.50	**14.57**	GPT-4o	36.24	91.76	13.04	GPT-4o	19.74	89.47	16.12
GPT-o1	51.32	96.26	8.03	GPT-o1	44.9	92.57	8.08	GPT-o1	38.85	94.25	7.15
GPT-o3	67.49	**96.51**	5.16	GPT-o3	52.16	**93.61**	7.83	GPT-o3	**42.16**	**89.72**	6.75
Gemini1.5-Prc	45.79	93.97	7.52	Gemini1.5-Prc	37.52	87.46	6.19	Gemini1.5-Prc	37.52	87.46	6.19
Gemini2.5-Prc	**69.15**	91.23	3.51	Gemini2.5-Prc	**56.15**	**90.15**	6.94	Gemini2.5-Prc	40.09	88.51	7.02
Llama3.3-70B	38.58	86.52	12.53	Llama3.3-70B	27.40	85.63	11.06	Llama3.3-70B	13.60	92.15	**16.21**
Deepseek	47.32	90.04	12.58	Deepseek	38.12	91.23	**13.12**	Deepseek	21.54	90.25	14.57
Deepseek R1	52.41	93.18	9.53	Deepseek R1	36.58	92.63	8.08	Deepseek R1	20.04	93.16	9.49
Claude 3.5	53.12	N/A	9.14	Claude 3.5	43.59	N/A	9.56	Claude 3.5	32.13	N/A	9.03
Claude 3.7	56.13	N/A	4.58	Claude 3.7	46.39	N/A	5.06	Claude 3.7	39.57	N/A	5.12
Grok	44.16	N/A	10.07	Grok	40.29	N/A	11.59	Grok	37.51	N/A	10.54

Findings III: Obvious **misalignment** demonstrated by LLMs is their notable deficiency in identifying essential information for constructing valid routes, frequently leading to the omission of critical initial absolute directional guidance. An absolute direction (e.g., east, northwest, or 100°) at the starting point is required to establish a viable route. Our experiments reveal that this inability to provide an actionable first step accounts for 4–16% of route generation failures. Furthermore, some LLM responses degenerate into complete failures, offering overly generic or non-navigational advice, exemplified by suggestions like, "From your starting location, simply walk back to your destination." Notably, unlike common alignment challenges, where iterative prompting can rectify errors, repeated emphasis on the necessity of an initial absolute direction yields only marginal improvements in this context. This persistent deficiency suggests a more fundamental limitation: LLMs may lack an internalized spatial framework crucial for geolocation-based reasoning, particularly in grasping the importance of "self-localization" and "sequential spatial linkage" for effective route construction.

Findings IV: Despite general inadequacies, more advanced LLMs demonstrate clear advantages in geospatial tasks. Gemini2.5-pro, previously recognized for its superior spatial cognitive capabilities (Yang et al., 2024b), notably outperforms other models across some metrics. The contrast between GPT-o3 and GPT-4o further illustrates this point: GPT-o3, considered more advanced in reasoning, surpasses GPT-4o by 5.9% in return rate and 13% in similarity on the easy dataset, alongside better robustness. However, it is interesting to note that chain-of-thought could make the model performance drop on difficult datasets, for example, Deepseek vs. R1. These observations confirm both the ongoing progress of LLMs in geospatial cognition and the discriminative effectiveness of our benchmark.

5 Geospatial Cognition Disorder Study

The failure pattern led us to identify two disorders unique to geospatial cognition: disorientation and superficiality. The distribution of the above disorders is shown in Table 4. Due to space limits, we select four typical models to analysis.

Table 4: Geospatial Cognition Disorders for Route Reversal (a route can suffer multiple disorders).

Disorder	Proportion
Inconsistency	47%
Superficiality	21%
Misalignment	12%
Disorientation	63%

Disorientation not only means that LLMs get lost in road networks, but also that they cannot identify current location from the starting point. According to the low return rate, most routes generated end up far from the ground truth. This indicates that LLMs have disorders in constructing the road network in their latent space. Moreover, we found that even within a single instruction, LLMs might behave in a confusing manner when questions are asked about the current location. In an example from GPT-4o, after “Turn west and continue for 100 m,” followed by “Turn right and go straight for 300 m,” GPT-4o claimed that the endpoint lay *southwest* of the origin, an impossible result. Repeating this probe on 300 easy routes and checking the answers manually, we found that GPT-4o identified the final direction correctly in only about 50 % of the cases (Table 5), even when the task only require an approximate direction.

Table 5: Disorientation of GPT4o for 300 Samples

Difficulty Level	Disorientation (%)
Easy (100)	32%
Medium (100)	56%
Hard (100)	85%

Superficiality is a tendency to employ task-

agnostic heuristics or exploit statistical patterns in the training data, rather than engaging in the problem-specific reasoning required by the prompt. In the context of route reversal tasks, genuine problem solving necessitates geospatial reasoning, i.e. understanding relative positions. However, LLMs often exhibit superficiality by resorting to semantic inversion—a simple reversal of directional terms ("north" to "south", "left" to "right") and step order—without necessarily processing the underlying spatial relationships. Crucially, this semantic inversion heuristic does not reliably guarantee accuracy and represents a shortcut around the intended computational process.

Our experiments reveal a notable correlation between this superficial behavior and the models' confidence: outputs generated via semantic inversion consistently exhibit higher token probabilities compared to outputs from geospatial reasoning. As illustrated in Table 6, we suggest that token probability analysis could serve as an indicator for detecting whether an LLM is employing domain-specific reasoning or relying on superficial heuristics.

Table 6: Average Confidence Differences of LLMs in Semantic Reversal (superficial approach) and Normal Geospatial Reasoning in 200 samples

Reasoning Mode	GPT4o	Llama3.3-70B	Gemini2.5 Pro	GPTo3
Semantic Reversal	98	99	98	99
Geospatial Thinking	92	91	93	90

5.1 Discussion

Our analysis suggests that poor LLM performance on route-reversal tasks is attributable to a fundamental architectural deficit in geospatial cogntion, rather than to easily remedied factors like prompt engineering. This claim is supported by our empirical results: low return rates and similarity, high variance across trials, and systematically inflated confidence for invalid solutions. Moreover, the marginal improvement from prompt refinements suggests the bottlenecks are systemic and lie deeper than surface-level instruction following.

Representation bottleneck. Sub-word tokenizers divide coordinates such as52.5167 into "52", ".", "516", "7", mapping them to nearly orthogonal vectors. This process breaks the relationship between a number and its representation, violating the principle of *metric continuity*. Consequently, a small numerical change becomes a large, unrelated jump in the embedding space. This lack of numerical sensitivity creates cascading errors in operations that depend on metric continuity, such as heading updates or cumulative offsets, amplifying small inaccuracies into large route deviations. Studies confirm that LLMs' grasp of ordered magnitude is weak and heavily dependent on the tokenizer (Yang et al., 2024a; Seßler et al., 2024). This explains why models generate semantically plausible but metrically divergent paths.

Objective misalignment. The next-token cross-entropy objective is agnostic to numerical proximity. For example, predicting 52.5168° incurs the same loss as 90.00° when the true value is 52.5167°. Because the loss function is defined over a discrete vocabulary, the optimization gradient cannot signal the *magnitude* of numerical error. The model is thus incentivized to match tokens for syntactic fluency, not to minimize the metric distance between its predicted coordinates and the ground truth. This explains the observed gap between high token probabilities (confidence) and low geometric fidelity. This finding aligns with prior work, indicating that LLMs degrade to random performance in tasks requiring sub-degree accuracy (Kazemi et al., 2023).

Data sparsity. High-precision coordinates and long-tail place names are exceedingly rare in web-scale text, causing their token frequencies to approach zero. Unlike specialized models that discretize space into learnable grid cells (e.g., H3) at the cost of precision (Schestakov and Gottschalk, 2024), LLMs lack a principled spatial hashing mechanism. Consequently, they default to coarse linguistic heuristics, such as inverting "left/right" and reordering steps. This strategy achieves textual plausibility but fails to reconstruct the fine-grained geometric path. This outcome is consistent with the performance drop on harder routes where small metric errors accumulate.

Failure of Geometric Compositionality. The self-attention mechanism, which computes a content-weighted average, is fundamentally a semantic aggregator, not a geometric transformer. It lacks the inductive bias for affine transformations like rotation, translation, and scaling that are foundational to path integration (Dziri et al., 2023). This architectural deficit directly causes the empirical failures observed in our benchmark, such as *disorientation* (failure to maintain a global heading) and initial *misalignment*. The model resorts to shallow symbolic flips, a phenomenon termed "linearized subgraph matching" (Dziri et al., 2023), such as swapping "left/right" without re-grounding the tra-

Table 7: Effect of adding a vector map to the prompt (GPT-4o, 200 easy routes, Toronto).

Condition	Return Rate (%)	Similarity
Original	6.4	41.06
Assisted	43.7	73.08

jectory. This indicates a reliance on surface patterns over a latent spatial scaffold.

Taken together, these factors offer a coherent explanation for the observed failures. Crucially, they clarify why stronger prompting yields only limited gains. Prompts can steer output format, but they cannot supply the numeric continuity, metric-aware objectives, dense spatial data, or geodesic operators that are missing from the model's architecture. Addressing these gaps is critical for enabling reliable route-level geospatial cognition.

6 Possible Mitigation

Full architectural changes (numeric-aware tokenizers, metric-aware losses, spatial operators) are costly. We therefore test a low-cost, inference-time aid without modifying our language-first benchmark. We attach a *vector map* to the original prompt. The map is rendered from OSM-style vector tiles, shows streets and salient POIs only. Reversal rules and output format remain unchanged. We report two settings: **Original (language-only)** for route cognition without visual confounds, and **Assisted (language + vector map)** as a minimal cue for inference. On 200 *easy* routes in Toronto (GPT-4o), the return rate rises from 6.4% to 43.7%, and similarity from 41.06 to 73.08 (Fig. 4, Table 7).

The gain shows that lightweight visual context reduces gross reversal failures. It does not establish the mechanism: improvements may come from spatial abstraction or from visual anchoring that bypasses textual reasoning. We use vector maps (topology and labels, but no imagery) to limit spurious cues while a full disentanglement is left for future analysis. Making maps the *primary* input would mix visual perception with route cognition and weaken attribution. We insist on a language-first solution because it isolates route reasoning and aligns with the landmark–route–survey hierarchy. There is evidence that route knowledge can be acquired without vision (Tinti et al., 2006) in cognition science research. So while we admit that the assisted setting can improves utility from engineering perspective, it does not address the architectural issues in Sec. 5.1.

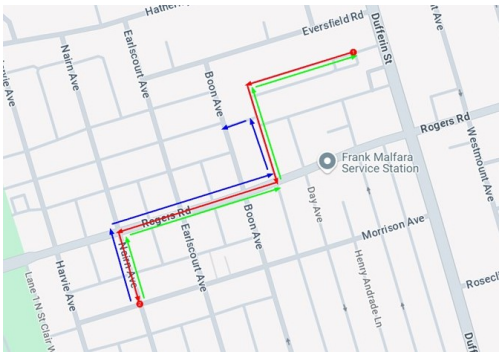


Figure 4: Original route (red), reverse path without image prompt (blue), and with the vector-map prompt (green) on an easy sample in Toronto.

7 Conclusion

We introduced a benchmark *TurnBack* for evaluation on geospatial route cognition. It reveals that SOTA LLMs are still far from reliable geospatial reasoners. On the *TurnBack*, no model solved the route reversal task properly. Even on "easy" dataset the return rate was below 12%, and performance collapsed to near zero on harder routes. Reverse routes often failed to reach the start point, and repeated queries produced divergent paths, revealing the absence of a stable internal geospatial representation. Moreover, models expressed unwarranted confidence, underscoring a deep misalignment between natural language generation and true geospatial cognition.

Models that pass the benchmark will prove competent in route-cognition tasks such as road-network knowledge graph and show potential in route planning. By turning these weaknesses into quantifiable analyses, we highlight three research avenues: (i) representations that preserve distance—numeric-aware tokenizers or geospatial cell vocabularies; (ii) objectives that punish metric error rather than pure string mismatch; and (iii) inductive biases that embed geodesic operators or call external map tools. Such advances can potentially enables LLMs achieve reliable geospatial cognition.

8 Limitations

Although our approach offers novel insights into evaluating LLMs' geospatial route cognition, several limitations warrant acknowledgment.

Data Constraints: Our dataset, comprising 36,000 routes across 12 cities on 6 continents, of-

fers substantial coverage. However, it does not encompass the full diversity of global urban road network typologies. Furthermore, other network types, such as those in rural areas or national parks, which could elicit different geospatial reasoning from LLMs, were not included. This exclusion was due to the significant challenges in data acquisition and quality assurance required for a comprehensive Earth-scale exploration within the constraints of this study. It is pertinent to note that our data generation methodology, leveraging OpenStreetMap (OSM), theoretically permits scaling to OSM's full data. However, this is beyond the practical scope of this research.

Model Constraints: Our empirical evaluation was necessarily confined to a selected set of LLMs due to computational resource limitations. Consequently, the findings may not directly generalize to all existing or future LLMs. The field of LLMs is characterized by rapid advancements, with new models and architectures emerging frequently. While we endeavored to include SOTA and representative models available during our experimental phase (early 2025), it is inevitable that some newer models may not be covered by the time of publication. Additionally, financial constraints precluded the large-scale evaluation of certain proprietary models, such as GPTo4. Nevertheless, we emphasize that our proposed methodology and its implementation are reproducible with the APIs of a vast majority of LLMs, ensuring the broader applicability of our evaluation framework.

PathBuilder Constraints Due to the topology diversity of the routes and the vague expression of natural language outputs from LLMs, it is challenging to implement a PathBuilder with 100% reproduction accuracy. Please refer to Appendix E for detailed explanations and examples.

Theoretical constraints Although route reversal is the most representative path knowledge task in our theory, however, we cannot deny that there still exist some other representative tasks such as the interrogation of geometric properties of routes. These tasks are either too trivial or not convincing enough. Ideally all representative tasks from Landmark-Route should be integrated, but such a workload is beyond the scope of this paper.

8.1 Ethical Considerations

To the best of our knowledge, we do not have any potential ethical concerns to disclose.

References

Samantha Allison and Denise Head. 2017. Route repetition and route reversal: Effects of age and encoding method. *Psychology and Aging*, 32(3):220.

Hannah Bast, Daniel Delling, Andrew Goldberg, Matthias Müller-Hannemann, Thomas Pajor, Peter Sanders, Dorothea Wagner, and Renato F Werneck. 2016. Route planning in transportation networks. *Algorithm engineering: Selected results and surveys*, pages 19–80.

Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. 2023. Are large language models geospatially knowledgeable? In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–4.

Weicheng Cai, Jinkun Chen, and Ming Li. 2018. Exploring the encoding layer and loss function in end-to-end speaker and language recognition system. *arXiv preprint arXiv:1804.05160*.

Antoine Coutrot, Ed Manley, Sarah Goodroe, Christoffer Gahnstrom, Gabriele Filomena, Demet Yesiltepe, Ruth Conroy Dalton, Jan M Wiener, Christoph Hölscher, Michael Hornberger, and 1 others. 2022. Entropy of city street networks linked to future spatial navigation ability. *Nature*, 604(7904):104–110.

Peng Ding, Jiading Fang, Peng Li, Kangrui Wang, Xiaochen Zhou, Mo Yu, Jing Li, Matthew R. Walter, and Hongyuan Mei. 2024. Mango: A benchmark for evaluating mapping and navigation abilities of large language models. *Preprint*, arXiv:2403.19913.

C Donald Heth, Edward H Cornell, and Tonya L Flood. 2002. Self-ratings of sense of direction and route reversal performance. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 16(3):309–324.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. *Preprint*, arXiv:2305.18654.

Jie Feng, Jun Zhang, Tianhui Liu, Xin Zhang, Tianjian Ouyang, Junbo Yan, Yuwei Du, Siqi Guo, and Yong Li. 2024. Citybench: Evaluating the capabilities of large language models for urban tasks.

Yu Feng, Linfang Ding, and Guohui Xiao. 2023. GeoQAMap - Geographic Question Answering with Maps Leveraging LLM and Open Knowledge Base. In *12th International Conference on Geographic Information Science (GIScience 2023)*, volume 277 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 28:1–28:7, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Patricia Frontiera, Ray Larson, and John Radke. 2008. A comparison of geometric approaches to assessing spatial similarity for gir. *International Journal of Geographical Information Science*, 22(3):337–360.

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. 2024. Scene-llm: Extending language model for 3d visual understanding and reasoning. *Preprint*, arXiv:2403.11401.

Zhongliang Fu, Liang Fan, Zhiqiang Yu, and Kaichun Zhou. 2018. A moment-based shape similarity measurement for areal entities in geographical vector data. *ISPRS International Journal of Geo-Information*, 7(6):208.

Paul Furgale and Timothy D Barfoot. 2010. Visual teach and repeat for long-range rover autonomy. *Journal of field robotics*, 27(5):534–560.

Hari Krishna Gadi, Lu Liu, and Liqiu Meng. 2024. Road networks matching supercharged with embeddings. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, GeoAI '24, page 27–37, New York, NY, USA. Association for Computing Machinery.

GIScience. 2024. openrouteservice. `https://github.com/GIScience/openrouteservice`. Version v6.2.

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. *arXiv preprint arXiv:2407.11014*.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on llm performance? *Preprint*, arXiv:2411.10541.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494.

Zongcai Huang, Peng Peng, Feng Lu, and He Zhang. 2025. An llm-based method for quality information extraction from web text for crowed-sensing spatiotemporal data. *Transactions in GIS*, 29(1).

Harun Karimpur, Florian Röser, and Kai Hamburger. 2016. Finding the return path: landmark position effects and the influence of perspective. *Frontiers in psychology*, 7:1956.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. 2023. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*.

Navid Khademi and Ramin Saedi. 2019. Latent learning and the formation of a spatiotemporal cognitive map of a road network. *Travel Behaviour and Society*, 14:66–80.

Hyebin Kim and Sugie Lee. 2024. Poi gpt: Extracting poi information from social media text data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48:113–118.

David King, Aya Aboudina, and Amer Shalaby. 2020. Evaluating transit network resilience through graph theory and demand-elastic measures: Case study of the toronto transit system. *Journal of Transportation Safety & Security*, 12(7):924–944.

Ashish Kumar, Saurabh Gupta, David Fouhey, Sergey Levine, and Jitendra Malik. 2018. Visual memory for robust path following. *Preprint*, arXiv:1812.00940.

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. 2025. St-llm: Large language models are effective temporal learners. In *European Conference on Computer Vision*, pages 1–18. Springer.

Yifan Liu, Chenchen Kuai, Haoxuan Ma, Xishun Liao, Brian Yueshuai He, and Jiaqi Ma. 2024. Semantic trajectory data mining with llm-informed poi classification. *arXiv preprint arXiv:2405.11715*.

Hanspeter A. Mallot. 2024. *From Geometry to Behavior: An Introduction to Spatial Cognition*. The MIT Press.

Ali Mansourian and Rachid Oucheikh. 2024. Chatgeoai: Enabling geospatial analysis for public through natural language, with large language models. *ISPRS International Journal of Geo-Information*, 13(10):348.

Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Geollm: Extracting geospatial knowledge from large language models. *Preprint*, arXiv:2310.06213.

Ida Momennejad, Hosein Hasanbeig, Felipe Vieira, Hiteshi Sharma, Robert Osazuwa Ness, Nebojsa Jojic, Hamid Palangi, and Jonathan Larson. 2023. Evaluating cognitive maps and planning in large language models with cogeval. *Preprint*, arXiv:2309.15129.

Peter Mooney, Wencong Cui, Boyuan Guan, and Levente Juhász. 2023a. Towards understanding the geospatial skills of chatgpt: Taking a geographic information systems (gis) exam. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, GeoAI '23, page 85–94, New York, NY, USA. Association for Computing Machinery.

Peter Mooney, Wencong Cui, Boyuan Guan, and Levente Juhász. 2023b. Towards understanding the geospatial skills of chatgpt: Taking a geographic information systems (gis) exam. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 85–94.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Yanwei Pang, Yazhao Li, Jianbing Shen, and Ling Shao. 2019. Towards bridging semantic gap to improve semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4230–4239.

Teriitutea Quesnot and Stéphane Roche. 2014. Measure of landmark semantic salience through geosocial data streams. *ISPRS International Journal of Geo-Information*, 4(1):1–31.

Stefan Schestakov and Simon Gottschalk. 2024. Trajectory representation learning on road networks and grids with spatio-temporal dynamics. *arXiv preprint arXiv:2411.14014*.

Raphael Schumann and Stefan Riezler. 2021. Generating landmark navigation instructions from maps as a graph-to-text problem. *Preprint*, arXiv:2012.15329.

Kathrin Seßler, Yao Rong, Emek Gözlüklü, and Enkelejda Kasneci. 2024. Benchmarking large language models for math reasoning tasks. *arXiv preprint arXiv:2408.10839*.

Claude E Shannon. 1993. *Claude elwood shannon: Collected papers*. IEEE press.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *Preprint*, arXiv:2309.15025.

Carla Tinti, Mauro Adenzato, Marco Tamietto, and Cesare Cornoldi. 2006. Visual experience is not necessary for efficient survey spatial cognition: evidence from blindness. *Quarterly journal of experimental psychology*, 59(7):1306–1328.

V Ya Tsvetkov. 2013. Information interaction as a mechanism of semantic gap elimination. *European researcher*, (4-1):782–786.

Hiroyuki Usui. 2018. Estimation of geometric route distance from its topological distance: application to narrow road networks in tokyo. *Journal of Geographical Systems*, 20(4):387–412.

Steffen Werner, Bernd Krieg-Brückner, Hanspeter A Mallot, Karin Schweizer, and Christian Freksa. 1997. Spatial cognition: The role of landmark, route, and survey knowledge in human and robot navigation. In *Informatik'97 Informatik als Innovationsmotor: 27. Jahrestagung der Gesellschaft für Informatik Aachen, 24.–26. September 1997*, pages 41–50. Springer.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, and 1 others. 2024. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*.

Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. 2023. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*.

Rongwu Xu, Brian S. Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. *Preprint*, arXiv:2312.09085.

Yibo Yan and Joey Lee. 2024. Georeasoner: Reasoning on geospatially grounded context for natural language understanding. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4163–4167.

Anran Yang, Cheng Fu, Qingren Jia, Weihua Dong, Mengyu Ma, Hao Chen, Fei Yang, and Hui Wu. 2025. Evaluating and enhancing spatial cognition abilities of large language models. *International Journal of Geographical Information Science*, pages 1–36.

Fei Yang, Zhonghui Wang, Haowen Yan, and Xiaomin Lu. 2022. Geometric similarity measurement method for micro scene generalization. *Applied Sciences*, 12(2):628.

Haotong Yang, Yi Hu, Shijia Kang, Zhouchen Lin, and Muhan Zhang. 2024a. Number cookbook: Number understanding of language models and how to improve it. *arXiv preprint arXiv:2411.03766*.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2024b. Thinking in space: How multimodal large language models see, remember, and recall spaces. *Preprint*, arXiv:2412.14171.

Zhizhuo Yin, Yuyang Wang, Theodoros Papatheodorou, and Pan Hui. 2024. Text2vrscene: Exploring the framework of automated text-driven generation system for vr experience. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 701–711. IEEE.

A Related Work (Full)

LLMs' Geospatial Cognition:

LLMs' Geospatial Cognition: Recent research has demonstrated that LLMs perform relatively well in **Landmark** cognition tasks, such as answering geographic questions (Bhandari et al., 2023). Conversely, LLMs generally struggle with **Survey** cognition tasks, including route planning and urban navigation (Mansourian and Oucheikh, 2024; Yan and Lee, 2024; Gupta et al., 2024). Although Gurnee and Tegmark (2023) indicate that LLMs can internalize real-world spatial representations (e.g., latitude, longitude, and timestamps), current progress clearly shows that LLM geospatial cognition is transitioning from Landmark knowledge toward Route knowledge, yet significant challenges remain.

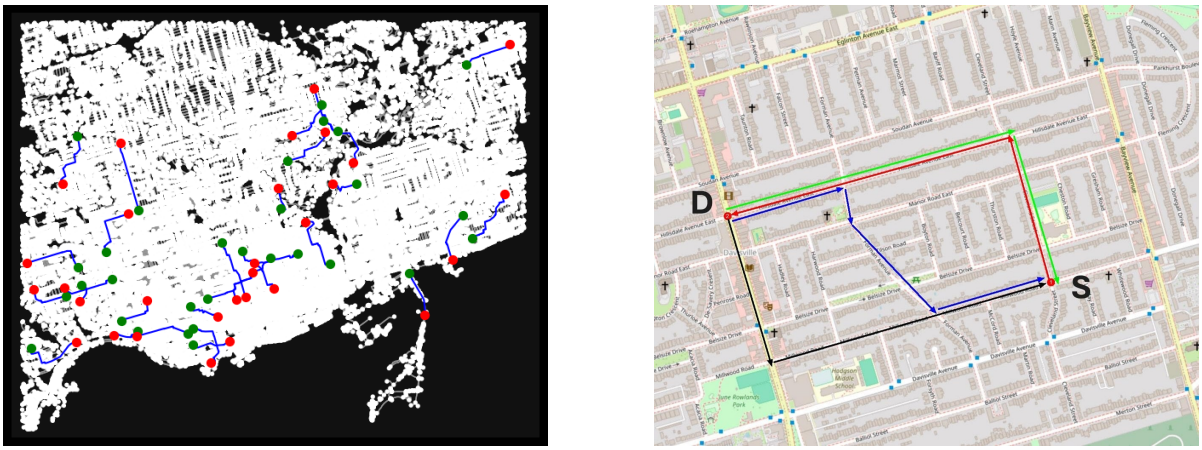
Route Reversal Benchmark: A few benchmarks have explored LLM geospatial cognition, yet they typically exhibit three limitations. First, **Landmark Overfocus**: due to the natural language-friendly nature of LLMs, many studies disproportionately emphasize repetitive landmark knowledge questions (Manvi et al., 2024; Mooney et al., 2023b). Second, **Premature Survey Inquiry**: studies such as (Ding et al., 2024) prematurely examine survey knowledge without recognizing that effective route navigation first requires robust route knowledge. Third, **Vision Confusion**: integrating vision into geospatial tasks complicates the clear attribution of LLM performance to either perceptual reactions or internal geospatial reasoning (Feng et al., 2024). In contrast, route reversal tasks have long served as essential metrics in geospatial cognition research, predating the emergence of LLMs (Mallot, 2024; Allison and Head, 2017; Donald Heth et al., 2002; Coutrot et al., 2022), as they strictly involve route-based cognition without reliance on vision or textual knowledge.

PathBuilder: Researchers have made many advances in virtual reality, fintech and text generation on the topic of how to transform formal language between natural language (White et al., 2024; Yin et al., 2024; OpenAI et al., 2024). In the field of route navigation, there are already mature solutions for convert route geometry to natural language (Bast et al., 2016; Schumann and Riezler, 2021). Since the introduction of LLMs, research in another direction has become particularly important - LLMs are not yet able to understand geometric languages directly. Our work bridges the gap in this area.

Geometric similarity calculation for polylines varies by application (Frontiera et al., 2008; Yang et al., 2022; Fu et al., 2018), and embedding methods can handle scale discrepancies in map data (Gadi et al., 2024).

LLMs confidence of generation has been shown possibly to be artificially interfered with by the prompt (Xu et al., 2024). In Huang et al. (2025), **robustness of LLMs generations** is explored to prove LLMs suffer from uncertain reponses in different task. For **misalignment**, pervious research have discussed different alignment methods and their effects which is important to the safety of LLMs (Shen et al., 2023; Wolf et al., 2023).

B Data Generation



(a) Route generation (b) Route reversal

Figure 5: (Left): 50 routes generated in Toronto between 500 and 2500 meters in length. (Right): The red route between S and D represents the original route. Because it is generated by the routing engine, its optimal reverse route is also itself. The other returned routes, though valid reversals, differ by varying degrees of similarity.

C Similarity Metrics

Table 8: Geographical Measurement Metrics and Their Descriptions

Metrics	Description
Length Ratio	The ratio of the length of a path or curve to a reference length, commonly used to compare the relative sizes of paths.
Hausdorff Distance	The maximum distance between two point sets, defined as the greatest distance from a point in one set to the nearest point in the other set. Used to assess the spatial deviation between two paths.
Fréchet Distance	A similarity measure between curves that accounts for both location and orientation, often described as the minimum "travel distance" required for two entities to traverse their respective curves simultaneously. It captures the overall similarity in the shape and traversal order of the paths.
Edit Distance	A metric that measures the number of single-point edits (insertions, deletions, or substitutions) required to transform one path's point sequence into another. It evaluates the sequential similarity between two paths.
Jaccard Index	A statistic for comparing the similarity and diversity of sample sets, defined as the size of the intersection divided by the size of the union of the sample sets. In path similarity, it measures the overlap between the regions covered by two paths.
Angle	The geometric measure of the rotation between two intersecting lines or vectors, typically expressed in degrees or radians. In path similarity, it quantifies the directional difference between corresponding segments of the paths.
Sum of Coordinate Offsets	The total sum of the differences between corresponding coordinates (typically in the x and y dimensions) of two geometric objects or paths. This measure is useful for assessing the overall displacement between the objects.

Table 9: Route properties for different levels and different cities.

Difficulty Level	City	Length (SE) [m]	Turns (SE)	Complexity
Easy	Toronto	917.23 (3.63)	4.12 (0.04)	0.22
	Denver	923.91 (3.72)	4.26 (0.03)	0.21
	Mexico City	912.03 (3.56)	4.50 (0.04)	0.24
	São Paulo	935.01 (3.78)	4.62 (0.03)	0.25
	London	903.10 (3.44)	4.38 (0.04)	0.23
	Munich	946.53 (3.69)	4.51 (0.04)	0.22
	Tokyo	877.73 (3.47)	5.15 (0.04)	0.27
	Singapore	895.37 (3.66)	4.89 (0.04)	0.26
	Sydney	947.95 (3.56)	4.23 (0.03)	0.22
	Auckland	919.00 (3.75)	4.38 (0.03)	0.23
	Cairo	908.00 (3.59)	4.71 (0.04)	0.25
	Cape Town	915.00 (3.50)	4.69 (0.03)	0.24
Medium	Toronto	1612.63 (5.41)	7.54 (0.05)	0.48
	Denver	1623.94 (5.28)	7.23 (0.05)	0.46
	Mexico City	1598.83 (5.56)	7.79 (0.06)	0.50
	São Paulo	1642.94 (5.66)	7.87 (0.05)	0.49
	London	1604.04 (5.16)	7.41 (0.05)	0.47
	Munich	1681.92 (5.53)	7.73 (0.06)	0.49
	Tokyo	1596.48 (5.06)	8.12 (0.06)	0.52
	Singapore	1612.33 (5.25)	7.96 (0.06)	0.51
	Sydney	1635.27 (5.72)	7.34 (0.05)	0.47
	Auckland	1627.00 (5.47)	7.56 (0.05)	0.48
	Cairo	1607.00 (5.13)	7.86 (0.05)	0.50
	Cape Town	1618.00 (5.41)	7.68 (0.05)	0.49
Hard	Toronto	2142.27 (8.03)	12.77 (0.07)	0.72
	Denver	2122.26 (7.66)	12.39 (0.06)	0.70
	Mexico City	2117.62 (7.88)	12.91 (0.08)	0.74
	São Paulo	2154.02 (8.06)	13.26 (0.07)	0.73
	London	2105.27 (7.50)	12.53 (0.07)	0.71
	Munich	2202.91 (8.00)	13.18 (0.08)	0.71
	Tokyo	2076.61 (7.34)	13.58 (0.07)	0.76
	Singapore	2095.83 (7.53)	13.32 (0.07)	0.75
	Sydney	2136.15 (7.59)	12.67 (0.07)	0.72
	Auckland	2129.00 (7.69)	12.68 (0.07)	0.71
	Cairo	2110.00 (7.47)	13.02 (0.07)	0.73
	Cape Town	2125.00 (7.66)	12.87 (0.07)	0.72

D PathBuilder discussion and evaluation

In this section, we will evaluate the PathBuilder using the generated dataset. We choose three cities: Tokyo, Munich, and Toronto as the dataset generation locations for this work. This is mainly due to the fact that these cities are spread over three continents, have large populations, and distinct urban scenes.

As shown in Table 10, the PathBuilder performs well in all cities. The success rate in these three cities also matches the complexity characteristics of their urban road networks. For example, Tokyo is often considered a city with a very narrow and complex road network (Usui, 2018). On the other hand, the design of the North American urban road network, represented by Toronto, is generally considered to be more regular (King et al., 2020). While the PathBuilder is able to handle the majority of cases, there are still some circumstances that prevent it from achieving perfect results. These include theoretically unattainable bottlenecks as well as technical problems that we have not yet solved such as roundabout, but the attained performance is sufficiently good for our task. The details are provided in Appendix E.

City	Number	Length (m)	Success (%)
Toronto	6000	1670	96
Tokyo	6000	1422	90
Munich	6000	1733	94

Table 10: Length is the average length of all routes; the success rate means it passes the similarity check with a threshold of 85%.

E PathBuilder Limitations

E.1 Information Loss and Semantic Ambiguity

There have been studies demonstrating the existence of loss of information transfer between (and within) different expression and coding systems (Tsvetkov, 2013; Pang et al., 2019; Cai et al., 2018; Shannon, 1993). In our work, geospatial information has actually suffered some loss before entering the PathBuilder: navigation information expressed in natural language does not fully express the original geometric information. In the navigation information from OpenRouteService, the path engine typically describes the magnitude of a turn as "slightly" or "sharply" (GIScience, 2024). For example, a turn with an angle between 11° and 44° is called "slightly" in the output instructions. In practice, such information is valid with the added tolerance of human vision in the human eye. However, this does not always apply when transforming from natural language to geometry. In our experiments, we found many incorrect corner cases at boundary values. This type of error can only be optimized by repeated experiments but is very difficult to eradicate. This explains most of the PathBuilder errors.

E.2 Extremely high density road network

For these reasons, PB cannot work accurately in a network with high density and accuracy requirements. As shown in Figure 6(a), in a train station network in Germany, PB performed poorly despite the fact that the design of the walkway is correct.

E.3 Mismatch between network nodes and actual streets

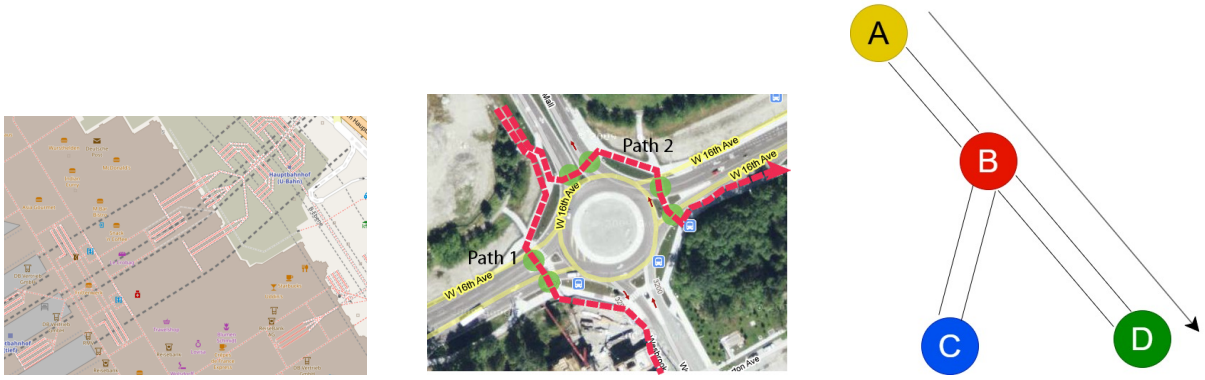
As shown in Figure 6(c). Although ABD can be considered as a single street for a pedestrian, point B is treated as a Node in the road network database which means that geometrically BC and BD are two parts. So in reality, if the PB is instructed to "go straight" from A to D through the entire street, it will have to stop at C because it will have to make a "turn" judgment here in the database. Although it is possible to approximate the "next street" implied by the similarity of the angles, but this inevitably weakens the accuracy.

E.4 Roundabout

Although rarer, traffic circles are actually scenarios that cannot be resolved in walking mode. The navigation instructions for traffic circles are highly dependent on visual information such as "third exit", which is very difficult to accomplish from a geospatial perspective. For example, as shown in Fig 6(b), current navigation instructions are highly related to vision information such as "Leave roundabout the first exit on your right."

F Models Confidence Check

It is well established that humans are more vulnerable to misinformation in complex road networks (Khademi and Saedi, 2019). Consequently, it is crucial to gauge the level of confidence LLMs have in their responses to better understand their susceptibility to misinformation.



(a) High Complexity Road Network (b) Roundabout (c) Node of Street

Figure 6: Some corner cases remain unsolved or have a low accuracy in the PathBuilder

A previous study by Xu et al. (2024) used token probabilities to estimate how confident LLMs are in their answers, thus offering insight into the models' susceptibility to misinformation. However, because route reversal is not a simple Boolean question, we cannot directly mimic that approach by extracting token probabilities for "yes" or "no."

In this paper, we adapt their method to accommodate for our task. Specifically, for each navigation instruction generated by LLMs, we compute the confidence by extracting the token probability of the direction word (e.g., "North") that follows the "turn" action. We then approximate the overall confidence for the entire set of navigation instructions by taking the average of all such direction-word probabilities as Equation 5 shows.

$$Confidence_{set} = \frac{1}{N} \sum_{i=1}^{N} DirectionProb_i \quad (5)$$

where N is the total number of direction instructions, and $DirectionProb_i$ is the token probability of the direction word in the i-th instruction. This probability can be calculated from the log probabilities record of LLMs.

G Experiment setup

The APIs used for all closed-source models are the official APIs for the February 1, 2025 model, with default parameters except for the temperature control, which is zero. For the open-source models, the experimental models were taken from the official version of Hugging Face, and the temperature was also set to 0. The open-source models were set in a training environment of 8 RTX 4090s.

The dataset generation time was about two weeks, mainly due to the limitation on the number of requests from openrouteservice.

H Route reversal performance for GPT4o under different temperatures

To investigate the impact of the temperature parameter on LLMs in the route reversal task, we performed an ablation study using GPT4o at various temperature settings, evaluated on the all city dataset at the easy difficulty level, see Table 11. As expected, lower temperature values produced more deterministic responses, yielding better overall performance. However, even at the lowest temperature (0.0), GPT4o exhibited notable randomness, particularly reflected in robustness and misalignment scores. Thus, we conclude that current LLMs, including GPT4o, lack sufficient spatial cognitive capability to reliably generate consistent route reversal responses. To minimize such randomness and ensure comparability across models, we adopt a temperature setting of 0.0 for all subsequent experiments where applicable. Additionally, as this benchmark aims to fairly assess baseline performance of existing LLMs, no domain adaptation or detailed prompt engineering was employed to artificially boost results.

I Landmark knowledge validation

Although there have been clear indications that LLMs are able to understand landmark knowledge and extract Points of Interest (POIs) (Kim and Lee, 2024; Liu et al., 2024), few datasets have been able to comprehensively examine their abilities. To prove that LLMs adequately grasp Landmark

Table 11: Route reversal performance for GPT4o under different temperature settings. The evaluation was conducted on the all city dataset at easy difficulty level.

Temperature	Return Rate (%)	Similarity	Robustness	Confidence	Misalignment
0.0	6.70	41.12	43.50	92.60	14.20
0.1	6.50	41.01	42.90	92.30	14.70
0.2	6.40	40.76	42.59	93.10	15.10
0.3	6.20	40.30	42.31	92.77	15.40
0.4	6.10	40.12	41.70	93.21	16.20
0.5	6.10	40.00	40.96	92.78	16.70
0.6	6.10	39.96	40.42	91.87	17.10
0.7	5.90	39.44	40.05	93.15	17.40
0.8	5.80	38.78	39.87	92.67	17.60
1.0	5.50	38.54	39.23	92.31	17.90

Knowledge, we decide to run a small scale of geospatial landmark knowledge test.

Our dataset contains 100 questions with three dimensions: detailed information about popular Landmarks (40%), simple information about common Landmarks (30%), and coordinates transformations related tests (30%). A large part of our questions is inspired by Yang et al. (2025).

All responses were manually verified and the results are shown in Table 12. The specifics of the questions are available in Appendix I.

Table 12: Performance of LLMs in landmark knowledge

LLM	Popular	Common	Geocode
ChatGPT-4o	100%	57%	0%
Llama3.3-70B	100%	40%	0%
Gemini2.5-pro	100%	80%	6%
ChatGPT-o3	100%	81%	3%
Claude3.5	100%	73%	0%
Claude3.7	100%	80%	3%
Deepseek	100%	70%	0%
Deepseek(R1)	100%	78%	3%
Grok	100%	66%	0%

As shown in Table 12, the LLMs show a good grasp of popular landmarks. However, they perform poorly when it comes to lesser known locations. The LLMs showed a lack of knowledge when the questions became more complicated, that would require geospatial cognition, i.e. route and survey level knowledge.

Landmark Knowledge Questions Selections

The following questions are selected from a pool of 50 questions, organized into three domains: Popular Landmarks, Common Landmarks, and Geocoding. Each domain contains 5 representative questions that demonstrate typical challenges and considerations in geospatial knowledge about landmark.

A. Popular Landmarks

This section contains questions related to well-known landmarks and their basic information.
Q1: Where is the Eiffel Tower?
Q2: How big is the white house?
Q3: How tall and how heavy is the Status of Liberty?
Q4: In which year did Colosseum built?
Q5: Where is the British Museum located in London?

B. Common Landmarks

This section focuses on everyday landmarks such as intersections, buildings, and natural features.
Q1: Where is the Cathedral Church of Our Lady in Germany?
Q2: What are the names of the streets that connect Russell Square in London?
Q3: What district of Paris is Rue Saint-Jacques in?
Q4: Is Clerkenwell Road on the north or south bank of the Thames?

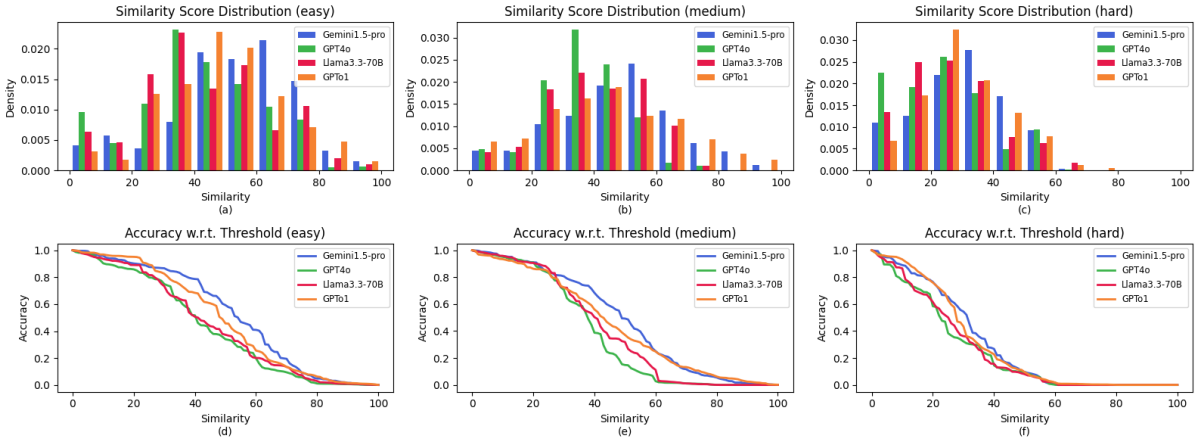


Figure 7: Performance visualization of different models across difficulty levels. The first row (a-c) shows the similarity score distributions for easy, medium, and hard datasets respectively. The second row (d-f) presents the accuracy at different similarity thresholds for each difficulty level. Each color represents a different model: Gemini2.5-pro (blue), GPT4o (green), Llama3.3-70B (red), and GPTo3 (orange).

Q5: How many McDonald's are there in Toronto?

C. Geocoding of Landmarks

This section addresses specific challenges in geocoding and coordinate verification.

Q1: Give the coordinates of any McDonald's in Toronto.

Q2: What are the corresponding coordinates for 317 Dundas St W, Toronto, Canada?

Q3: Where is 48°52'20.8"N, 2°18'15.0"E? What's nearby, the more precise the better?

Q4: How far is 21 West End Ave, New York from 20 West End Ave, New York?

Q5: The straight line connecting 47°21'13.2"N, 3°5'10.8"E with 43°43'14.3"N, 39°51'18.7"E passes through which countries on the surface?

J Prompt

As shown in Figure 8, the guide prompt aims to help LLMs locate road networks and set the rules for route reversal. Note that we emphasized twice that the instruction in the first step must include the absolute direction, since we found in early experiments that the probability of LLMs violating the rule was much higher with just one emphasis. In addition to this, we require that LLMs use the same language style as the examples in prompt as well as no semantic inversions. However, in the final result, LLMs only adhered well to the former. It is also worth mentioning that we asked the LLMs to output some information about the surface visible along the way. In the manual evaluation, we did not find any correct examples, and in fact, LLMs tended not to answer this question.

As shown in Figure 9, the reason for providing the latitude and longitude is to help LLMs smoothly locate the original route. Also, this further enhances the rigor of the evaluation of the geospatial aspects of LLMs. In the literature previously addressed, LLMs do possess the ability to extract some information from geographic systems. If this ability does come from an understanding of spatial relationships, and not just from textual training, then they should be able to accurately recognize the location of the routes.

Guide Prompt for Route Reversal

Generate a road network for `[CITY NAME, COUNTRY NAME]` based on your knowledge.

The following task involves reversing a navigation route from destination (D) back to start point (S). Follow these key requirements:

1. **Start with absolute direction.** Use precise cardinal directions (North, South, East, West). Avoid ambiguous terms like "head backward."
2. **No simple inversion.** Understand the route thoroughly and create logical return directions rather than merely reversing steps.
3. **Maintain consistent format.** Use standard navigation terms ("head," "turn," "continue," "arrive") as in the original directions.
4. **Reference landmarks.** Include nearby points of interest (POI) to demonstrate geographical context.
5. **Begin with absolute direction.** The first instruction must specify an absolute direction (non-negotiable).

Example: . . . . . .

Figure 8: Guide prompt template for route reversal task, detailing the key requirements for generating reversed navigation instructions. The template emphasizes absolute directions, logical route understanding, format consistency, step matching, landmark referencing, and mandatory absolute direction start.

Instruction Prompt for Route Reversal

Start Point: 43°38'47"N, 79°26'11.5"W

1. Head west, continue for 75.9 meters.
2. Turn slight right, continue for 37.7 meters.
3. Turn left, continue for 11.3 meters.
4. Turn right, continue for 126.3 meters.
5. Keep right, along Queen Street, continue for 91.7 meters.
6. Keep right, continue for 146.6 meters.
7. Turn slight left, continue for 2.3 meters.
8. Turn right, continue for 18.8 meters.
9. Turn right, continue for 198.7 meters.
10. Turn left, continue for 4.8 meters.
11. Turn right, continue for 18.7 meters.
12. Turn right, continue for 2.1 meters.
13. Keep left, continue for 26.4 meters.
14. Straight ahead, then arrive at your destination.

Figure 9: Example navigation instructions for route reversal task, showing a detailed route in Toronto with precise coordinates and step-by-step directions including distance measurements.

Figure 10: Prompts used in the route-reversal experiment: (a) guide prompt template with detailed requirements and (b) concrete example instructions illustrating the task.