MobQA: A Benchmark Dataset for Semantic Understanding of Human Mobility Data through Question Answering

Anonymous ACL submission

Abstract

This paper presents MobQA, a benchmark dataset designed to evaluate the semantic understanding capabilities of large language models (LLMs) for human mobility data through natural language question answering. While existing models excel at predicting human movement patterns, it remains unobvious how much they can interpret the underlying reasons or semantic meaning of those patterns. MobQA provides a comprehensive evaluation framework for LLMs to answer questions about diverse human GPS trajectories spanning daily to weekly granularities. It comprises 5,800 high-quality question-answer pairs across three complementary question types: factual retrieval (precise data extraction), multiple-choice reasoning (semantic inference), and free-form explanation (interpretive description), which all require spatial, temporal, and semantic reasoning. Our evaluation of major LLMs reveals strong performance on factual retrieval but significant limitations in semantic reasoning and explanation, with trajectory length substantially impacting effectiveness. These findings demonstrate both achievements and limitations of state-of-the-art LLMs for semantic mobility understanding.

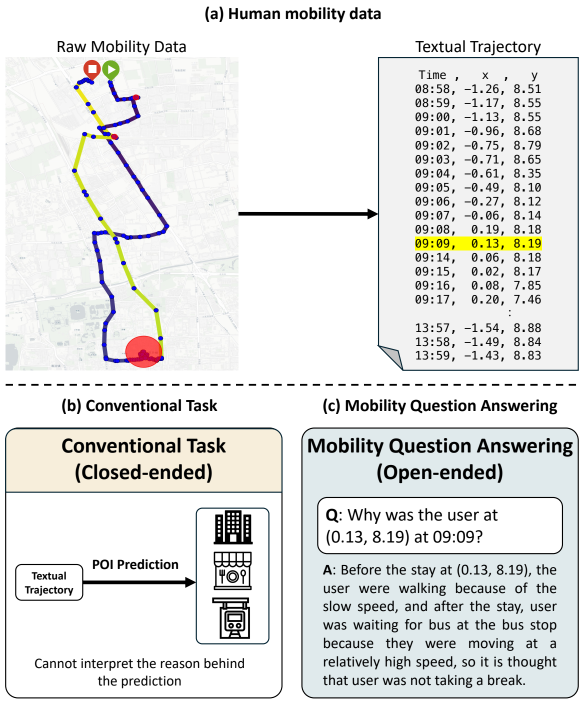


Figure 1: MobQA presents the mobility question answering benchmark. (a) Raw GPS trajectory is anonymized and converted into a textual sequence. (b) Conventional tasks are prediction-oriented and closed-end. (c) Mobility question answering evaluates factual and semantic understanding through natural language.

1 Introduction

Modern positioning technologies have enabled large-scale collection of high-resolution human mobility data, but their numerical form makes it difficult for non-experts to interpret or exploit such data (Parent et al., 2013; Albanna et al., 2015). When examining timestamped GPS trajectories like the one shown in Fig. 1a, it is not straightforward to recognize the mode of transportation, the purpose of the trip, or the reasons behind certain movements. Developing intelligent systems that can understand mobility data from such semantic perspectives would enhance various applications in urban analysis, transportation, and public health (Shang et al., 2011; Askarizad and Safari, 2020; Badr et al., 2020; Chang et al., 2021) by providing deeper insights into individual daily activities beyond simple crowd movement statistics.

Prior work has focused primarily on closed-ended, predictive modeling tasks such as next-location prediction and point-of-interest (POI) prediction (Liu et al., 2016; Feng et al., 2020; Liang et al., 2022; Xu et al., 2024) (see also Fig. 1b). These methods are effective at identifying patterns and trends in movement data. However, they typically do not address the underlying reasons behind these patterns, such as user motivations or trip purposes (Xue and Salim, 2023; Gong et al., 2024). Indeed, recent literature attributes this limitation to the lack of dedicated datasets and benchmark protocols for addressing the semantic aspects of human mobility data (Luca et al., 2021; Pappalardo et al., 2023; Mokbel et al., 2024; Zhang et al., 2024).

To address this gap, we present *MobQA*, a new human mobility dataset designed to evaluate intelligent systems through natural language question answering, *i.e.*, *mobility question answering*. Natural language is a flexible and expressive modality for interacting with human mobility data. It allows users to pose open-ended questions about mobility patterns, such as *"Why was the user at this location at this time?"* as illustrated in Fig. 1c. Furthermore, large language models (LLMs) have recently shown remarkable capabilities in understanding and generating natural language sentences, making them well-suited for interpreting complex mobility data when properly represented as text.

The MobQA Dataset is built upon the Geolife dataset (Zheng et al., 2008, 2010, 2011), selected for its public availability, diverse transportation modes, and dense spatiotemporal sampling, and comprises 5,800 high-quality question-answer pairs spanning daily and weekly granularities. The questions fall into three types: 2,000 factual retrieval questions, 2,000 multiple-choice questions for semantic inference, and 1,800 free-form questions for interpretive explanation. We also establish comprehensive evaluation protocols using accuracy for factual and multiple-choice tasks, and an LLM-as-a-judge framework (Zheng et al., 2023; Gu et al., 2024) to assess the faithfulness and informativeness of free-form answers, following established practices (Sai et al., 2022; Adlakha et al., 2024).

We conducted extensive experiments with various LLMs, including GPT-4o series, Gemini 1.5/2.0, o3-mini, and ten open-source models from the Llama 3 series (Meta, 2024), Qwen 3 (Team, 2025), and DeepSeek-R1 series (DeepSeek-AI, 2025), to evaluate their performance on mobility question answering tasks. Our results reveal that while these models achieve strong performance on factual retrieval question answering, they demonstrate significant limitations in semantic reasoning and explanation tasks, with performance substantially degrading for longer trajectory sequences. We believe that MobQA will encourage the research community to develop more explainable and accessible mobility understanding technologies. [1]

[1]The MobQA dataset is available at `https://anonymous.4open.science/r/mobqa`.

2 Mobility Question Answering

2.1 Task Description

We address the task of mobility question answering. This task challenges LLMs to reason over human mobility data and answer natural language questions that require spatial, temporal, and semantic understanding of movement patterns. Formally, given a question Q about a user's movement trajectory T, the goal is to generate a natural language answer A as $(Q, T) \rightarrow A$.

The trajectory T serves as the primary data source, represented as a textual sequence of timestamped coordinates such as `(09:02, -0.75, 8.79)` (see also Fig. 1a). LLMs must infer spatiotemporal patterns and semantic cues *solely from this numerical sequence* T to produce accurate answers A to the question Q.

2.2 Trajectories

Human trajectories are typically long sequences; collecting GPS data every minute yields 1,440 points for a single day. Even after removing stationary moments to reduce redundancy, daily trajectories can still comprise several hundred points. This poses challenges for current LLMs that can struggle with such long sequences of numerical data (Kim et al., 2024; Liu et al., 2024; Yoon et al., 2024). To systematically examine the impact of input sequence length on LLM performance, we consider two trajectory granularities:

- **Daily Trajectory**: A single 24-hour period, capturing intraday behaviors such as commuting, shopping, and short-term activities.
- **Weekly Trajectory**: A seven-day sequence providing richer behavioral context and revealing weekly habits or anomalies.

While weekly trajectories provide richer behavioral evidence, they include more data points. Evaluating models on questions with those two granularities allows us to quantify the trade-off between contextual richness and sequence length.

2.3 Questions

To evaluate LLMs across different reasoning demands, we design three complementary question types inspired by existing benchmarks (Baradaran et al., 2022; Zhong et al., 2022; Fang et al., 2024a): *factual retrieval*, *multiple choice*, and *free form*. Fig. 2 shows statistics and examples for each type.

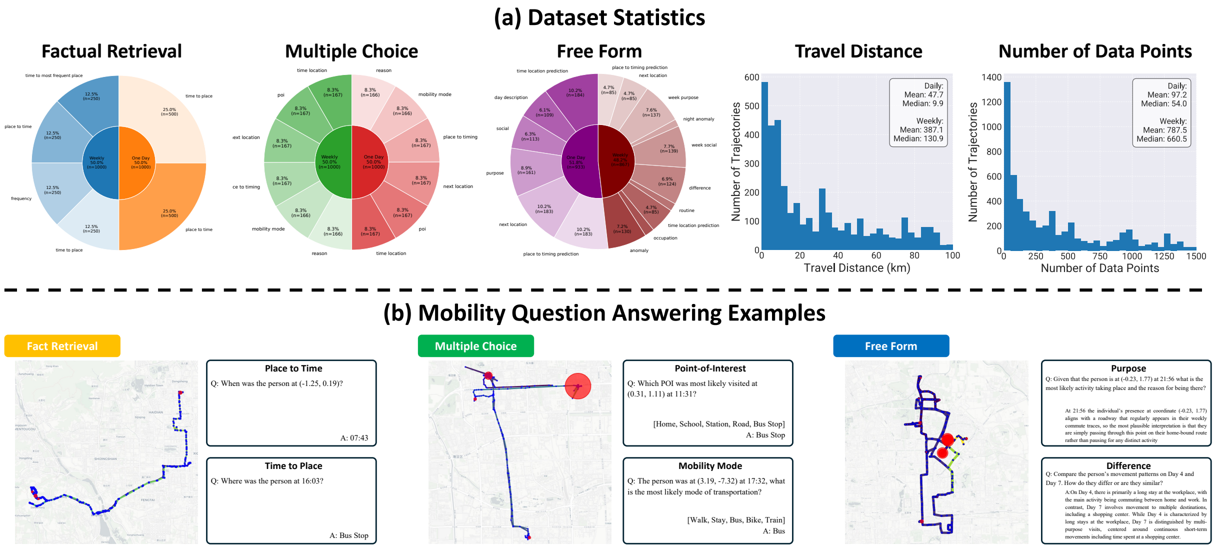


Figure 2: (a) Distribution of questions across the three categories: factual retrieval, multiple choice, and free form, as well as the distributions of trajectory length (number of data points) and travel distance in the MobQA dataset. (b) Representative examples of questions and answers for each category, along with the corresponding trajectories.

- **Factual retrieval questions** evaluate an LLM's ability to *locate and extract verifiable facts* from a trajectory. The key challenge is pinpointing the relevant data point in a long sequence of similar entries. The input-output interface follows $(Q, T) \rightarrow A$, where the answer is a precise value (*e.g.*, coordinates, times, counts). Example: *"Where is the user at 10:15?"*

- **Multiple-choice questions** ask LLMs to *infer hidden patterns or semantic attributes*, such as mobility mode or the next POI, from numerical trajectories. Each instance extends the standard (Q, T) pair with five candidate answers (one correct and four distractors), requiring the LLMs to choose the most plausible option. Example: *"What is the mobility mode at 10:15 at (-0.15, 4.91)?"* with candidates `[Walk, Stay, Bus, Bike, Train]`.

- **Free-form questions** require LLMs to generate *natural language descriptions or explanatory interpretations* from the trajectory. Models must abstract complex patterns, detect anomalies, and infer underlying intentions. The output is unconstrained text following the interface $(Q, T) \rightarrow A$. Example: *"Summarize Day 7's mobility and explain any deviations from the weekly routine."*

Consequently, our benchmark offers a comprehensive assessment with tasks ranging from basic data retrieval to complex spatiotemporal reasoning and description generation.

2.4 Evaluation Methodology

We evaluate the performance of LLMs on mobility question answering using metrics tailored to each question type. For factual retrieval and multiple-choice questions with precise answers, we use the *accuracy* metric, which calculates if the answer exactly matches the ground truth values or choices. On the other hand, we adopt *faithfulness* and *informativeness* metrics (Sai et al., 2022; Adlakha et al., 2024) for the free-form questions because they are subjective and require evaluating both factual accuracy and semantic richness:

- **Faithfulness** assesses how accurately the generated answers reflect factual details (*e.g.*, coordinates, timestamps) from the trajectory, ensuring alignment and no contradictions.

- **Informativeness** evaluates the completeness and richness of the response, verifying that it comprehensively addresses the question by incorporating relevant movement patterns and inferred semantic insights.

To evaluate free-form responses, we employ an *LLM-as-a-judge framework* (Zheng et al., 2023; Gu et al., 2024) that uses powerful LLMs to score generated responses (see Sec. 4.1 for more details.)

3 MobQA Dataset

We construct the MobQA dataset from an existing human trajectory dataset (hereafter, the base dataset) following the pipeline illustrated in Fig. 3.

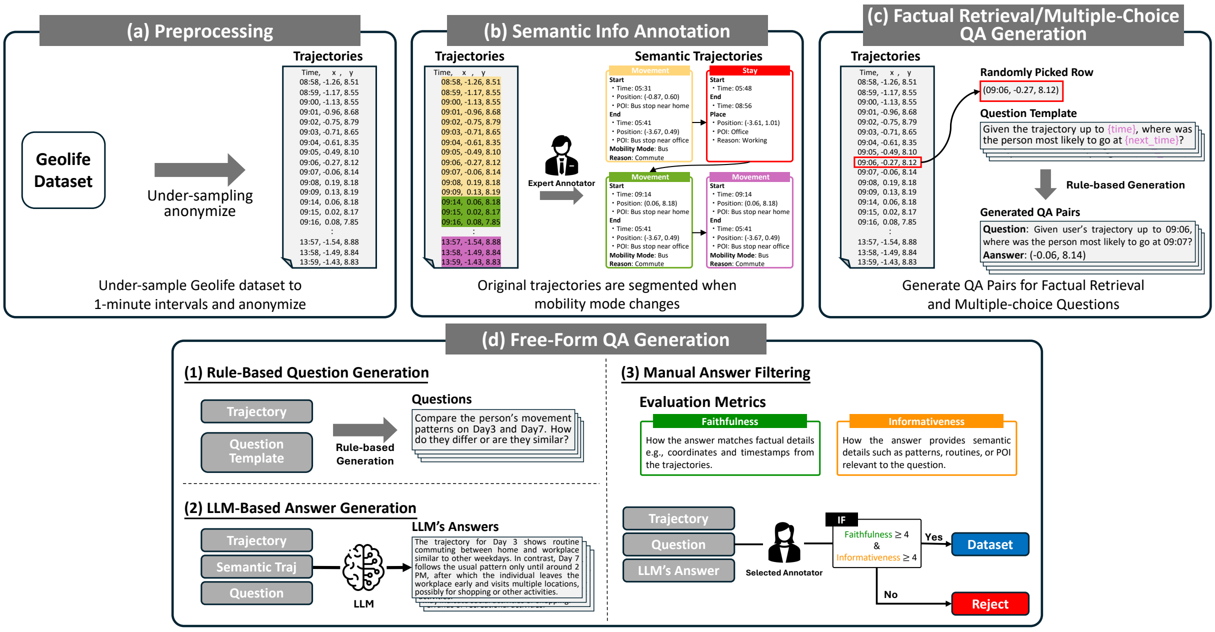


Figure 3: Overview of the MobQA dataset construction process, illustrating the stages of data preprocessing, semantic trajectory construction, question generation, answer generation, and answer filtering.

3.1 Constructing Trajectory Data

Requirements for Base Dataset To construct a benchmark capable of evaluating semantic understanding of human mobility, the underlying trajectory data must satisfy three essential criteria:

- **R1: Dense Sampling over Long Horizons.** The data must possess high spatiotemporal resolution (e.g., minute-level GPS logs) to capture fine-grained behaviors while covering long time horizons to analyze routine patterns.
- **R2: Multi-mode Mobility Data.** The dataset must include diverse transportation modes beyond a single vehicle type (*e.g.*, taxi, bus).
- **R3: Public Availability.** The dataset must be publicly accessible so that the community can reproduce and verify the results.

Dataset Selection We systematically evaluate existing datasets and relevant literature (Luca et al., 2021) against the aforementioned requirements. As summarized in Tab. 1, no existing dataset category simultaneously satisfies all three requirements. Appendix B provides a detailed evaluation.

Therefore, we concluded that the **Geolife GPS dataset** (Zheng et al., 2008, 2010, 2011) is the **only publicly available dataset that satisfies all criteria**. Geolife offers a unique combination of high spatiotemporal resolution GPS trajectories recorded at frequent intervals, verified multi-modal

Table 1: Comparison of publicly available human mobility dataset categories against MobQA requirements.

Dataset Category	Representative Examples	R1	R2	R3
Single-Mode Datasets	NYC Taxi, T-Drive	×	×	✓
Sparse Check-in Logs	Foursquare, Gowalla	×	×	✓
Infrastructure Traces	D4D, MIT Reality Mining	×	×	✓
Coarse-Grained Trajectory	YJMob100K	×	✓	✓
Restricted Mob Dataset	Priva'Mov, Nokia MDC	✓	✓	×
Geolife	-	✓	✓	✓

transportation labels across 11 transport modes (*e.g.*, walking, biking, subway), and diverse mobility behaviors spanning commutes, leisure, and long-distance trips. With longitudinal coverage exceeding 5 years for 182 users, the dataset uniquely enables the generation of semantically rich question-answer pairs that demand understanding of complex spatiotemporal movement patterns.

Preprocessing We create the input textual trajectory T by converting raw GPS data into a structured format suitable for LLMs. From the Geolife dataset, we retain 69 participants with user-reported mobility-mode labels from the original 182 users. Raw data consists of spatiotemporal movement records represented as ordered triples of geographic coordinates and timestamps: $\{(x_i, y_i, t_i) \mid i = 1, \ldots, n\}$, where (x_i, y_i) denote the latitude and longitude coordinates of the i-th GPS observation and t_i represents the corresponding timestamp. We downsample logs to 1-minute intervals and merge stationary periods into single events, reducing data volume while preserving temporal resolution. Despite these preprocessing steps, our dataset features long sequences; daily trajectories average 97.2 data points, while weekly trajectories average 787.5 data points, as shown in Fig. 2.

Privacy-Preserving Processing To enhance privacy protection and standardize trajectory coordinate, all geographic coordinates are transformed into anonymized offsets (x'_i, y'_i) relative to a randomly selected user-specific reference point $(x_{\text{ref}}, y_{\text{ref}})$, computed as $(x'_i, y'_i) = (x_i - x_{\text{ref}}, y_i - y_{\text{ref}})$. This coordinate anonymization protects users' privacy by removing personally identifiable location information, while preserving the essential geometric structure and spatial relationships necessary for trajectory analysis.

Textual Trajectory Representation Each trajectory is converted into plain text sequences to be compatible with LLM processing. Specifically, we represent each data point as a tuple of the form (hh:mm, x'_i, y'_i), where hh:mm denotes the time in hours and minutes, and x'_i, y'_i are the anonymized coordinates rounded to two decimal places (approximately 10-meter resolution). The resulting representation is a sequence of such tuples, separated by newline characters (\n), for example:

```
(09:02, -0.75, 8.79)\n(09:03, -0.71,
  8.65)\n. . . (14:00, -1.38, 8.78).
```

See Fig. 12 in Appendix D.3 for a complete daily trajectory example.

3.2 Annotating Semantic Information

We enrich a subset of the trajectory data (eight consecutive days from 19 users, selected to maximize geographic coverage, routine diversity, and transport mode variety) with manual annotations to generate question-answer pairs requiring detailed semantic understanding of mobility data.

Three professional annotators specialized in mobility-data annotation performed the labeling using a dedicated interactive annotation tool we developed to facilitate minute-level precision (see Appendix E.1 for more details). As shown in Fig 3b, each daily trajectory is segmented when the *mobility mode* changes (*e.g.*, from walking to bus) or a *stay phase* (*e.g.*, at home or work) begins or ends. The annotators then label each segment with:

- **Point of Interest (POI):** semantic location types (*e.g.*, home, workplace, shopping center)
- **Mobility Mode:** transportation methods (*e.g.*, walking, bus, subway, car)[2]
- **Reason:** underlying purposes for the movement or stay (*e.g.*, commuting, leisure, shopping)

All the annotators were thoroughly briefed on the study purpose and provided informed consent for their annotations to be included in the publicly available dataset. The annotation protocol involved two structured practice phases using pre-annotated exemplar trajectories.

The lead author performed final validation for quality assurance of all annotations to ensure consistency in mobility modes, precise segment boundary alignment, and completeness. This process required approximately 26 minutes per daily trajectory and 60 hours total, yielding high-fidelity semantic annotations covering 11 verified transport modes including walking, biking, subway, and car, spanning diverse mobility behaviors from commutes and leisure activities to long-distance trips. See Appendix E for details.

3.3 Creating Question-Answer Pairs

Factual Retrieval Questions Factual retrieval questions assess the ability of LLMs to extract precise information directly from trajectory data. These questions require exact data retrieval without interpretation, such as determining a person's location at a specific time or counting visit frequencies. Specifically, we employ a template-driven pipeline using predefined question templates:

- `time_to_place`: "Where was the person at {time}?"
- `frequency`: "How many times did the person visit the location at coordinates {place}?"

Our pipeline involves: (1) uniformly sampling a trajectory suitable for the target question type, (2) selecting a data point from the sampled trajectory, and (3) filling template placeholders with extracted attributes to formulate the question. Answers are produced deterministically by copying requisite fields or applying simple rule-based operations.

Multiple-Choice Questions Multiple-choice questions evaluate semantic inference capabilities, requiring LLMs to understand context and make reasoned selections from provided options. These

[2] We re-annotated mobility modes for this subset of trajectories because the original, self-reported mobility modes of Geolife dataset are often temporally inaccurate (off by up to one minute) or sometime missing.

questions assess understanding of mobility modes, points of interest, purposes, and predictive reasoning. Similar to factual retrieval, we use templates for question generation.

- `poi`: "Given that the person was at {place} at {time}, which POI did they most likely visit?"
- `next_location`: "Given the trajectory up to {time}, where was the person most likely to go at {next_time}?"

We sample questions from the full corpus of trajectories if they can be answered from raw GPS traces. For questions requiring semantic inference (*e.g.*, mobility mode, activity purpose), we use the semantically enriched subset from Sec. 3.2 to ensure that the necessary semantic information is available for verifying the answers. For each question, we crafted four plausible yet incorrect distractors. Numeric distractors are drawn from the same trajectory, while semantic distractors are selected from the established vocabulary.

Free-Form Questions Free-form questions require comprehensive analysis and interpretation, assessing the LLMs' ability to describe mobility patterns, and infer underlying reasons for movements. We use template generation and sample questions from semantically enriched trajectories:

- `routine`: "Analyze the weekly trajectory to identify and summarize the individual's regular movement patterns and routine activities."
- `difference`: "Compare the person's movement patterns on {day1} and {day2}. How do they differ or are they similar?"

Unlike factual-retrieval or multiple-choice questions, creating answers for free-form questions cannot be done by predefined rules alone. As shown in the Fig 3d, we first leverage GPT-4o to generate fluent, evidentially grounded responses. Specifically, we feed the question tuple (Q, T) as well as semantic information annotated in Sec. 3.2 to GPT-4o to obtain candidate answers.

To ensure the quality of the dataset, we manually filtered the generated answers. We first recruited nine annotators from a crowd-sourcing platform for a pilot task and retained the three best-performing annotators for the main filtering phase. They scored the candidate answers for faithfulness and informativeness (Sec. 2.4), accepting only those with scores exceeding 4 on both criteria. See Appendix F for manual filtering process details.

Table 2: Correlation coefficients (Pearson and Spearman) between LLM and human evaluations for faithfulness and informativeness metrics

Model	Faithfulness		Informativeness	
	Daily	Weekly	Daily	Weekly
GPT-4o-mini	0.502 / **0.562**	0.278 / 0.349	0.554 / 0.605	0.402 / 0.425
GPT-4o	**0.507** / 0.524	0.401 / 0.492	0.521 / 0.511	0.513 / 0.537
Gemini 1.5	0.370 / 0.378	**0.513 / 0.548**	0.544 / 0.593	0.502 / 0.521
Gemini 2.0	0.368 / 0.401	0.494 / 0.535	**0.584 / 0.608**	**0.577 / 0.596**

4 Experiments

We conducted experiments to evaluate LLMs on the MobQA dataset. We first evaluate closed LLMs in Sec. 4.2 and open models in Sec. 4.3, then investigate how trajectory length impacts LLM performance in Sec. 4.4.[3]

4.1 Experimental Setup

Data Partitioning We partitioned the MobQA dataset (5,800 samples) into training (80%) and test (20%) sets. The training data were used for fine-tuning open models, while all evaluations were conducted on the test set.

Memory Constraints Due to GPU memory limitations, our evaluation was restricted to trajectories containing fewer than 1,000 data points. [4] All experiments were conducted on a single NVIDIA A100 GPU. Full closed LLMs' results are in Tab. 7.

Prompting Strategy We designed a prompt (see Fig. 8 for details) with five components: (1) mobility data description, (2) task-specific instructions, (3) textual trajectory representation, (4) the target question, and (5) trajectory data reference.

LLM-as-a-Judge We selected Gemini-2.0-Flash as our LLM-as-a-judge evaluator based on its strong correlation with human ratings (see also Sec. 2.4 and Tab. 2).[5] To examine potential self-enhancement bias (Zheng et al., 2023; Dai et al., 2024; Ye et al., 2025), where LLMs favor responses from their own model family, we re-scored the answers with GPT-4o. As shown in Tab. 6, the ratings suggest that this bias was negligible in our task.

[3]We also investigate the effect of semantic information on LLM performance in Sec. D.2.

[4]Although the precise token count depends on numerical tokenization, a trajectory with 1,000 points typically corresponds to approximately 20,000 tokens including prompt.

[5]Our pilot experiment confirmed that conventional metrics, such as BLEU and BERTScore, showed poor alignment with human evaluations. See Appendix C for details.

Table 3: Performance comparisons of closed LLMs

Model	Factual Retrieval		Multiple-Choice		Faithfulness		Informativeness	
	Daily	Weekly	Daily	Weekly	Daily	Weekly	Daily	Weekly
GPT-4o-mini	0.690	0.462	0.294	0.310	2.00	2.12	1.96	2.13
GPT-4o	0.812	0.550	0.413	0.457	2.09	2.35	2.10	2.43
o3-mini	0.824	0.673	0.472	**0.505**	2.45	2.78	2.42	2.86
Gemini 1.5	0.963	**0.880**	**0.490**	0.449	**3.15**	2.86	**3.13**	2.97
Gemini 2.0	**0.991**	0.858	0.403	0.405	2.96	**3.10**	2.90	**3.23**

Table 4: Performance comparisons of open LLMs and their fine-tuned versions

Model	Factual Retrieval		Multiple-Choice		Faithfulness		Informativeness	
	Daily	Weekly	Daily	Weekly	Daily	Weekly	Daily	Weekly
Open Source Models								
Llama-3.2-1B	0.000	0.000	0.045	0.019	1.00	1.15	1.00	1.15
Llama-3.2-3B	0.615	0.168	0.210	0.087	1.01	1.48	1.01	1.53
Llama-3.1-8B	0.840	**0.679**	0.295	0.237	1.01	1.25	1.02	1.40
Qwen3-1.7B	0.110	0.007	0.045	0.048	1.07	1.09	1.07	1.09
Qwen3-4B	0.815	0.350	**0.560**	0.375	1.47	**1.94**	1.48	1.96
Qwen3-8B	**0.920**	0.482	0.495	0.319	1.22	1.35	1.26	1.36
Qwen3-14B	0.835	0.460	0.265	0.237	1.90	1.31	1.90	1.31
R1-1.5B	0.090	0.000	0.015	0.013	1.28	1.49	1.28	1.50
R1-7B	0.570	0.015	0.030	0.000	1.79	1.37	1.78	1.35
R1-14B	0.875	0.387	0.050	0.150	**2.36**	1.68	**2.33**	1.67
Fine-tuned Models								
Llama-3.2-1B (FT)	0.485	0.241	0.435	0.444	2.11	1.69	2.16	1.81
Llama-3.2-3B (FT)	0.730	0.328	0.505	0.562	2.10	1.61	2.18	1.67
Qwen3-1.7B (FT)	0.650	0.307	0.510	0.494	2.18	1.50	**2.33**	1.64
Qwen3-4B (FT)	0.830	0.460	**0.560**	**0.606**	2.14	1.62	2.23	**1.84**

4.2 Evaluation of Closed LLMs

We benchmark five recent LLMs: GPT-4o-mini-2024-07-18, GPT-4o-2024-11-20, o3-mini-2025-01-31, Gemini-1.5-Flash-002, and Gemini-2.0-Flash-001. All models were accessed through their public API. Details are in Appendix G.

Tab. 3 reveals distinct performance patterns across tasks and temporal granularities. For factual retrieval, Gemini models demonstrated superior accuracy, with Gemini-2.0-Flash achieving 0.991 on daily trajectories. However, all models showed performance degradation on longer weekly trajectories, highlighting challenges in maintaining factual recall as sequence length increases, which we analyze in Sec. 4.4. In contrast, multiple-choice results showed comparable performance between daily and weekly inputs across most models, suggesting task-dependent sensitivity to input length.

Free-form questions proved most challenging, with faithfulness and informativeness scores typically ranging between 2.0 and 3.0 out of 5. Notably, o3-mini's state-of-the-art reasoning capabilities did not translate to strong performance across all tasks. This aligns with recent observations that general-purpose reasoning mechanisms, such as chain-of-thought, may not readily transfer to domains requiring nuanced interpretation of dense, sequential data like time series (Zhou and Yu, 2025). These results show that generating high-quality, faithful, and informative free-form responses from mobility data remains a significant challenge for current LLMs. See Appendix I for qualitative examples.

4.3 Evaluation of Open Models

We benchmarked ten open models from the Llama 3.x (Meta, 2024), Qwen 3 (Team, 2025), and DeepSeek-R1 distilled model series (DeepSeek-AI, 2025). We also finetuned four of these models via Supervised Fine-Tuning (SFT) (Ouyang et al., 2022; Wang et al., 2023b) to evaluate specialized mobility performance. See Appendix G for details.

The evaluation results in Tab. 4 show a positive correlation between model size and performance. Smaller models, particularly in the 1B class, struggled with benchmark tasks. While larger models demonstrated improved performance, their capabilities remained limited, especially for weekly trajectories requiring long-term contextual understanding. This difficulty is evident in multiple-choice questions, where most open models performed worse on weekly trajectories compared to closed models, suggesting limited effectiveness for extracting information from longer trajectories. The model series showed varying strengths: the Qwen series excelled at multiple-choice tasks, while Qwen3-8B and Llama-3.1-8B performed well on daily and weekly factual retrieval tasks, respectively. The R1 series performed best on daily free-form tasks, highlighting our benchmark's effectiveness in assessing diverse model capabilities.

Fine-tuning demonstrated clear performance improvements over the base models, particularly in multiple-choice tasks where fine-tuned models achieved the highest scores. While factual retrieval also improved, it did not surpass larger base models. On free-form questions, the two series exhibit different trends. Llama models consistently improved through fine-tuning, whereas Qwen series shows limited gains, with the base model sometimes outperforming its fine-tuned version on weekly trajectories. Nevertheless, overall performance on free-form tasks remains limited across all models, highlighting the task's inherent difficulty, likely due to the high-level reasoning required that may be difficult to acquire through SFT alone.

4.4 Impact of Trajectory Lengths

Previous studies have shown that LLMs struggle with longer input sequences (He et al., 2024; Kim et al., 2024; Yoon et al., 2024; Lee et al., 2025). To examine this problem more closely, we investigated how trajectory length affects performance by dividing daily and weekly data into bins (Fig. 4).

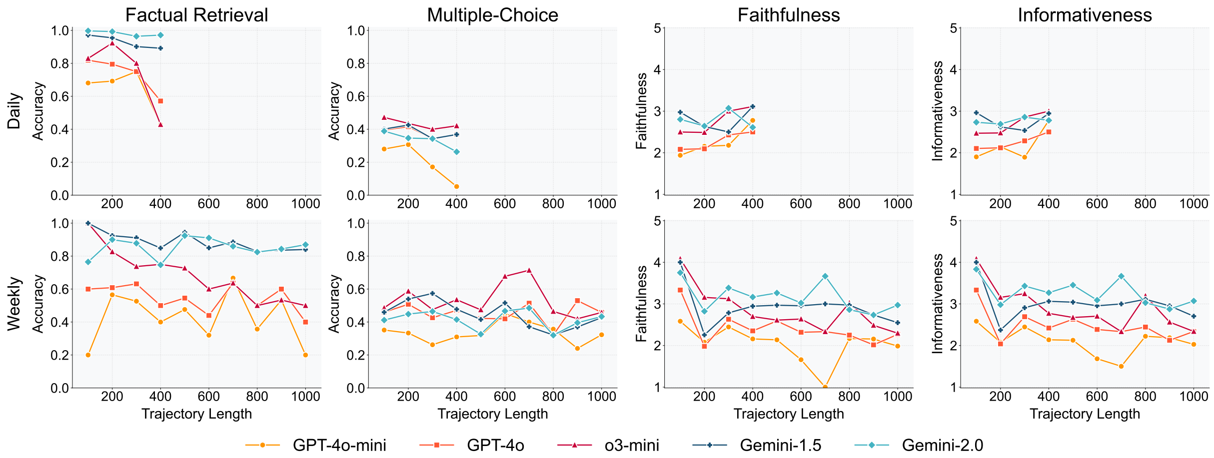


Figure 4: Performance comparisons of closed LLMs across varying trajectory lengths

For factual retrieval questions, GPT models exhibit a sharp accuracy drop as trajectory length increases, likely because longer trajectories make it harder to focus on relevant information. In contrast, Gemini models show more gradual decline or stable performance across different lengths.

For multiple-choice questions, trajectory length has less impact. Most models maintain relatively stable accuracy regardless of input length. Interestingly, some models achieve slightly better accuracy on longer weekly trajectories, possibly due to richer temporal context. These results suggest that optimal context length for multiple-choice reasoning varies by model. We also observed consistent performance degradation with increasing trajectory length for open models (see Appendix D.1).

5 Related Work

Spatiotemporal Reasoning Mobility question answering requires sophisticated spatiotemporal reasoning. Temporal reasoning research has centered on event ordering (Ning et al., 2020; Naik et al., 2019), frequency or duration estimation (Rajpurkar et al., 2016; Zhou et al., 2019; Virgo et al., 2022), and temporal causality (Mirza, 2014; Zhou et al., 2021). Even recent LLMs struggle with such tasks (Gupta et al., 2023; Yang et al., 2023; Jia et al., 2024; Chu et al., 2024; Wang and Zhao, 2024; Xiong et al., 2024; Fatemi et al., 2025).

Spatial reasoning is equally demanding. Vision-language models perform well when spatial cues are explicit in images (Johnson et al., 2017; Mirzaee et al., 2021; Mirzaee and Kordjamshidi, 2022; Kamath et al., 2023; Cheng et al., 2024a,b; Shiri et al., 2024; Lei et al., 2025) or video (Jang et al., 2017; Lei et al., 2020; Grunde-McLaughlin et al., 2021), yet still struggle with nuanced relations (Cheng et al., 2024c; Fu et al., 2025).

MobQA requires LLMs to perform spatiotemporal reasoning from numerical inputs alone,which is known to be challenging in embodied reasoning tasks (Majumdar et al., 2024).

LLMs for Mobility Data Mobility data have been analyzed with deep learning models (Liu et al., 2016; Feng et al., 2020; Liang et al., 2022). Applying LLMs to mobility data is a growing frontier. Recent studies tackle trajectory prediction (Wang et al., 2023a; Liang et al., 2024; Xu et al., 2024; Li et al., 2024) and synthetic data generation (Asano et al., 2024; Wang et al., 2024), often leveraging visual or geospatial signals (Feng et al., 2025a,b). Yet most work remains confined to closed-ended tasks, short sequences (often under 100 events) (Kim et al., 2024; Yoon et al., 2024), or enriched inputs such as check-in POIs (Wang et al., 2023a). Handling long streams of raw coordinates that lack explicit semantics continues to challenge current LLMs (Fang et al., 2024b; Yoon et al., 2024). MobQA uniquely targets the semantic understanding of human mobility through natural language.

6 Conclusion

We introduced MobQA, the first benchmark for evaluating LLMs' semantic understanding of human mobility through question answering. Our experiments reveal capabilities and limitations of state-of-the-art LLMs across factual retrieval, multiple-choice reasoning, and free-form explanation tasks. We believe MobQA will advance research in LLM-based mobility analysis.

Limitations

Dataset Scope The MobQA dataset is built on the Geolife dataset,[6] which provides rich urban trajectory data collected in Beijing. Due to privacy concerns associated with collecting and releasing human mobility data, we chose to use this publicly available dataset that was explicitly consented for research use. While the Geolife dataset offers a valuable starting point, the geographic and demographic scope remains limited. Moreover, our dataset is currently constructed in Japanese. Nevertheless, recent findings suggest that human mobility exhibits common, transferable features across cities (Schläpfer et al., 2021), and that cross-city transfer learning is feasible for mobility prediction tasks (Jin et al., 2022). This indicates that our benchmark, despite being based on a single city, represents an important step toward more diverse benchmarks. Furthermore, our data construction framework is inherently language-agnostic and can be extended to other languages and cultural contexts seamlessly without major changes in the annotation and question-answer generation processes. Future work will seek to build multilingual versions of the dataset with more diversity in geography and demography. This will support broader global applicability and enable future studies on how such factors influence mobility understanding.

Evaluation Scope This study focuses on three core tasks: factual retrieval, multiple-choice reasoning, and free-form explanation, as a starting point for semantic mobility understanding. While these cover essential capabilities, they do not exhaust the full space of mobility-related reasoning. Future extensions could incorporate dialogue-based interactions, long-term behavior summarization, or subjective tasks such as intent or preference estimation, enabling richer assessment of model understanding.

Ethical Considerations

Human mobility data, while valuable for research, carries inherent potential risks, particularly concerning individual privacy. Misuse of such data could lead to significant privacy infringements. To address these concerns, our study exclusively utilizes *anonymized trajectories* as introduced in Sec. 3.1. This anonymization is a critical measure to ensure that individual identities are protected and that the data cannot be traced back to specific persons, thereby resolving the primary privacy issues associated with using detailed movement data in our experiments.

[6] https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

Basma H Albanna, Ibrahim F Moawad, Sherin M Moussa, and Mahmoud A Sakr. 2015. Semantic trajectories: A survey from modeling to application. In Vasily Popovich, Christophe Claramunt, Manfred Schrenk, Kyrill Korolenko, and Jérôme Gensel, editors, *Information Fusion and Geographic Information Systems (IF&GIS' 2015): Deep Virtualization for Mobile GIS*, pages 59–76. Springer International Publishing.

Hikaru Asano, Ryo Yonetani, Taiki Sekii, and Hiroki Ouchi. 2024. Text2Traj2Text: Learning-by-synthesis framework for contextual captioning of human movement trajectories. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 289–302. Association for Computational Linguistics.

Reza Askarizad and Hossein Safari. 2020. The influence of social interactions on the behavioral patterns of the people in urban spaces (case study: The pedestrian zone of rasht municipality square, iran). *Cities*, 101:102687.

Hamada S Badr, Hongru Du, Maximilian Marshall, Ensheng Dong, Marietta M Squire, and Lauren M Gardner. 2020. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *The Lancet Infectious Diseases*, 20(11):1247–1254.

Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. 2022. A survey on machine reading comprehension systems. *Natural Language Engineering*, 28(6):683–732.

Sonia Ben Mokhtar, A Boutet, L Bouzouina, P Bonnel, Olivier Brette, L Brunie, M Cunche, Stephane D 'Alu, Vincent Primault, Patrice Raveneau, H Rivano, and Razvan Stanica. 2017. PRIVA'MOV: Analysing human mobility through multi-sensor datasets.

Vincent D Blondel, Markus Esch, Connie Chan, Fabrice Clerot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. 2012. Data for development: the D4D challenge on mobile phone data. *arXiv [cs.CY]*.

Francesco Calabrese, Laura Ferrari, and Vincent D Blondel. 2015. Urban sensing using mobile phone network data: A survey of research. *ACM Computing Surveys (CSUR)*, 47(2):1–20.

Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. 2021. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024a. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024b. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 37, pages 135062–135093. Curran Associates, Inc.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024c. SpatialRGPT: Grounded spatial reasoning in vision-language models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Eunjoon Cho, Seth A Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*. ACM.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1204–1228. Association for Computational Linguistics.

Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. 2024. Neural retrievers are biased towards LLM-generated content. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*, page 526–537. Association for Computing Machinery.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv [cs.CL]*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nathan Eagle and Alex (Sandy) Pentland. 2006. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268.

Xi Fang, Weijie Xu, Fiona Anting Tan, Ziqing Hu, Jiani Zhang, Yanjun Qi, Srinivasan H Sengamedu, and Christos Faloutsos. 2024a. Large language models (LLMs) on tabular data: Prediction, generation, and understanding - a survey. *Transactions on Machine Learning Research*.

Xi Fang, Weijie Xu, Fiona Anting Tan, Ziqing Hu, Jiani Zhang, Yanjun Qi, Srinivasan H Sengamedu, and Christos Faloutsos. 2024b. Large language models (LLMs) on tabular data: Prediction, generation, and understanding - a survey. *Transactions on Machine Learning Research*.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2025. Test of time: A benchmark for evaluating LLMs on temporal reasoning. In *Proceedings of the International Conference on Learning and Representation (ICLR)*.

Jie Feng, Tianhui Liu, Yuwei Du, Siqi Guo, Yuming Lin, and Yong Li. 2025a. CityGPT: Empowering urban spatial cognition of large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pages 591–602. ACM.

Jie Feng, Zeyu Yang, Fengli Xu, Haisu Yu, Mudan Wang, and Yong Li. 2020. Learning to simulate human mobility. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*, page 3426–3433. Association for Computing Machinery.

Jie Feng, Jun Zhang, Tianhui Liu, Xin Zhang, Tianjian Ouyang, Junbo Yan, Yuwei Du, Siqi Guo, and Yong Li. 2025b. CityBench: Evaluating the capabilities of large language models for urban tasks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pages 5413–5424. ACM.

Jie Feng, Mingyang Zhang, Huandong Wang, Zeyu Yang, Chao Zhang, Yong Li, and Depeng Jin. 2019. DPLink: User identity linkage via deep neural network from heterogeneous mobility data. In *The World Wide Web Conference*, pages 459–469. ACM.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2025. BLINK: Multimodal large language models can see but not perceive. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 148–166. Springer Nature Switzerland.

Letian Gong, Yan Lin, Xinyue Zhang, Yiwen Lu, Xuedi Han, Yichen Liu, Shengnan Guo, Youfang Lin, and Huaiyu Wan. 2024. Mobility-LLM: Learning visiting intentions and travel preference from human mobility data with large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2021. AGQA: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11287–11297.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on LLM-as-a-judge. *arXiv [cs.CL]*.

Vivek Gupta, Pranshu Kandoi, Mahek Vora, Shuo Zhang, Yujie He, Ridho Reinanda, and Vivek Srikumar. 2023. TempTabQA: Temporal question answering for semi-structured tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2431–2453. Association for Computational Linguistics.

Junqing He, Kunhao Pan, Xiaoqun Dong, Zhuoyang Song, LiuYiBo LiuYiBo, Qianguosun Qianguosun, Yuxin Liang, Hao Wang, Enming Zhang, and Jiaxing Zhang. 2024. Never lost in the middle: Mastering long-context question answering with position-agnostic decompositional training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13628–13642. Association for Computational Linguistics.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhen Jia, Philipp Christmann, and Gerhard Weikum. 2024. TIQ: A benchmark for temporal question answering with implicit time constraints. In *Companion Proceedings of the ACM Web Conference 2024*, page 1394–1399. Association for Computing Machinery.

Yilun Jin, Kai Chen, and Qiang Yang. 2022. Selective cross-city transfer learning for traffic prediction via source city region re-weighting. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 731–741, New York, NY, USA. Association for Computing Machinery.

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

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175. Association for Computational Linguistics.

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-LLM: Large language models for health prediction via wearable sensor data. In *Proceedings of the fifth Conference on Health, Inference, and Learning*, volume 248, pages 522–539. PMLR.

Ani Kristo. 2021. NYC Yellow Taxi Trips Dataset.

J K Laurila, Daniel Gatica-Perez, I Aad, Blom J., Olivier Bornet, Trinh-Minh-Tri Do, O Dousse, J Eberle, and M Miettinen. 2012. The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing*, Newcastle.

Taewhoo Lee, Chanwoong Yoon, Kyochul Jang, Donghyeon Lee, Minju Song, Hyunjae Kim, and Jaewoo Kang. 2025. ETHIC: Evaluating large language models on long-context tasks with high information coverage. 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 5497–5512. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225. Association for Computational Linguistics.

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. 2025. Scaffolding coordinates to promote vision-language coordination in large multimodal models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2886–2903. Association for Computational Linguistics.

Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. 2024. UrbanGPT: Spatio-temporal large language models. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*, page 5351–5362. Association for Computing Machinery.

Yuebing Liang, Yichao Liu, Xiaohan Wang, and Zhan Zhao. 2024. Exploring large language models for human mobility prediction under public events. *Computers, environment and urban systems*, 112(102153):102153.

Yuxuan Liang, Kun Ouyang, Yiwei Wang, Xu Liu, Hongyang Chen, Junbo Zhang, Yu Zheng, and Roger Zimmermann. 2022. TrajFormer: Efficient trajectory classification with transformers. In *Proceedings of* the 31st ACM International Conference on Information & Knowledge Management*, page 1229–1237. Association for Computing Machinery.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2016. Predicting the next location: a recurrent model with spatial and temporal contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, page 194–200. AAAI Press.

Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. 2021. A survey on deep learning for human mobility. *ACM Comput. Surv.*, 55(1):1–44.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, and 5 others. 2024. OpenEQA: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16488–16498.

Meta. 2024. The llama 3 herd of models. *arXiv [cs.AI]*.

Paramita Mirza. 2014. Extracting temporal and causal relations between events. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 10–17. Association for Computational Linguistics.

Roshanak Mirzaee and Parisa Kordjamshidi. 2022. Transfer learning with synthetic corpora for spatial role labeling and reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6148–6165. Association for Computational Linguistics.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598. Association for Computational Linguistics.

Mohamed Mokbel, Mahmoud Sakr, Li Xiong, Andreas Züfle, Jussara Almeida, Taylor Anderson, Walid Aref, Gennady Andrienko, Natalia Andrienko, Yang Cao, Sanjay Chawla, Reynold Cheng, Panos Chrysanthis, Xiqi Fei, Gabriel Ghinita, Anita Graser, Dimitrios Gunopulos, Christian S Jensen, Joon-Seok Kim, and 29 others. 2024. Mobility data science: Perspectives and challenges. *ACM Trans. Spatial Algorithms Syst.*, 10(2).

Aakanksha Naik, Luke Breitfeller, and Carolyn Rose. 2019. TDDiscourse: A dataset for discourse-level temporal ordering of events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 239–249. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in neural information processing systems*, volume 35, page 27730–27744. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 311–318. Association for Computational Linguistics.

Luca Pappalardo, Ed Manley, Vedran Sekara, and Laura Alessandretti. 2023. Future directions in human mobility science. *Nature Computational Science*, 3(7):588–600.

Christine Parent, Stefano Spaccapietra, Chiara Renso, Gennady Andrienko, Natalia Andrienko, Vania Bogorny, Maria Luisa Damiani, Aris Gkoulalas-Divanis, Jose Macedo, Nikos Pelekis, Yannis Theodoridis, and Zhixian Yan. 2013. Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)*, 45(4):1–32.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for NLG systems. *ACM Comput. Surv.*, 55(2):1–39.

Markus Schläpfer, Lei Dong, Kevin O'Keeffe, Paolo Santi, Michael Szell, Hadrien Salat, Samuel Anklesaria, Mohammad Vazifeh, Carlo Ratti, and Geoffrey B West. 2021. The universal visitation law of human mobility. *Nature*, 593(7860):522–527.

Shuo Shang, Bo Yuan, Ke Deng, Kexin Xie, and Xiaofang Zhou. 2011. Finding the most accessible locations: reverse path nearest neighbor query in road networks. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 181–190. Association for Computing Machinery.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. 2024. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21440–21455. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. *arXiv [cs.CL]*.

Felix Virgo, Fei Cheng, and Sadao Kurohashi. 2022. Improving event duration question answering by leveraging existing temporal information extraction data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4451–4457. European Language Resources Association.

Jiawei Wang, Renhe Jiang, Chuang Yang, Zengqing Wu, Makoto Onizuka, Ryosuke Shibasaki, Noboru Koshizuka, and Chuan Xiao. 2024. Large language models as urban residents: An LLM agent framework for personal mobility generation. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Xinglei Wang, Meng Fang, Zichao Zeng, and Tao Cheng. 2023a. Where would I go next? large language models as human mobility predictors.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508. Association for Computational Linguistics.

Yuqing Wang and Yun Zhao. 2024. TRAM: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6389–6415. Association for Computational Linguistics.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10452–10470. Association for Computational Linguistics.

Xiaohang Xu, Renhe Jiang, Chuang Yang, Zipei Fan, and Kaoru Sezaki. 2024. Taming the long tail in human mobility prediction. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 37, pages 54748–54771. Curran Associates, Inc.

Hao Xue and Flora D Salim. 2023. Human mobility question answering (vision paper). *arXiv [cs.CL]*.

Takahiro Yabe, Kota Tsubouchi, Toru Shimizu, Yoshihide Sekimoto, Kaoru Sezaki, Esteban Moro, and Alex Pentland. 2024. YJMob100K: City-scale and longitudinal dataset of anonymized human mobility trajectories. *Scientific Data*, 11(1):397.

Dingqi Yang, Daqing Zhang, and Bingqing Qu. 2016. Participatory cultural mapping based on collective behavior data in location-based social networks. *ACM transactions on intelligent systems and technology*, 7(3):1–23.

Dingqi Yang, Daqing Zhang, Vincent W Zheng, and Zhiyong Yu. 2015. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE transactions on systems, man, and cybernetics. Systems*, 45(1):129–142.

Sen Yang, Xin Li, Lidong Bing, and Wai Lam. 2023. Once upon a *time* in *graph*: Relative-time pretraining for complex temporal reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11879–11895. Association for Computational Linguistics.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2025. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*.

Hyungjun Yoon, Biniyam Aschalew Tolera, Taesik Gong, Kimin Lee, and Sung-Ju Lee. 2024. By my eyes: Grounding multimodal large language models with sensor data via visual prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2219–2241. Association for Computational Linguistics.

Jing Yuan, Yu Zheng, Chengyang Zhang, Wenlei Xie, Xing Xie, Guangzhong Sun, and Yan Huang. 2010. T-drive: driving directions based on taxi trajectories. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 99–108. Association for Computing Machinery.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning and Representation (ICLR)*.

Zijian Zhang, Yujie Sun, Zepu Wang, Yuqi Nie, Xiaobo Ma, Peng Sun, and Ruolin Li. 2024. Large language models for mobility in transportation systems: A survey on forecasting tasks. *arXiv [cs.LG]*.

Kai Zhao, Sasu Tarkoma, Siyuan Liu, and Huy Vo. 2016. Urban human mobility data mining: An overview. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1911–1920. IEEE.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Yu Zheng, Hao Fu, Xing Xie, Wei-Ying Ma, and Quannan Li. 2011. *Geolife GPS trajectory dataset - User Guide*, geolife gps trajectories 1.1 edition.

Yu Zheng, Longhao Wang, Ruochi Zhang, Xing Xie, and Wei-Ying Ma. 2008. GeoLife: Managing and understanding your past life over maps. In *The Ninth International Conference on Mobile Data Management (mdm 2008)*, pages 211–212.

Yu Zheng, Xing Xie, Wei-Ying Ma, and Others. 2010. GeoLife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39.

Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022. Video question answering: Datasets, algorithms and challenges. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6439–6455. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. “going on a vacation” takes longer than “going for a walk”: A study of temporal commonsense understanding. 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 3363–3369. Association for Computational Linguistics.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371. Association for Computational Linguistics.

Zihao Zhou and Rose Yu. 2025. Can LLMs understand time series anomalies? In *Proceedings of the International Conference on Learning and Representation (ICLR)*.

A Reproducibility Measures

The MobQA dataset is composed of anonymized trajectories, questions, and reference answers, which are derived from the Geolife GPS traces. The Geolife dataset is distributed under the Microsoft Research License Agreement (MSR-LA). We employ the dataset purely for academic and benchmarking purposes, strictly adhering to its terms for non-commercial research use. Crucially, following direct communication with the original Geolife authors, we have obtained explicit written permission to release the generated question-answer pairs, semantic trajectory annotations, and corresponding user IDs. This authorization ensures full compliance with the original dataset's licensing terms while the original GPS traces remain protected under the license agreement.

To ensure full reproducibility under MSR-LA, we adopt an approach where each researcher can individually download and obtain the Geolife license, then reconstruct the complete MobQA dataset using our released components. Specifically, we open-source: (i) the generated question-answer pairs with corresponding user IDs and trajectory dates, (ii) the semantic trajectory annotations, (iii) reconstruction code that generates the complete MobQA dataset from the downloaded Geolife dataset and our released QA pairs and trajectory information, and (iv) all evaluation pipelines and reproduction scripts, including answer generation and evaluation code (prompts for each task, scoring scripts) required to reproduce our experiments. Note however that differences in LLM API versions may introduce an unavoidable ceiling on exact replicability.

B Justification for Base Dataset Selection

Our goal is to enable semantic question answering over human mobility trajectories. As outlined in Sec. 3, we require datasets that satisfy three essential criteria: (R1) dense sampling over long horizons to capture fine-grained behaviors and routine patterns, (R2) multi-mode mobility data beyond single vehicle types, and (R3) public availability for reproducibility.

We reviewed existing surveys covering publicly available human mobility datasets (Calabrese et al., 2015; Zhao et al., 2016; Luca et al., 2021). These datasets can be broadly categorized as follows, though most are unsuitable for our goal as they fail to simultaneously satisfy all three requirements.

- **Single-Mode Datasets** (*e.g.*, NYC Yellow Taxi (Kristo, 2021), T-Drive (Yuan et al., 2010)) are limited to a single mode of transport, primarily taxis. This makes them unsuitable for analyzing multi-mode mobility patterns or personal routine behaviors, violating R2. Moreover, these datasets record only pickup and dropoff locations rather than continuous trajectories, or capture only short sequences that lack the extended temporal coverage needed to understand routine behaviors and long-term patterns, thus also violating R1.

- **Sparse Check-in Logs** (*e.g.*, Foursquare (Yang et al., 2015, 2016), Gowalla (Cho et al., 2011)) consist of check-in logs recorded only upon arrival at specific locations. This creates extremely sparse data points. While useful for tasks like point-of-interest recommendation, they lack the continuous path information necessary for understanding the *how* and *why* of movement between locations, violating R1. Moreover, the absence of intermediate trajectory data makes it difficult to infer the transportation modes used between check-in points, also violating R2.
- **Network Infrastructure Traces** (*e.g.*, ISP-based logs (Feng et al., 2019), Orange D4D (Blondel et al., 2012), MIT Reality Mining (Eagle and , Sandy)) rely on connections to fixed network nodes such as cellular towers or Wi-Fi access points. Their spatial accuracy is inherently limited by infrastructure coverage, and temporal sampling is often sparse or event-driven (*e.g.*, triggered only during calls). This lack of fine-grained spatiotemporal resolution violates R1. Moreover, inferring transportation modes from coarse cell-tower transitions is unreliable, also violating R2.
- **Coarse-Grained Trajectory** encompasses datasets where trajectories are spatially aggregated into fixed grid cells or regions, primarily to preserve privacy or reduce data volume. Examples include YJMob100K (Yabe et al., 2024), which discretizes continuous GPS traces into large mesh codes (*e.g.*, 500m × 500m) and long time intervals. While effective for macroscopic flow analysis, this artificial coarsening removes the precise geometric details required to reconstruct complex micro-mobility behaviors, violating R1.
- **Restricted Mobility Data** (*e.g.*, Priva'Mov (Ben Mokhtar et al., 2017), Nokia MDC (Laurila et al., 2012)) represent high-quality datasets that often contain dense, multi-transportation mode records satisfying both R1 and R2. However, access to these datasets is typically limited, such as by strict non-disclosure agreements (NDAs). This lack of public accessibility (violating R3) hinders reproducibility and benchmarking for the wider community.

Indeed, our preliminary experiments with sparse data, including Foursquare check-ins and YJ-Mob100K's coarse-grained GPS, confirmed these limitations. We found that from these abstract trajectories, even human experts were unable to reliably infer the semantic nuances required for our QA task. While such datasets may be sufficient for trajectory prediction or generation, they cannot support the deep, context-aware understanding that our natural language-based benchmark is designed to evaluate.

This systematic elimination identified the *Geolife GPS dataset* (Zheng et al., 2008, 2010, 2011) as the only publicly available option that meets our stringent requirements. Geolife is the ideal foundation for our MobQA dataset for the following reasons:

- It provides high spatiotemporal resolution, enabling detailed trajectory analysis.
- It includes verified labels for 11 transport modes (*e.g.*, walking, biking, subway), covers diverse behaviors (*e.g.*, commutes, leisure), and spans over five years for 182 users.
- It is openly accessible and comprehensively documented, ensuring full replicability.

C Evaluation Methodology Details

C.1 Evaluation Metric for Free-Form Tasks

To select an appropriate evaluation metric for free-form questions, we compared conventional automatic metrics with the recent LLM-as-a-judge framework. Specifically, we benchmarked following widely used metrics against human judgments:

- **BLEU** (Papineni et al., 2002): A precision-oriented metric that measures n-gram overlap between generated and reference texts, widely adopted in machine translation evaluation.
- **ROUGE** (Lin, 2004): A recall-focused metric that computes n-gram overlap (*e.g.*, ROUGE-2) or the longest common subsequence (ROUGE-L), commonly used for summarization tasks.
- **BERTScore** (Zhang* et al., 2020): A semantic-similarity metric that leverages contextualized BERT embeddings to capture meaning beyond surface-level token matches.

Tab. 5 presents the results from benchmarking these metrics on both daily and weekly subsets of our

Table 5: Comparing Automatic Metrics and LLM-as-a-Judge: Correlation Coefficients for Faithfulness and Informativeness

Eval Method	Faithfulness		Informativeness	
	One Day	Weekly	One Day	Weekly
Automatic Metrics				
BLEU	0.081 / 0.206	0.185 / 0.230	0.076 / 0.175	0.113 / 0.154
ROUGE-2	-0.191 / -0.093	-0.024 / 0.001	-0.272 / -0.206	-0.017 / 0.010
ROUGE-L	-0.003 / -0.006	0.110 / 0.136	-0.122 / -0.151	0.167 / 0.183
BERTScore	0.239 / 0.252	0.086 / 0.114	0.279 / 0.294	0.136 / 0.151
LLM-as-a-Judge				
GPT-4o-mini	0.502 / **0.562**	0.278 / 0.349	0.554 / 0.605	0.402 / 0.425
GPT-4o	**0.507** / 0.524	0.401 / 0.492	0.521 / 0.511	0.513 / 0.537
Gemini 1.5	0.370 / 0.378	**0.513** / **0.548**	0.544 / 0.593	0.502 / 0.521
Gemini 2.0	0.368 / 0.401	0.494 / 0.535	**0.584** / **0.608**	**0.577** / **0.596**

data. Our findings indicate that traditional automatic metrics correlate poorly with human judgments. ROUGE-2 and ROUGE-L show particularly weak or even negative correlations, while BLEU and BERTScore achieve modest positive correlations that are nevertheless substantially lower than those of LLM-based evaluators.

In contrast, LLM-as-a-judge evaluators show substantially stronger agreement with human assessments. GPT-4o-mini achieves the highest faithfulness correlation on daily data, while Gemini models excel on weekly data. Gemini 2.0 yields the highest informativeness correlations, exceeding 0.57 for both temporal scales. Overall, LLM-based judges achieve correlations in the 0.40-0.60 range, substantially outperforming automatic metrics. Given the inadequacy of conventional metrics in this context, we adopt LLM-as-a-judge assessments for faithfulness and informativeness throughout our study. This approach aligns with recent best practices for evaluating open-ended generation tasks (Zheng et al., 2023; Gu et al., 2024).

C.2 Self-Enhancement Bias in LLM-as-a-Judge

Recent studies have highlighted self-enhancement bias in LLM-as-a-judge evaluations, where evaluators may favor responses from their own model family (Zheng et al., 2023; Dai et al., 2024; Ye et al., 2025). To mitigate this bias and ensure the validity of our results,, we conducted a comparative analysis using both our primary evaluator (Gemini-2.0-Flash) and GPT-4o as a secondary evaluator from a different model family.

Our analysis revealed a consistency in the assessment patterns of both evaluators. As shown in Tab. 6, the performance ranking remained stable for both faithfulness and informativeness metrics,

Table 6: Performance of LLMs on Free-form Questions (Faithfulness & Informativeness). Results are shown as Gemini-2.0-Flash/GPT-4o evaluation scores. **bold + Underline** indicates the best score, **Bold** indicates the second best, and Underline indicates the third best.

Model	Faithfulness		Informativeness	
	One Day	Weekly	One Day	Weekly
GPT-4o-mini	2.01/2.07	1.98/1.92	1.97/1.82	2.01/1.83
GPT-4o	2.15/2.16	2.27/2.25	2.15/2.07	2.35/2.23
o3-mini	2.46/2.42	2.74/**2.70**	2.43/2.08	2.81/2.46
Gemini 1.5	**3.12**/**2.94**	**2.81**/2.58	**3.08**/**2.42**	**2.91**/**2.38**
Gemini 2.0	**2.96**/**2.79**	**3.09**/**2.69**	**2.91**/**2.30**	**3.22**/**2.48**

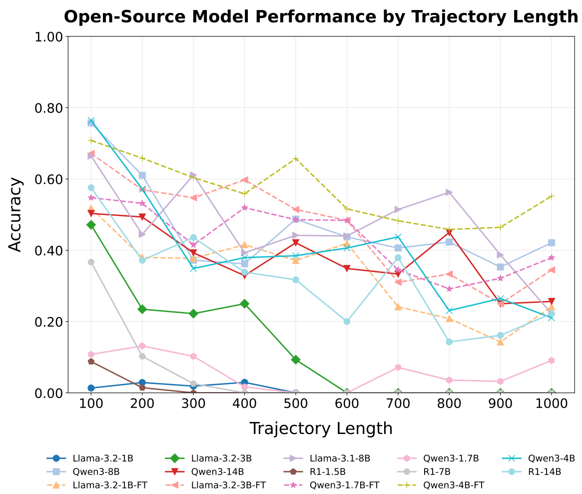


Figure 5: Performance comparisons of open-source LLMs and their fine-tuned versions across varying trajectory lengths

regardless of whether Gemini-2.0-Flash or GPT-4o served as judge. Notably, the Gemini family consistently received higher scores than other models across the board.

To statistically validate this inter-evaluator agreement, we analyzed the correlation between their scores. The results confirmed a very strong positive correlation, with a Pearson correlation coefficient of $r = 0.837$ and a Spearman's rank correlation of $\rho = 0.845$. This high correlation provides compelling evidence that our evaluation framework is robust and that the observed performance differences reflect genuine model capabilities rather than artifacts of self-enhancement bias.

D Additional Analyses

This section provides additional experimental analyses to further investigate the performance and limitations of existing LLMs. Specifically, we examine

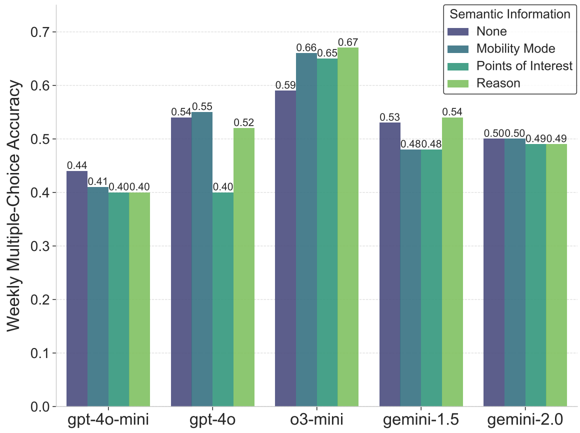


Figure 6: Effect of semantic information for multiple-choice question answering performances

(i) the impact of trajectory length on open-source models, (ii) the effect of incorporating semantic information, and (iii) the full dataset results, offering deeper insights into model robustness and performance variation.

D.1 Impact of Trajectory Lengths on Open Models

We performed a trajectory length analysis on the open models (Fig. 5), which presents the combined results for factual retrieval and multiple-choice tasks. Similar to the trends observed with the closed models, all ten open models showed a consistent degradation in performance as the trajectory length increased. This negative impact was more pronounced in smaller models. Although fine-tuning consistently improved performance over the base models, it failed to resolve the issue of performance degradation with longer sequences. This highlights the inherent difficulty of the underlying problem, which remains a challenge even for specialized models.

D.2 Effect of Semantic Information

Mobility data in real-world applications can sometimes include semantic information such as transportation modes, POIs, or reasons for visits, which may be beneficial for mobility question answering. To investigate this, we evaluated multiple-choice accuracy for weekly trajectory data with four input configurations: raw trajectories (None) and trajectories augmented with mobility modes, POIs, or inferred reasons of visit.

As reported in Fig. 6, o3-mini demonstrated substantial improvement with semantic augmentation, achieving its highest accuracy (0.67) when reasons of visit were included. In contrast, gpt-4o-mini performed worse with all types of semantic inputs, potentially due to longer sequences making it difficult for the model to focus on relevant information. Gemini-1.5 and Gemini-2.0 remained relatively stable with minimal gains or losses. While this stability suggests robustness, it also indicates that these models may not fully exploit the added semantic cues under current settings. These results indicate that semantic cues benefit performance only when LLMs can integrate them effectively. Without proper handling, such augmentations may even degrade performance, particularly in smaller models.

D.3 Full Dataset Result

Tab. 7 shows the evaluation results of closed-source LLMs for all the test data including longer trajectories that exceeds our GPU memory limit for open-source models.

Table 7: Performance comparisons of closed-source LLMs

Model	Factual Retrieval		Multiple-Choice		Faithfulness		Informativeness	
	Daily	Weekly	Daily	Weekly	Daily	Weekly	Daily	Weekly
GPT-4o-mini	0.690	0.427	0.294	0.352	2.01	1.98	1.97	2.01
GPT-4o	0.812	0.580	0.413	**0.476**	2.15	2.27	2.15	2.35
o3-mini	0.824	0.644	0.472	0.473	2.46	2.74	2.43	2.81
Gemini 1.5	0.963	**0.885**	**0.490**	0.473	**3.12**	2.81	**3.08**	2.91
Gemini 2.0	**0.991**	0.874	0.403	0.439	2.96	**3.09**	2.91	**3.22**

E Semantic Annotation Details

This section provides detailed information about the semantic annotation process described in Sec. 3.2. Three expert annotators specialized in mobility-data annotation processed a total of 136 trajectories (17 users $\times$ 8 consecutive days), spending approximately 60 hours in total (excluding training and instruction), averaging around 26 minutes per day-trajectory.

E.1 Annotation Tool

The custom annotation tool (Fig. 7) was developed based on Streamlit[7], providing an interactive environment to facilitate precise semantic labeling. The interface consists of two main components: a contextual view of eight consecutive days displayed on the left pane, and a detailed visualization of the selected day's trajectory on the right pane. The interactive trajectory visualization was built using MovingPandas[8], allowing annotators to intuitively inter-

[7] https://streamlit.io/
[8] https://movingpandas.org/

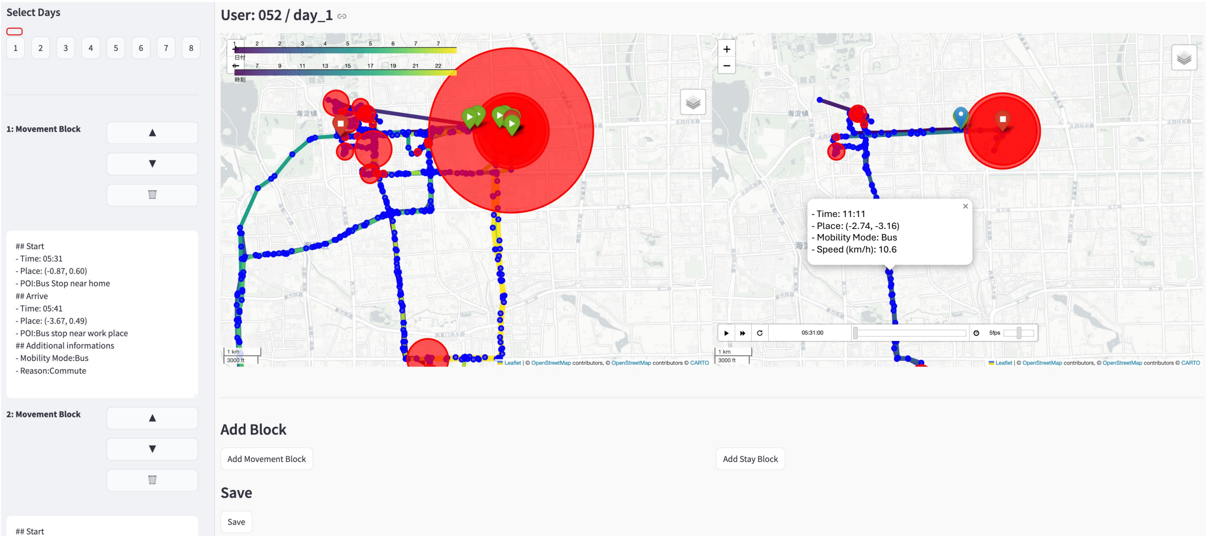


Figure 7: Screenshot of the Streamlit-based semantic trajectory annotation interface. The leftmost pane lets annotators switch among eight consecutive days, providing constant week-level context. In the main map view, automatically detected stay points appear as red circles whose radii reflect duration, whereas individual GPS fixes are shown as blue dots; inter-dot spacing intuitively conveys movement speed. Hover tooltips reveal timestamp, coordinates, speed, and user-reported mobility mode, and playback controls enable frame-by-frame inspection of the trajectory. Below the map, editable movement and stay blocks can be inserted, reordered, or deleted, streamlining fine-grained semantic refinement.

pret movement characteristics. Specifically, rule-based automatically identified stationary points are depicted by red circles, whose sizes indicate the duration of stays.[9] Blue dots represent individual GPS fixes, with inter-dot spacing visually encoding travel speed—dense clusters signify slower movements or stops, whereas greater spacing indicates faster travel segments. The interface also allows annotators to zoom in and out of areas of interest, enabling detailed examination of specific locations and movements while maintaining the broader context of the trajectory. Additionally, annotators can access precise information, such as timestamps, geographic coordinates, speed, and user-reported mobility modes through hover tooltips. Playback functionalities, such as frame-by-frame navigation, further support accurate segmentation.

E.2 Annotator Details

The annotation team consisted of three professional annotators (two female, one male) employed within the authors' organization. All annotators were thoroughly briefed on the purpose of the study, including how the annotated data would be used and its intended public release. They provided informed consent for their annotations to be included in the publicly available dataset. While we determined that the annotation process posed minimal risks to the annotators (primarily limited to standard ergonomic considerations associated with computer work), we ensured transparent communication about all aspects of the project. The annotators acknowledged their understanding of both the project goals and the public nature of their contributions to the mobility dataset.

E.3 Annotation Process

The annotation protocol involved two structured training phases. Initially, annotators participated in a one-hour session, including detailed instructions and hands-on practice guided individually by the lead author using pre-annotated exemplar trajectories. Feedback on tool usability and instruction clarity was collected after this phase. Based on this input, guidelines were clarified, and the annotation tool interface was refined accordingly. Subsequently, annotators underwent a second one-hour instructional session before proceeding with the main annotation tasks.

Quality assurance was conducted rigorously, with the lead author performing a final validation of all semantic trajectories. This review ensured consistency in mobility modes, segment boundaries aligned precisely to minute-level resolution,

[9] These stay durations were carefully redefined by annotators during the annotation process to ensure accuracy.

and completeness by identifying and correcting potential omissions such as brief transitions between mobility modes. This meticulous verification step guaranteed high-fidelity semantic trajectories suitable for reliable downstream applications in mobility-related question answering.

F Manual Answer Filtering

To ensure the quality of ground-truth answers to free-form questions, we employ a manual filtering process, which verifies whether the generated answers are faithful to the original trajectory data and contain information beneficial to the user.

F.1 Annotators

Annotators for the answer filtering task were recruited through a crowdsourcing platform named Lancers.[10] Initially, we hired nine annotators via an open call and conducted a preliminary annotation task. We then selected the three annotators who demonstrated exceptionally high annotation quality to undertake the main annotation work in this filtering phase. Every annotator was explained on the study purpose and data usage, acknowledged risks, and consented to their annotation being included in the public dataset.

The compensation for the annotators was a total of 36,652 JPY for the preliminary task involving the initial nine annotators. For the main phase, each of the three selected annotators was paid 18,568 JPY, resulting in a total cost of 92,356 JPY. All annotators were informed beforehand that their annotations might be publicly released as part of the dataset, and all provided their consent.

F.2 Annotation Tool

We extended the annotation tool developed in Sec. E.1 to visualize additional information necessary for the answer filtering process, such as stop points, POIs, mobility modes, as well as questions and generated answers. This customized interface enabled annotators to efficiently assess answer quality by simultaneously referencing the original trajectory data along with its semantic context.

F.3 Evaluation Process and Metrics

The quality of each answer for the free form questions was evaluated based on faithfulness and informativeness scores defined in Sec. 2.4. Prior to the annotation task, annotators were informed the evaluation protocol with a document that explained the use of the annotation tool, detailed the evaluation metrics, and provided three illustrative examples for different scores of faithfulness and three for informativeness to clarify the scoring criteria. Furthermore, annotators were encouraged to maintain open communication to address any questions or ambiguities they encountered during the process.

Each response was rated on a 1-5 scale for both faithfulness and informativeness. To be finally included in the dataset, an answer had to achieve a score of 4 or higher on both metrics. Annotators first underwent a calibration phase using gold-standard examples scored by experts to ensure a unified understanding of the evaluation criteria. The actual evaluation was conducted using the aforementioned custom interface, where annotators could simultaneously view the raw trajectory data, the semantic trajectory, the question, and the model-generated answer while scoring. As a result of this filtering process, approximately 90% (around 1,800) of the initially generated 2,000 answers met the acceptance criteria. This high acceptance rate is likely attributed to the benefit of providing semantic information alongside raw trajectory data to the LLM (GPT-4o) during answer generation.

G Experiment Details

G.1 Model Details

We benchmarked five recent closed-source large language models (LLMs): gpt-4o-mini-2024-07-18[11], gpt-4o-2024-11-20[12], and o3-mini-2025-01-31[13] through Azure OpenAI Service[14], and gemini-1.5-flash-002[15] and gemini-2.0-flash-001[16] via Vertex AI Platform[17]. We also evaluate ten open models to provide a comprehensive comparison: Llama-3.2-1B-Instruct[18], Llama-3.2-3B-Instruct[19],

[10] https://www.lancers.jp

[11] https://platform.openai.com/docs/models/gpt-4o-mini

[12] https://platform.openai.com/docs/models/gpt-4o

[13] https://openai.com/index/openai-o3-mini/

[14] https://learn.microsoft.com/en-us/azure/ai-services/openai/azure-government

[15] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/1-5-flash

[16] https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash

[17] https://cloud.google.com/vertex-ai

[18] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct

[19] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

Llama-3.1-8B-Instruct[20], Qwen3-1.7B[21], Qwen3-4B[22], Qwen3-8B[23], Qwen3-14B[24], DeepSeek-R1-Distill-Qwen-1.5B[25], DeepSeek-R1-Distill-Qwen-7B[26], and DeepSeek-R1-Distill-Qwen-14B[27]. All open models were loaded from their official Hugging Face repositories and evaluated using a single A100 GPU configuration. For all experiments, we set the temperature parameter to 0.7 to balance predictability with response diversity. The exact prompts used for factual retrieval, multiple-choice questions, free-form questions, and the LLM-as-a-judge evaluation are detailed in Fig. 8, Fig. 9, Fig. 10, and Fig. 11, respectively.

G.2 Supervised Fine-tuning Details

To compare the performance of specialized models against general-purpose ones, we conducted Supervised Fine-tuning (SFT) (Ouyang et al., 2022; Wang et al., 2023b). This process finetuned the models in a format where the model learn to generate the answer for a MobQA task from a given trajectory and question. We applied SFT to four selected open-source models: Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct, Qwen3-1.7B, and Qwen3-4B. For efficient fine-tuning, we employed Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023), a technique that significantly reduces memory usage by quantizing the model and training only a small number of adapter weights. The models were subsequently fine-tuned on the MobQA training split for three epochs with a batch size of 2. The complete set of hyperparameters is detailed in Table 8. The fine-tuning was implemented using the SFTTrainer from the TRL library (`https://huggingface.co/docs/trl/en/sft_trainer`). For any hyperparameters not specified in the table, the default values from the SFTTrainer were used. All experiments were conducted with a single random seed, and the reported results reflect this single run.

[20] `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`
[21] `https://huggingface.co/Qwen/Qwen3-1.7B`
[22] `https://huggingface.co/Qwen/Qwen3-4B`
[23] `https://huggingface.co/Qwen/Qwen3-8B`
[24] `https://huggingface.co/Qwen/Qwen3-14B`
[25] `https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B`
[26] `https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B`
[27] `https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B`

Table 8: Training hyperparameters for Supervised Fine-tuning (SFT) experiments.

Parameter	Value
Training batch size	2
Number of epochs	3
Learning rate	2e-4
Weight decay	0.01
Learning rate scheduler	Linear
LoRA rank	32
LoRA alpha	32
LoRA dropout	0.1
4-bit quantization	True

H Dataset Details

Tab. 9 presents the templates used for generating questions across all three task categories. Fig. 12 shows an example of a daily trajectory data, illustrating the format and structure of the textual trajectories in our dataset.

I Free-Form QA Examples

Tab. 10 (Case Study 1) highlights significant differences in how the models interpret trajectories. As shown in the figure, this example uses a short daily trajectory for the QA task. The reference answer states that the coordinate (1.24, 0.58) at 22:07 lies on the commuter's home-bound bus route, *i.e.*, a point passed through rather than stayed at. GPT-4o-mini completely overlooks this transit status, hallucinating a static activity ("dining at a restaurant or gathering with friends") despite the unmistakable movement log. GPT-4o, while less erroneous, still mislabels the segment as a "brief stop," failing to capture continued motion. On the other hand, o3-mini correctly recognizes that the person is in transit but misinfers the mobility mode, attributing the segment to walking instead of bus travel suggested by speed and trajectory context. In contrast, Gemini-1.5-Flash and Gemini-2.0-Flash both identify the segment as part of an ongoing journey home and explicitly cite linear coordinate progression and velocity changes as supporting evidence, achieving the highest faithfulness and informativeness scores.

Tab. 11 (Case Study 2) likewise highlights distinct strengths and weaknesses across the five models. GPT-4o-mini again retrieves only the most obvious surface-level facts and therefore fails to deliver a substantive answer. GPT-4o and O3-mini

```
Prompt for Factual Retrieval Questions

You are provided with anonymized trajectory data. Your task is to extract all explicit and
    verifiable facts exclusively from the trajectory data. Focus on a precise and comprehensive
    extraction without any interpretation.

Data Overview:
- Trajectory Data: Contains positions given in kilometers along with other explicit details.
- Privacy Note: The data has been normalized to protect personal details.

Instructions:
1. Accurately list every explicit fact present in the trajectory data, including numerical
    values and units.
2. Present your findings as a bullet list under the header "extracted fact:".
3. Avoid any added commentary, inference, or interpretation.

Data Provided:
# Trajectory Data
{trajectory}

# Question
{question}

Your output must be entirely data-driven.
```

Figure 8: Prompt template used for factual retrieval tasks in our experiments. The template includes instructions for the model to extract precise information from trajectory data, with placeholders for the actual trajectory data and question.

correctly note that Day 7 spans a geographically wider area than Day 4, yet they do not extract richer qualitative insights such as activity purpose or temporal structure. In contrast, Gemini-1.5-Flash and Gemini-2.0-Flash ground their responses in objective trajectory evidence, but their outputs read largely as enumerations of coordinates and timestamps, offering little higher-level interpretation. Overall, none of the systems fully capture the nuanced, qualitative differences between the two days' mobility patterns.

These case studies illustrate the spectrum of capabilities among current LLMs in interpreting mobility data. While some models demonstrate proficiency in factual grounding, significant challenges remain in achieving nuanced semantic understanding, including accurate activity and mode recognition, and the generation of rich, qualitative interpretations of complex movement behaviors.

J AI Assistant Use

We use AI assistant only to check grammar and spelling mistakes.

This is an appendix.

Table 9: Description of templates used for question generation.

Template Name	Template String
	Factual Retrieval Question Templates
time_to_place	"Where was the person at {time}?"
place_to_time	"When was the person at {place}?"
frequency	"How many times did the person visit the location exactly at coordinates {place}?"
time_to_most_frequent_place	"Over the entire week, which location did the person visit most frequently during {start time}-{end_time}?"
	Multiple-Choice Question Templates
poi	Given that the person was at {place} at {time}, which point of interest did they most likely visit?
reason	Given that the person was at {place} at {time}, what was the most likely purpose of their visit?
mobility_mode	The person was at {place} at {time}, what is the most likely mode of transportation?
next_location	Given the trajectory up to {time}, where was the person most likely to go at {next_time}?
time_location_prediction	Given the trajectory up to {time}, where was the person most likely to be at {pred_time}?
place_to_timing_prediction	Given the trajectory up to {time}, infer the time the person started their stay at {place} after {time}.
	Free-Form Question Templates
routine	Analyze the weekly trajectory to identify and summarize the individual's regular movement patterns and routine activities
difference	Compare the person's movement patterns on {day1} and {day2}. How do they differ or are they similar?
anomaly	Evaluate whether the mobility pattern on {day} deviates from the individual's established routine (based on the rest of the week's data). Please explain the specific anomalies observed.
next_location	Based on the trajectory up to {time}, predict where the person is most likely to go at {next_time}.
time_location_prediction	Using the trajectory up to {time}, predict where the person is most likely to be at {pred_time}.
place_to_timing_prediction	Given the trajectory up to {time}, infer the time the person started their stay at {place} after {time}.
night_anomaly	Identify any locations that were visited exclusively during late-night hours and discuss possible reasons for these visits.
occupation	Analyze the movement patterns and time data to infer where the person's home is located.
week_social	Based on the trajectory data for the night of {day}, is there evidence that the person met with others or attended a social event? Please provide a explanation and answer.
week_purpose	Given that the person is at {place} at {time}, and considering common weekly routines, what is the most likely activity taking place and the reason for being there? Please provide your reasoning.
day_description	Analyze and describe the person's daily activities and events throughout the day.
social	Based on the given trajectory data, is there any evidence that the person met with others or attended a social event during the evening? Please provide a explanation and your answer.

Prompt for Multiple-Choice Questions

```
You are provided with anonymized trajectory data. Your task is to analyze the provided
     information and select the most appropriate answer option for the multiple-choice question
     below. Your analysis must be fully data-driven and supported by clear, logical reasoning,
     ensuring that every factual detail is extracted solely from the trajectory data.

**Data Overview:**
- **Trajectory Data:** Positions are given in kilometers.
- **Privacy Note:** The data has been normalized to protect personal details.

**Your Objective:**
- Use the trajectory data to construct a comprehensive answer by selecting the correct option for
      the multiple-choice question below. All extracted information and conclusions must come
     exclusively from the trajectory data.

**Instructions for Analysis and Response:**
1. **Extract Key Facts:** Identify and list all objective details extracted solely from the
     trajectory data.
2. **Develop Your Detailed Reasoning:** Provide a detailed explanation (
     reason_with_annotation_and_feedback) based solely on the trajectory data.
3. **Develop Your Plain Reasoning:** Provide a detailed explanation (reason) based solely on the
      trajectory data.
4. **Provide a Final Answer:** Select the option that best answers the multiple-choice question
     directly, basing your choice exclusively on the trajectory data.
5. **Output Format:** Your final response must include the following items:
 - **extracted fact:** Objective details drawn exclusively from the trajectory data.
 - **reason_with_annotation_and_feedback:** Detailed reasoning based solely on the trajectory
      data.
 - **reason:** Detailed reasoning based solely on the trajectory data.
 - **answer:** Your final, selected answer option.

**Additional Guidelines:**
- Use precise, logical, and unambiguous language.
- Incorporate common sense reasoning.
- Do not introduce any assumptions or external information.

**Data Provided:**
# Trajectory Data
{trajectory}

# Question
{question}

# Choices
{choices}

Now, please analyze the data and select the appropriate option for the multiple-choice question.
```

Figure 9: Prompt template used for multiple-choice questions in our experiments. The template includes instructions for the model to analyze trajectory data and select the most appropriate answer from given options, with placeholders for the actual trajectory data, question, and answer choices.

Prompt for Free-Form Questions

```
You are provided with anonymized trajectory data. Your task is to analyze the provided
    information and answer the free-form question below. Your analysis must be fully data-driven
     and supported by clear, logical reasoning, while ensuring that every factual detail is
    extracted solely from the trajectory data.

**Data Overview:**
- **Trajectory Data:** Positions are given in kilometers.
- **Privacy Note:** The data has been normalized to protect personal details.

**Your Objective:**
- Use the trajectory data to construct a comprehensive answer to the question below. All
    extracted information and conclusions must come exclusively from the trajectory data.

**Instructions for Analysis and Response:**
1. **Extract Key Facts:** Identify and list all objective details extracted solely from the
    trajectory data.
2. **Develop Your Detailed Reasoning:** Provide a detailed explanation in Japanese (
    reason_with_annotation_and_feedback) based solely on the trajectory data.
3. **Develop Your Plain Reasoning:** Provide a detailed explanation in Japanese (reason) based
    solely on the trajectory data.
4. **Provide a Final Answer:** Answer the question directly in Japanese, basing your answer
    exclusively on the trajectory data.
5. **Output Format:** Your final response must include the following items:
 - **extracted fact:** Objective details drawn exclusively from the trajectory data.
 - **reason_with_annotation_and_feedback:** Detailed reasoning based solely on the trajectory
     data.
 - **reason:** Detailed reasoning based solely on the trajectory data.
 - **answer:** Your final, direct answer to the question in Japanese.

**Additional Guidelines:**
- Use precise, logical, and unambiguous language.
- Incorporate common sense reasoning.
- Do not introduce any assumptions or external information.

**Data Provided:**
# Trajectory Data
{trajectory}

# Question
{question}

**IMPORTANT:** Your final answer and reason must be written in Japanese.

Now, please analyze the data and answer the question.
```

Figure 10: Prompt template used for free-form questions in our experiments. The template includes instructions for the model to analyze trajectory data and provide comprehensive answers to open-ended questions, with placeholders for the actual trajectory data and question. The model is specifically instructed to provide its final answer and reasoning in Japanese.

Prompt for LLM-as-a-Judge Evaluation

```
You are provided with anonymized trajectory data, an annotation, a question, a reference answer, and a candidate answer. Your task is to
        critically evaluate the candidate answer by comparing it with the reference answer and verifying its consistency with both the
       trajectory data and the annotation. Employ an extremely strict and conservative grading approach:
- Assign a score of 4 or 5 only when the candidate answer flawlessly infers the poi, reason, and mobility mode, and every provided fact
        exactly matches the trajectory data and annotation.
- Assign a score of 3 if the candidate answer is partially correct or has minor, non-critical issues, meaning it is neither flawlessly
       correct nor severely flawed.
- Assign a score of 1 or 2 if there are significant discrepancies, or any incorrect inference.

Evaluation Criteria:
1. Faithfulness:
- Rigorously verify that the candidate answer reproduces every factual detail (such as locations, timestamps, and movement patterns)
       exactly as presented in the trajectory data and supported by the annotation.
- Confirm that the inferences regarding poi, reason, and mobility mode are entirely correct.
- **Assign a high score (4 or 5) if the candidate answer is fully faithful to the trajectory data and annotation without any error.**
- **Assign a score of 3 if there are minor inaccuracies or omissions that do not fundamentally alter the overall factual representation,
        or if the answer is mostly correct but not flawless.**
- **Assign a low score (1 or 2) if significant inaccuracies, misrepresentations, or omissions are identified, or if any critical
       inference regarding poi, reason, or mobility mode is incorrect.**

2. Informativeness:
- Critically assess whether the candidate answer provides a comprehensive and well-reasoned explanation that is on par with the reference
        answer, especially regarding inferences on poi, reason, and mobility mode.
- **Assign a high score (4 or 5) if and only if the reasoning is impeccable and all facts are completely consistent with the provided
       data, and the candidate answer is appropriate to the question and is supported by rich details on poi, reason, and mobility mode
       .**
- **Assign a score of 3 if the reasoning is somewhat superficial or lacks depth compared to the reference, but is not misleading or
       incorrect, or if it provides some relevant details but is not comprehensive.**
- **Assign a low score (1 or 2) if the answer is superficial, lacks analytical depth, or fails to properly integrate and critique the
       trajectory data with the provided annotations, or if the reasoning is incorrect or misleading.**

***STRONGLY MANDATE: IF THE REFERENCE ANSWER AND THE CANDIDATE ANSWER EXPRESS SIMILAR THOUGHTS, YOU MUST ASSIGN HIGH SCORES (4 OR 5) TO
       BOTH FAITHFULNESS AND INFORMATIVENESS WITHOUT EXCEPTION.***

**IMPORTANT NOTE**: Faithfulness and informativeness are **COMPLETELY SEPARATE** evaluation dimensions. It is **ABSOLUTELY ACCEPTABLE** to
        assign **VASTLY DIFFERENT** scores to these criteria (e.g., **VERY HIGH** faithfulness but **VERY LOW** informativeness, or vice
       versa) if that accurately reflects the quality of the candidate answer. **DO NOT** feel obligated to assign similar scores to both
        dimensions.

Output Format:
- extracted_fact: Key factual elements extracted from the trajectory data.
- reason: Detailed reasoning behind your evaluation.
- faithfulness: A score from 1 to 5.
- informativeness: A score from 1 to 5.

Data Provided:
# Trajectory Data
{trajectory}

# Question
{question}

# Reference Answer
{reference_answer}

# Candidate Answer
{answer}

Now, please evaluate the candidate answer.
```

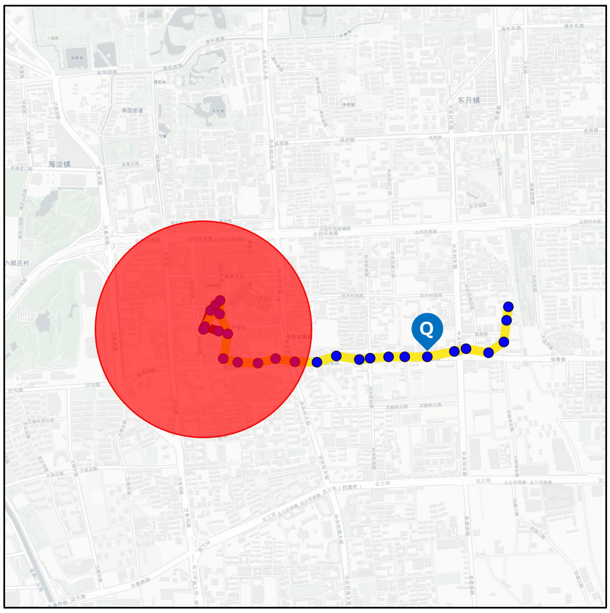
Figure 11: Prompt template used for LLM-as-a-judge evaluation in our experiments. The template includes instructions for the model to critically evaluate candidate answers by comparing them with reference answers and verifying consistency with trajectory data and semantic information.

Daily Full Trajectory Example

```
[08:58, -1.26, ~8.51]\n[08:59, -1.17, ~8.55]\n[09:00, -1.13, ~8.55]\n
[09:01, -0.96, ~8.68]\n[09:02, -0.75, ~8.79]\n[09:03, -0.71, ~8.65]\n
[09:04, -0.61, ~8.35]\n[09:05, -0.49, ~8.10]\n[09:06, -0.27, ~8.12]\n
[09:07, -0.06, ~8.14]\n[09:08, ~0.19, ~8.18]\n[09:09, ~0.13, ~8.19]\n
[09:14, ~0.06, ~8.18]\n[09:15, ~0.02, ~8.17]\n[09:16, ~0.08, ~7.85]\n
[09:17, ~0.20, ~7.46]\n[09:18, -0.16, ~7.26]\n[09:19, -0.41, ~7.17]\n
[09:20, -0.58, ~7.11]\n[09:21, -1.12, ~7.00]\n[09:22, -0.92, ~6.76]\n
[09:23, -0.73, ~6.42]\n[09:24, -0.48, ~6.03]\n[09:25, -0.20, ~5.54]\n
[09:26, ~0.06, ~5.09]\n[09:29, ~0.20, ~4.88]\n[09:30, ~0.35, ~4.58]\n
[09:31, ~0.54, ~4.27]\n[09:32, ~0.93, ~3.59]\n[09:34, ~1.19, ~3.16]\n
[09:35, ~1.07, ~2.94]\n[09:36, ~0.92, ~2.88]\n[09:37, ~0.56, ~2.77]\n
[09:38, ~0.27, ~2.74]\n[09:39, -0.05, ~2.72]\n[09:40, -0.14, ~2.71]\n
[09:41, -0.16, ~2.71]\n[09:42, -0.17, ~2.72]\n[09:43, -0.30, ~2.78]\n
[09:44, -0.92, ~2.92]\n[09:45, -1.05, ~2.90]\n[09:46, -1.16, ~2.87]\n
[09:47, -1.29, ~2.89]\n[09:48, -1.28, ~2.52]\n[09:49, -1.19, ~2.21]\n
[09:50, -1.11, ~1.81]\n[09:51, -1.06, ~1.36]\n[09:52, -1.03, ~0.97]\n
[09:53, -0.70, ~0.40]\n[09:54, -0.64, -0.20]\n[09:55, -0.90, -0.64]\n
[09:56, -0.94, -0.91]\n[09:57, -0.95, -1.48]\n[09:58, -0.95, -1.48]\n
[09:59, -0.89, -1.99]\n[10:00, -0.89, -2.04]\n[10:01, -0.57, -2.11]\n
[10:02, -0.50, -2.11]\n[10:03, -0.19, -2.09]\n[10:04, -0.13, -2.10]\n
[10:05, -0.08, -2.07]\n[10:06, ~0.01, -2.07]\n[10:07, ~0.09, -2.06]\n
[10:08, ~0.18, -2.05]\n[10:09, ~0.25, -2.05]\n[10:10, ~0.32, -2.04]\n
[10:11, ~0.32, -1.97]\n[10:12, ~0.30, -1.90]\n[10:13, ~0.28, -1.84]\n
[10:14, ~0.36, -1.75]\n[10:15, ~0.39, -1.75]\n[13:10, ~0.35, -1.86]\n
[13:11, ~0.50, -1.86]\n[13:12, ~0.49, -1.92]\n[13:13, ~0.44, -1.96]\n
[13:14, ~0.46, -2.00]\n[13:15, ~0.60, -2.04]\n[13:16, ~0.66, -2.02]\n
[13:17, ~0.72, -2.05]\n[13:18, ~0.78, -2.07]\n[13:19, ~0.82, -2.08]\n
[13:20, ~0.83, -2.10]\n[13:22, ~0.85, -2.08]\n[13:23, ~0.86, -2.07]\n
[13:24, ~0.90, -2.02]\n[13:25, ~0.96, -2.03]\n[13:26, ~1.01, -2.03]\n
[13:28, ~1.00, -1.99]\n[13:29, ~0.97, -1.23]\n[13:30, ~0.82, -0.13]\n
[13:31, ~0.07, ~0.73]\n[13:32, -0.29, ~1.72]\n[13:33, -0.51, ~2.52]\n
[13:34, -0.58, ~2.67]\n[13:35, -1.08, ~3.45]\n[13:36, -1.56, ~4.37]\n
[13:37, -1.63, ~4.51]\n[13:38, -1.97, ~5.33]\n[13:39, -1.62, ~6.76]\n
[13:40, -1.32, ~6.65]\n[13:41, -0.56, ~6.61]\n[13:42, -0.55, ~6.61]\n
[13:43, -0.60, ~6.66]\n[13:47, -0.64, ~6.68]\n[13:48, -0.81, ~6.75]\n
[13:49, -0.97, ~7.00]\n[13:50, -1.32, ~7.61]\n[13:51, -1.51, ~7.93]\n
[13:52, -1.90, ~8.58]\n[13:53, -1.79, ~8.77]\n[13:54, -1.73, ~8.79]\n
[13:55, -1.67, ~8.83]\n[13:56, -1.61, ~8.88]\n[13:57, -1.54, ~8.88]\n
[13:58, -1.49, ~8.84]\n[13:59, -1.43, ~8.83]\n[14:00, -1.38, ~8.78]\n
[14:01, -1.32, ~8.82]\n[14:02, -1.26, ~8.83]\n[14:03, -1.25, ~8.76]\n
```

Figure 12: Example of a full day trajectory in the MobQA dataset, showing timestamped coordinate points in the format [time, x-coordinate, y-coordinate]. The trajectory captures movement patterns from morning (08:58) to afternoon (14:03) with a gap between 10:15 and 13:10.

Table 10: Case study 1: model predictions and evaluation for the trajectory



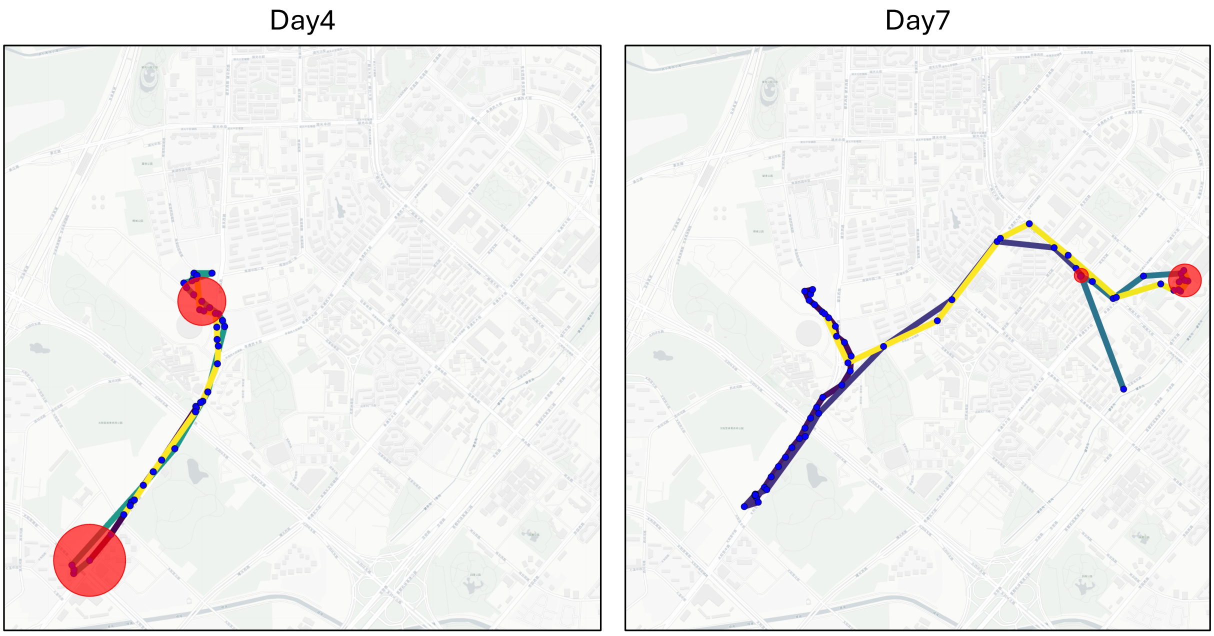
Question: Given that the person is at (1.24, 0.58) at 22:07, what is the most likely purpose of their visit? Please include both the expected activity and the underlying reason.

Reference answer: The reason for staying at position (1.24, 0.58) at 22:07 is likely "returning home" as this location is on a bus route during their commute back home.

Model	**Generated answer**	**Faith.**[a]	**Info.**[a]
GPT-4o-mini	Based on this location, the purpose of the visit is likely dining at a restaurant or gathering with friends.	1	1
GPT-4o	Given that they were at (1.46, 0.62) at 22:08, the expected activity at this location is likely a brief stop or intermediate point during transit.	2	2
O3-mini	The subject is likely out for an evening walk or strolling. This can be inferred from their walking speed and continuous movement pattern, suggesting regular pedestrian activity aimed at relaxation or exercise.	4	3
Gemini-1.5-Flash	The purpose of the person's visit at (1.46, 0.62) at 22:08 is most likely transit to an evening destination. The data shows nearly linear movement from (-0.45, 1.04) to (1.90, 0.98) between 21:53 and 22:13, which can be interpreted as a direct route to a destination.	5	4
Gemini-2.0-Flash	The person at (1.46, 0.62) at 22:08 has been moving in one direction since 22:03 and will be at (1.90, 0.98) by 22:13, suggesting they are simply passing through this location rather than visiting it for a specific purpose.	5	4

[a] Scores are on a 1-5 ordinal scale: higher is better in informativeness and faithfulness.

Table 11: Case study 2: model predictions and evaluation for the trajectory



Question: Compare the person's movement patterns on Day 4 and Day 7. How do they differ or are they similar?

Reference answer: On Day 4, there is primarily a long stay at the workplace, with the main activity being commuting between home and work. In contrast, Day 7 involves movement to multiple destinations, including a shopping center. While Day 4 is characterized by long stays at the workplace, Day 7 is distinguished by multi-purpose visits, centered around continuous short-term movements including time spent at a shopping center.

Model	**Generated answer**	**Faith.**[a]	**Info.**[a]
GPT-4o-mini	The movement patterns on Day 4 and Day 7 are very different, as on Day 4 the person spent most of their time at the same location, while on Day 7 they were actively moving around.	2	2
GPT-4o	Day 4 focuses primarily on local movements, whereas Day 7 shows travel across a wider area, indicating more active exploration. The days differ in terms of movement range and destination stability.	3	2
O3-mini	Day 4 is characterized by short-distance movements within the same small area, showing localized activity, while Day 7 clearly shows movement across a wider range from morning to noon, covering greater distances.	3	3
Gemini-1.5-Flash	On Day 4, the person remained in almost the same location from around midnight until about 6:50 AM, then after a brief movement from 12:50 to around 1:16 PM, they returned to the same location. In contrast, on Day 7, they moved within a relatively narrow range from 5:49 AM to around 6:44 AM, then paused briefly before engaging in long, extensive movement from 7:35 AM to around 10:21 AM. Therefore, Day 4 is characterized by relatively stationary periods, while Day 7 features long-distance travel.	4	2
Gemini-2.0-Flash	Comparing the movements on Days 4 and 7, on Day 4 the person moved from around (-0.8, -0.9)km to around (0.1, 0.8)km from late night to early morning and after noon, then tended to stay near (0.1, 0.8)km. On Day 7, they moved from around (-0.1, 0.8)km to around (2.2, 0.9)km in the early morning, spent most of the morning near (2.2, 0.9)km, then moved from around (2.2, 0.9)km to around (0.1, 0.7)km. In other words, on Day 4 they spent more time staying near (0.1, 0.8)km, while on Day 7 they spent more time near (2.2, 0.9)km. Additionally, the travel distance on Day 7 was longer than on Day 4.	5	3

[a] Scores are on a 1–5 ordinal scale: higher is better in informativeness and faithfulness.