TravelBehaviorQA: A Benchmark Dataset for Behavioral Interpretation of GPS Trajectories

Dongyang Zhen[1], Niping Duan[1], Huan Zhou[1], Qingbin Cui[†1]
[1]University of Maryland, College Park
Correspondence: cui@umd.edu

Abstract

GPS trajectories encode rich behavioral information about how people move, organize activities, and form daily routines. Recent advances in large language models (LLMs) raise a natural question: can such models infer and summarize travel behavior directly from mobility traces? This paper introduces *TravelBehaviorQA*, a large-scale benchmark dataset that reframes trajectory analysis as a language-based behavioral understanding task. The dataset links raw GPS trajectories with human-grounded question-answering (QA) pairs that capture travel intensity, temporal structure, activity patterns, mode usage, and behavioral routines. Unlike prior mobility datasets focused on prediction or classification, *TravelBehaviorQA* emphasizes semantic interpretation through a unified mix of deterministic and open-ended questions. In this benchmark, we construct over 143k QA instances spanning users and years, and evaluate a broad range of state-of-the-art LLMs under controlled settings. Our results reveal substantial gaps between factual extraction and genuine behavioral reasoning, showing that model scale alone is insufficient and that trajectory representation is a primary bottleneck. *TravelBehaviorQA* exposes critical limitations of current models and establishes a rigorous benchmark for advancing language-based understanding of human mobility behavior. The dataset is available at `https://github.com/dongyangzhen/TravelBehaviorQA`

1 Introduction

Recent advances in large language models (LLMs) have enabled strong performance on reasoning, summarization, and dialogue tasks across diverse domains (Liu et al., 2025c; Wu et al., 2025; Wang et al., 2025). However, non-textual behavioral traces such as GPS trajectories present a distinct challenge (Kong et al., 2025). Unlike natural language, trajectories are composed of continuous spatiotemporal observations rather than discrete tokens

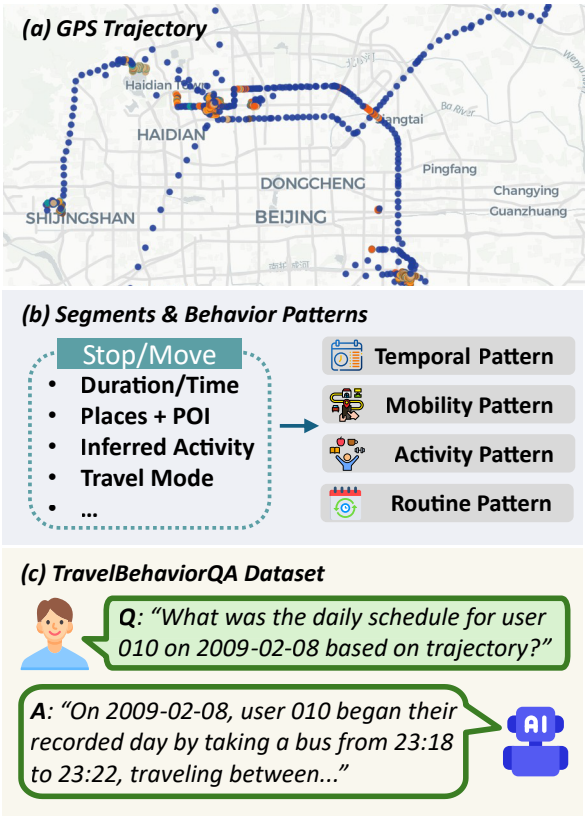

Figure 1: Overview of the TravelBehaviorQA dataset.

and are typically devoid of explicit semantic annotations (Liu et al., 2025a; Yang et al., 2025b). In transportation research, GPS trajectories are highly informative, capturing individuals' movements through space and time and implicitly reflecting travel modes, schedules, and habitual routines (Zaroujtaghi et al., 2025). Yet this information is encoded in a low-semantic form: raw latitude–longitude coordinates with timestamps require substantial interpretation to yield meaningful behavioral insights (Zhu et al., 2023).

Existing mobility datasets and modeling efforts predominantly target prediction or classification tasks, such as next-location forecasting and travel mode detection (Gong et al., 2024; Molloy et al.,

Findings of the Association for Computational Linguistics: ACL 2026, pages 32053–32071
July 2-7, 2026

2023; Yabe et al., 2024). As a result, they provide limited insight into whether LLMs can interpret trajectories at a behavioral level or generate human-interpretable summaries of travel patterns (Wang et al., 2024). In contrast, this benchmark formulates GPS trajectory understanding as a language-based reasoning task rather than a transportation modeling problem. Model inputs consist of serialized spatiotemporal traces, augmented with structured abstractions such as trip segmentation and place semantics, expressed in textual form. Outputs are natural-language answers whose structure and complexity vary by question type. *TravelBehaviorQA* spans both low-entropy tasks (deterministic question), which require factual aggregation and temporal querying, and high-entropy tasks (open-ended question), which require multi-sentence summarization and discourse-level reasoning over extended behavioral contexts. These tasks explicitly probe NLP challenges, including temporal ordering, salience selection, implicit inference of intent, and faithfulness to non-linguistic evidence. By grounding LLMs in spatiotemporal traces and jointly evaluating factual accuracy and narrative abstraction, *TravelBehaviorQA* targets fundamental capabilities in structured reasoning and evidence-based generation that are not assessed by conventional text-only benchmarks.

2 Related Work

2.1 LLMs for mobility inference and trajectory understanding

Recent work has explored the use of LLMs and foundation models for semantic inference from human mobility data (Shen et al., 2025; Qiu et al., 2023; Ghosh et al., 2024). Rather than operating directly on raw coordinates, most approaches combine GPS trajectories with auxiliary information such as points of interest (POI), maps, human annotations, or textual context to infer higher-level semantics (Huang et al., 2024; Xiao et al., 2023; Wang et al., 2023). LLMs have been applied to tasks including POI classification (Cheng et al., 2025), activity and trip-purpose inference (Li et al., 2024), next-location reasoning (Chen et al., 2025), and semantic trajectory summarization (Shao et al., 2025). Several studies reformulate mobility analysis as a language reasoning problem, prompting LLMs to interpret sequences of visited locations or stay points and generate explanations of travel behavior (Zhang and Xu, 2025; Ge et al., 2025; Liu et al., 2025b). Others use LLMs as agents or controllers to integrate multiple mobility models (Lan et al., 2024), or to generate realistic synthetic trajectories that reflect human routines (Yang et al., 2025a). While these approaches demonstrate the promise of LLMs for mobility understanding, they are evaluated on simple objectives such as classification accuracy or prediction performance. Thus, existing work offers limited evaluation of whether LLMs can serve as general behavioral interpreters of mobility trajectories, motivating our study to assess trajectory understanding through a unified QA framework grounded in human interpretations.

2.2 Mobility benchmarks and datasets

Several datasets and benchmarks have been developed to evaluate semantic understanding of human mobility data (Stanford et al., 2024; Lai et al., 2022; Nilforoshan et al., 2023; Abbiasov et al., 2024). Traditional mobility datasets primarily support supervised learning tasks such as transportation mode detection (Fu et al., 2025), activity recognition (Xiao and Tong, 2025), or trip-purpose classification (Yin et al., 2026), relying on predefined labels and task-specific evaluation protocols. More recently, QA–based benchmarks have been introduced to assess higher-level reasoning over trajectories, including factual retrieval, temporal reasoning, and narrative explanation (Hsiao et al., 2025; Reichman et al., 2025). These efforts have highlighted that while language models can handle simple fact-based queries, they often struggle with deeper semantic interpretation and coherent behavioral summarization (Asano et al., 2025). However, most existing benchmarks focus on deterministic factual retrieval from mobility data, with limited emphasis on inferring higher-level behavioral patterns, intent, or travel purpose. In contrast, *TravelBehaviorQA* integrates both deterministic and open-ended questions within a unified evaluation framework and aligns raw GPS trajectories with human-grounded behavioral descriptions. This design enables systematic assessment of factual accuracy alongside interpretive reasoning, providing a more comprehensive benchmark for evaluating trajectory understanding and mobility behavior interpretation.

3 Dataset Description

3.1 Overall dataset composition

TravelBehaviorQA comprises 143,238 QA pairs grounded in GeoLife GPS trajectories (Zheng et al.,

2011), spanning 68 users and 6,305 user-days over a four-year period. The dataset is organized into 11 behavioral categories and 30 fine-grained question types, with summary statistics reported in Table 1. It combines deterministic, fact-based questions with open-ended questions requiring narrative synthesis, supporting both controlled evaluation and semantic reasoning.

Table 1: Overall dataset statistics.

Statistic	Value
Total QA instances	143,238
Users	68
User-days	6,305
Temporal span	2007–2011
Behavioral categories	11
Question types	30
Open questions	35,666 (24.9%)
Deterministic questions	107,572 (75.1%)

3.2 Behavioral question categories

TravelBehaviorQA is designed to capture complementary aspects of human travel behavior while explicitly separating questions by their reasoning demands (Figure 2). The benchmark organizes behavioral queries along dimensions such as temporal organization, activity structure, spatial semantics, and mode usage (Figure 3). This separation enables targeted assessment of different forms of reasoning over the same underlying mobility evidence.

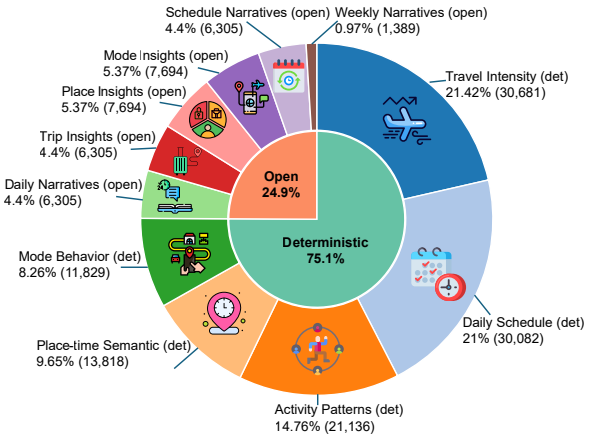

Figure 2: Category distribution.

Deterministic questions are constructed to evaluate factual grounding and consistency, requiring models to identify, aggregate, or quantify behavioral attributes that are directly supported by trajectory data. These questions reflect low-entropy inference, where correct answers are uniquely determined by observable spatiotemporal patterns. In contrast, open-ended questions probe higher-level behavioral interpretation, asking models to synthesize information across trips, activities, and time to produce coherent narratives or insights. Such questions introduce inherent ambiguity and demand abstraction, summarization, and selective emphasis rather than direct computation.

3.3 Comparison with other datasets

Recent mobility-related QA datasets such as MobQA (Asano et al., 2025) and POI-QA (Han et al., 2025) have taken important first steps toward grounding natural language questions in spatio-temporal data. However, these datasets primarily focus on localized factual queries or short explanations tied to individual trajectories or POIs, and thus predominantly evaluate low-entropy retrieval or shallow reasoning. From an NLP standpoint, they largely test whether models can map structured mobility signals to isolated textual facts, rather than whether models can construct coherent, faithful, and temporally grounded behavioral interpretations.

Table 2: Comparison of *TravelBehaviorQA* with existing mobility-related QA datasets from an NLP perspective.

Dataset	Primary Task Scope	Evaluated NLP Capabilities	# QA Pairs
MobQA	Single-trajectory semantics	Fact retrieval, localized explanation	∼5.8K
POI-QA	POI-centered spatio-temporal queries	Spatial grounding, temporal lookup	∼10K
TravelBehaviorQA	**Multi-scale behavioral understanding**	**Factual extraction, temporal abstraction, narrative summarization, faithfulness**	**∼140K**

In contrast, *TravelBehaviorQA* is designed to evaluate semantic trajectory understanding as a language reasoning problem. Its primary contribution lies in framing mobility data as a source of extended behavioral context that supports both deterministic questions and high-entropy, open-ended narratives spanning daily, weekly, and longitudinal time scales. This design enables the assessment of core NLP capabilities that prior datasets do not capture, including temporal abstraction, salience selection, discourse-level summarization, and faithfulness to non-linguistic evidence. By explicitly pairing structured trajectory representations with natural-language behavioral explanations at scale,

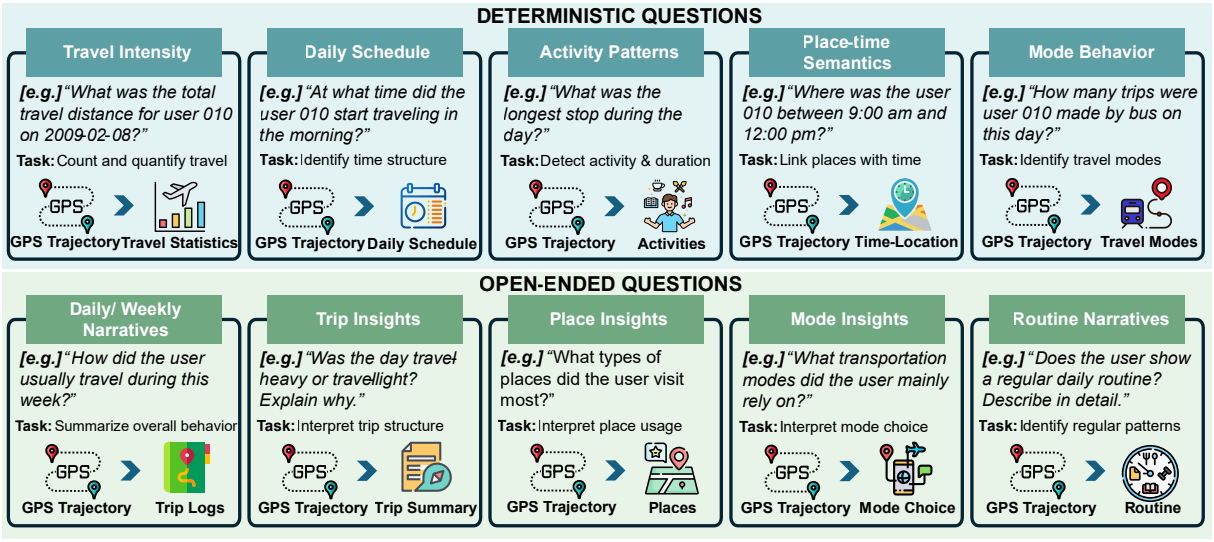

Figure 3: Examples of deterministic and open-ended question types in *TravelBehaviorQA*.

TravelBehaviorQA provides a substantially richer benchmark for diagnosing the limitations of current language models in grounded reasoning and narrative generation beyond text-only settings.

4 Methodology

4.1 Data construction

The dataset is constructed through a structured pipeline (Figure 4) that converts raw GPS trajectories into a behavior-centered QA benchmark, progressively abstracting low-level mobility traces into interpretable representations suitable for systematic evaluation of trajectory understanding.

Raw GPS trajectory data. The raw GPS trajectories are drawn from the GeoLife dataset (Zheng et al., 2011), comprising 17,621 trajectories collected from 178 users over a period exceeding four years. The data span approximately 1.25 million kilometers and 48,000 hours of movement, with trajectories represented as sequences of time-stamped GPS points. Most trajectories are densely sampled and provided as raw spatiotemporal records with limited semantic annotation, capturing diverse real-world mobility behaviors across multiple cities.

Trip segmentation. To enable behavioral interpretation of raw GPS trajectories, we first segment continuous point sequences into semantically meaningful *stop* and *move* units. Such segmentation is a necessary preprocessing step for transforming low-level spatiotemporal observations into higher-level behavioral insights (activities and trips). Given its demonstrated accuracy and robustness for GPS-based trip segmentation, we adopt ST-DBSCAN to identify stop regions using joint spatial and temporal constraints (Birant and Kut, 2007). Formally, a trajectory is represented as an ordered sequence of GPS points $\{(x_i, y_i, t_i)\}$. For each point p_i, ST-DBSCAN defines a spatiotemporal neighborhood

$$\mathcal{N}(p_i) = \{p_j \mid d_{\text{space}}(p_i, p_j) \leq \varepsilon_{\text{space}}, \; d_{\text{time}}(p_i, p_j) \leq \varepsilon_{\text{time}}\} \quad (1)$$

where $d_{\text{space}}(\cdot)$ denotes geographic distance and $d_{\text{time}}(\cdot)$ denotes temporal separation. A point is considered a core point if $|\mathcal{N}(p_i)| \geq \text{minPts}$. Starting from each core point, clusters are expanded by iteratively adding density-connected neighbors that satisfy the same spatiotemporal constraints, while points that fail to meet the density criterion are labeled as noise.

Travel mode prediction. Each *move* segment is assigned a transportation mode using a supervised gradient-boosted decision tree model implemented with XGBoost, selected for its strong empirical performance in our evaluation (Appendix A.3). We extract aggregate kinematic and geometric features for each segment, including distance, duration, speed, acceleration, jerk, point count, and bearing-rate variability. Ground-truth mode labels from the GeoLife dataset are used when available, and the trained model is applied to infer modes for unlabeled segments.

POI annotation. To provide semantic context for *stop* locations, each detected stop is enriched with nearby points of interest using the Foursquare

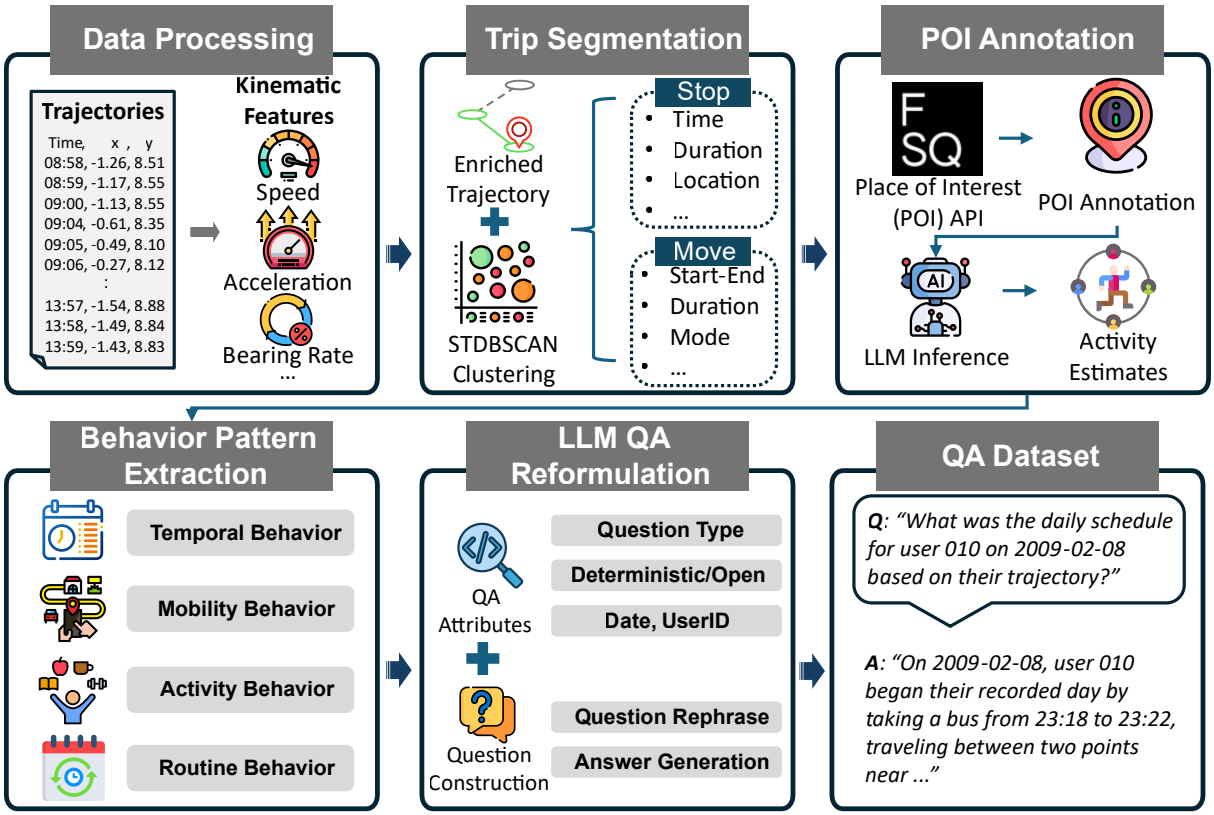

Figure 4: *TravelBehaviorQA* dataset construction pipeline.

POI API. For each stop, we retrieve the three closest POIs along with their categories and distances, using the earliest available records to maintain temporal consistency with the historical trajectories. These POI annotations serve as contextual cues rather than definitive labels and support downstream activity inference and grounded question generation.

QA generation. In the final stage, enriched trajectories are aggregated to derive higher-level behavioral patterns and construct the QA dataset. Structured summaries capturing temporal organization, activity engagement, place usage, and mobility intensity are provided as inputs to LLMs, which generate natural-language QA pairs under predefined templates. Deterministic questions are derived directly from computed attributes, while open-ended questions prompt narrative summaries that are constrained to remain faithful to the underlying trajectory evidence.

4.2 Human evaluation

We conduct a human evaluation with five domain experts, each independently rating 100 randomly sampled QA instances across users, behavioral categories, and question types, with deterministic and open-ended questions evaluated separately to reflect their distinct objectives.

Deterministic questions are evaluated for correctness on a five-point Likert scale. Open-ended questions are evaluated along three dimensions: correctness, clarity, and informativeness, using the same scale. All instances are independently rated by the five evaluators, and scores are averaged across raters. Mean scores with 95% confidence intervals are reported. Inter-rater reliability is measured using Krippendorff's α, defined as $\alpha = 1 - D_o/D_e$, where D_o and D_e denote observed and expected disagreement. Table 3 summarizes the results, showing high correctness and strong agreement for deterministic questions and consistently strong ratings with substantial agreement for open-ended questions, confirming the dataset's reliability and narrative quality.

Table 3: Human evaluation results with 95% confidence intervals and inter-rater agreement.

Question Type	Dimension	Score (95% CI)	α
Deterministic	Correctness	4.63 [4.52, 4.74]	0.81
Open-ended	Correctness	4.31 [4.18, 4.44]	0.74
	Clarity	4.52 [4.41, 4.63]	0.78
	Informativeness	4.18 [4.05, 4.31]	0.71

4.3 Benchmark definition and evaluation setup

We define *TravelBehaviorQA* as a dual-task benchmark that evaluates both factual accuracy and behavioral reasoning in trajectory-based question answering. Deterministic questions admit a single correct answer and are evaluated using exact-match accuracy. Open-ended questions require narrative responses and are evaluated using behavior-aware metrics designed to assess semantic grounding and coherence beyond surface-level similarity.

For open-ended evaluation, we introduce three metrics that operate exclusively on paired natural-language answers, without requiring access to underlying trajectories or structured annotations. Let a^* denote the reference answer and $\hat{a}$ the model-generated answer.

Slot-Level Factual Agreement (SFA). SFA measures whether the generated answer reproduces atomic behavioral facts explicitly stated in the reference narrative. For each question type, we define a slot schema $\mathcal{S} = \{s_1, \ldots, s_K\}$ capturing interpretable attributes such as times, modes, counts, and activity categories. Slot values are independently extracted from a^* and $\hat{a}$ and compared as

$$\mathrm{SFA} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}\big(\hat{v}_k \approx v_k^*\big), \tag{2}$$

where $\approx$ denotes exact match for categorical slots and tolerance-based match for numerical or temporal slots. Higher SFA indicates better factual agreement.

Temporal Ordering Consistency (TOC). TOC evaluates whether the relative chronological structure of events in the generated narrative matches that of the reference. Both answers are converted into ordered event sequences, and consistency is measured over all comparable event pairs:

$$\mathrm{TOC} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \mathbb{I}\big(\mathrm{order}(e_i^*, e_j^*) = \mathrm{order}(\hat{e}_i, \hat{e}_j)\big). \tag{3}$$

Higher TOC indicates stronger temporal coherence.

Reference-Based Contradiction Rate (RCR). RCR measures the proportion of factual claims in the generated answer that contradict the reference narrative. Let $\mathcal{C}(\hat{a})$ denote extracted claims from $\hat{a}$. The contradiction rate is defined as

$$\mathrm{RCR} = \frac{1}{|\mathcal{C}(\hat{a})|} \sum_{c\in\mathcal{C}(\hat{a})} \mathbb{I}_{\mathrm{contr}}(c), \tag{4}$$

where lower RCR indicates better faithfulness.

Unlike standard text summarization, trajectory-grounded QA evaluates whether a model can faithfully recover structured behavioral facts (e.g., times, modes, counts, and ordering) from non-linguistic evidence rather than reproduce stylistic or lexical variation. Consequently, reference answers in *TravelBehaviorQA* function as fact-structured behavioral records rather than paraphrase targets, making SFA, TOC, and RCR necessary and complementary to surface-similarity metrics such as BLEU and BERTScore, which are insensitive to factual omissions and hallucinated mobility claims.

5 Experimental Results

5.1 Overall benchmark performance

Current LLMs demonstrate limited capability in understanding and interpreting GPS-based travel behavior, as reflected by their performance (Table 4) on both deterministic and open-ended tasks. Deterministic accuracy remains moderate across models, indicating that even low-entropy factual extraction from trajectory representations is not consistently reliable. This suggests that identifying and aggregating basic spatiotemporal attributes continues to pose challenges when inputs are serialized and presented in natural language form.

Table 4: Overall benchmark performance on *TravelBehaviorQA*.

	Det.	Open-ended		
Model	**Acc.↑**	**SFA↑**	**TOC↑**	**RCR↓**
gpt-5.2	0.58	**0.82**	0.96	0.15
gpt-5-mini	0.14	0.68	0.96	0.19
gpt-5-nano	0.14	0.62	0.94	0.22
claude-sonnet-4.5	**0.65**	0.80	0.94	0.20
qwen3-max	0.35	0.74	0.98	0.13
qwen-plus	0.42	0.76	0.98	**0.12**
qwen-flash	0.37	0.69	0.98	0.16
deepseek-chat	0.39	0.65	0.98	0.14
grok-4-1-fast	0.34	0.75	0.95	0.15
gemini-2.5-pro	0.18	0.57	**0.99**	0.14

For open-ended questions, behavior-aware metrics reveal clear and systematic limitations. Models generally achieve high TOC scores, indicating that coarse temporal structure in reference narratives is often preserved. In contrast, SFA is notably

Table 5: Model performance by question category on *TravelBehaviorQA*. Deterministic categories are evaluated using accuracy; open-ended categories are reported as SFA, TOC and RCR.

Deterministic Question Categories (Accuracy ↑)					
Model	**Travel Inten.**	**Daily Schedule**	**Activity Pattern**	**Place–Time Sem.**	**Mode Behavior**
gpt-5.2	**0.64**	0.46	0.63	**0.66**	0.59
gpt-5-mini	0.14	0.12	0.04	0.23	0.30
gpt-5-nano	0.15	0.11	0.16	0.16	0.22
claude-sonnet-4.5	0.58	**0.60**	**0.73**	**0.66**	**0.96**
qwen3-max	0.29	0.23	0.42	0.51	0.57
qwen-plus	0.32	0.34	0.45	0.64	0.68
qwen-flash	0.28	0.31	0.48	0.50	0.41
deepseek-chat	0.32	0.28	0.51	0.52	0.49
grok-4-1-fast	0.26	0.22	0.48	0.49	0.54
gemini-2.5-pro	0.10	0.10	0.26	0.39	0.27
Open-Ended Question Categories (SFA ↑ / TOC ↑ / RCR ↓)					
Model	**Daily Narrative**	**Trip Insights**	**Place Insights**	**Mode Insights**	**Weekly Narrative**
gpt-5.2	**0.86** / 0.99 / 0.09	0.86 / 0.95 / 0.17	0.79 / 0.91 / **0.19**	**0.79** / 0.94 / 0.13	**0.82** / 0.98 / 0.17
gpt-5-mini	0.77 / 0.99 / 0.11	0.84 / 0.95 / 0.20	0.76 / 0.89 / 0.22	0.76 / **1.00** / 0.13	0.80 / **1.00** / 0.18
gpt-5-nano	0.69 / 0.97 / 0.17	0.82 / 0.95 / 0.22	0.59 / 0.87 / 0.22	0.75 / **1.00** / 0.13	0.71 / 0.97 / 0.22
claude-sonnet-4.5	0.84 / 0.98 / 0.13	0.78 / 0.95 / 0.26	**0.81** / 0.90 / 0.20	0.77 / 0.94 / 0.20	0.81 / 0.94 / 0.23
qwen3-max	0.80 / 0.99 / 0.09	0.82 / 0.97 / 0.12	0.62 / 0.95 / 0.22	0.67 / 0.97 / **0.08**	0.81 / 0.99 / 0.12
qwen-plus	0.81 / 0.99 / 0.08	**0.87** / 0.98 / **0.10**	0.66 / 0.96 / 0.20	0.69 / 0.97 / **0.08**	0.81 / 0.99 / **0.11**
qwen-flash	0.70 / **1.00** / 0.12	0.84 / 0.99 / 0.14	0.50 / 0.94 / 0.27	0.67 / 0.96 / 0.13	0.80 / 0.99 / 0.14
deepseek-chat	0.66 / 0.99 / 0.11	0.67 / 0.97 / 0.18	0.64 / **0.97** / 0.21	0.51 / 0.98 / 0.09	0.80 / 0.99 / 0.13
grok-4-1-fast	0.79 / 0.99 / 0.11	0.76 / 0.95 / 0.18	0.70 / 0.90 / 0.20	0.70 / 0.93 / 0.12	**0.82** / 0.98 / 0.16
gemini-2.5-pro	0.71 / **1.00** / 0.10	0.63 / **0.99** / 0.13	0.35 / 0.96 / 0.28	0.49 / 0.97 / **0.08**	0.69 / 0.99 / 0.12

lower, and non-trivial RCR values persist across all models, reflecting frequent inconsistencies and factual mismatches in generated narratives. Together, these results show that while current models can maintain basic chronological coherence, they struggle to faithfully reproduce and integrate behavioral facts, highlighting a substantial gap between deterministic extraction and robust semantic behavioral reasoning.

5.2 Model scale effects

Model scaling exhibits heterogeneous effects across evaluation dimensions. Deterministic accuracy shows limited and non-monotonic gains with increasing model size, indicating early saturation for low-entropy factual reasoning. In contrast, open-ended behavioral reasoning benefits more consistently from scale. Larger models achieve higher slot-level factual agreement and improved faithfulness, while temporal ordering remains strong even at smaller scales, suggesting that coarse temporal structure is not the primary bottleneck. Despite these improvements, contradiction rates remain non-negligible at the largest scales, underscoring that increased capacity alone is insufficient to fully resolve semantic grounding challenges in trajectory understanding. Overall, the observed trends are consistent across both GPT and Qwen families, indicating that performance gains are primarily driven by scale rather than family-specific architectural differences (Figure 5).

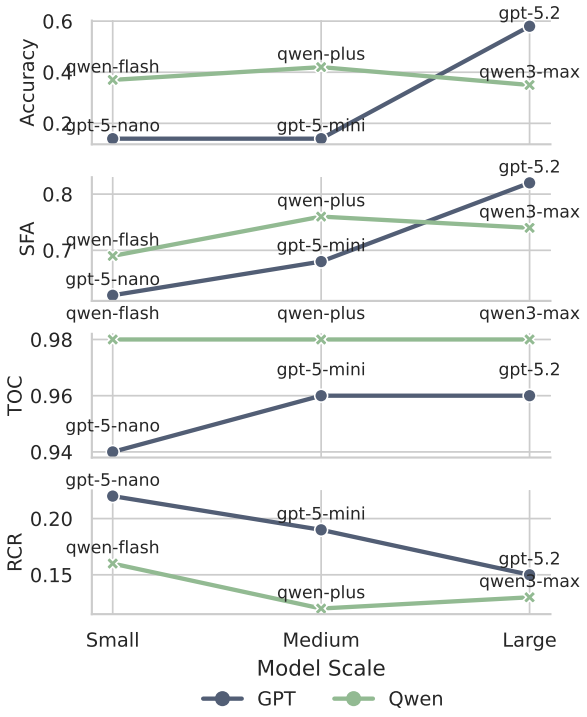

Figure 5: Performance comparison across LLM scales.

5.3 Impact of trajectory input length

The length of the trajectory input window has a non-monotonic effect on deterministic reasoning performance. Accuracy improves as temporal context expands from very short trajectories to moderately sized input windows, reflecting the benefit of additional contextual information for factual inference. However, performance declines slightly when the input window becomes excessively long, suggesting that redundant or distracting information can interfere with rule-based reasoning. This trade-off illustrates that both insufficient and excessive context can hinder effective trajectory understanding. A similar pattern has been reported in prior evaluations (Asano et al., 2025), suggesting that this consequence reflects a general characteristic of trajectory-based question answering rather than a dataset-specific artifact (Figure 6).

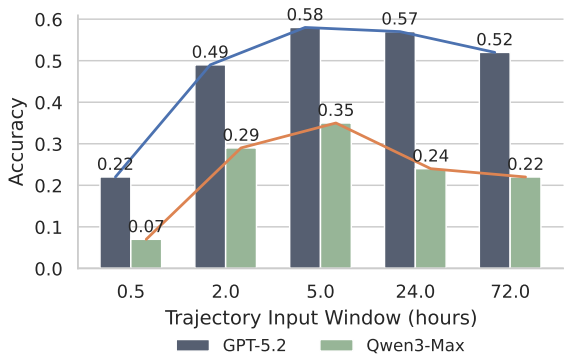

Figure 6: Impact of trajectory input length on model accuracy for deterministic questions.

5.4 Trajectory representation format

The representation of GPS trajectories provided to the language model has a significant impact on performance. Raw coordinate sequences lack semantic structure and are poorly suited for direct processing by language models, resulting in consistently weak reasoning and summarization outcomes. When trajectories are transformed into representations that more closely resemble natural language through temporal segmentation and POI grounding, performance improves substantially, reflecting better alignment with the model's inductive biases. Among these approaches, trip-level summaries achieve the strongest overall results. However, even segmented trajectories and structured trip logs offer only limited improvements for higher-level travel behavior summarization, showing that while semantically grounded representations are necessary, further abstraction and task-specific structuring are required to fully capture complex mobility patterns (Table 6).

Table 6: Effect of trajectory representation on GPT-5.2 performance.

	Det.	Open-ended		
Model	**Acc.↑**	**SFA↑**	**TOC↑**	**RCR↓**
Raw GPS only	0.05	0.82	0.91	0.22
Raw GPS + POI	0.18	0.68	0.94	0.19
Segmented + POI	0.58	0.82	0.96	0.15
Trip Narra. + POI	**0.61**	**0.83**	**0.96**	**0.15**

5.5 Effect of Decoding Temperature on Deterministic Evaluation

To assess the impact of stochastic decoding on deterministic question answering evaluation, we conducted a controlled decoding stability experiment on a deterministic subset ($n = 30{,}000$ per run), comparing deterministic decoding (`temperature` $= 0.0$) against stochastic decoding (`temperature` $= 1.0$) across three independent trials per configuration. Deterministic decoding consistently achieves higher mean accuracy under both exact and flexible matching criteria, with lower variance across runs (Table 7). Furthermore, deterministic decoding yields superior output reproducibility, reflected in higher all-same and pairwise agreement rates. These results confirm that stochastic decoding introduces avoidable randomness that undermines both accuracy and reproducibility in deterministic QA tasks.

Table 7: Decoding accuracy and reproducibility under deterministic vs. stochastic decoding (3 runs each).

Temp.	**Exact Acc**		**Flex. Acc**		**Reproducibility**	
	Mean	Std	Mean	Std	All-Same	Pairwise Agr.
0.0	**0.06**	0.02	**0.38**	0.04	**0.50**	**0.59**
1.0	0.03	0.00	0.31	0.02	0.47	0.54

5.6 Ablation Study

To assess whether upstream preprocessing errors propagate into QA labels and confound the distinction between reasoning failure and inherited noise, we conduct a controlled per-component ablation study, holding QA instances, prompts, model, decoding configuration, and evaluation metrics fixed while removing one preprocessing component at a time. As shown in Table 8, segmentation is the dominant structural dependency, and its removal

produces the largest performance drop across both evaluation metrics. Mode labels and POI information introduce moderate and minor degradations, respectively, each localized to semantically relevant question categories. Critically, the divergence between open-ended and deterministic degradation patterns across variants confirms that reasoning performance is not merely inherited from preprocessing artifacts, upstream components affect specific question categories in predictable, bounded ways, and the benchmark retains discriminative power throughout.

Table 8: Per-component ablation results. Δ denotes difference from the full pipeline.

Variant	Det. Acc.	ΔDet.	SFA	ΔSFA
Full	0.35	—	0.7528	—
No Mode	0.23	-0.12	0.6316	-0.1212
No POI	0.21	-0.14	0.7413	-0.0115
Fix Seg	0.10	-0.25	0.4115	-0.3413

6 Conclusion

This work reframes trajectory understanding as a language-centered behavioral reasoning problem rather than a traditional mobility analytics task. By introducing *TravelBehaviorQA*, we establish a benchmark that systematically links raw spatiotemporal traces with human-grounded behavioral interpretation, enabling principled evaluation of how language models reason over non-linguistic evidence. The unified integration of deterministic and open-ended questions exposes a fundamental gap between factual extraction and genuine behavioral abstraction, demonstrating that proficiency in low-entropy reasoning does not imply the ability to infer routines, intent, or coherent mobility narratives. In this sense, *TravelBehaviorQA* advances the evaluation of LLMs beyond prediction and classification, toward grounded reasoning, temporal abstraction, and faithful narrative generation over real-world behavioral signals.

More broadly, our findings indicate that progress in trajectory-based reasoning will not be driven by model scale alone, but by how spatiotemporal information is represented, abstracted, and aligned with linguistic inductive biases. The persistent limitations observed even under structured inputs suggest that trajectory understanding requires new representational and reasoning paradigms that bridge continuous mobility traces and symbolic behavioral descriptions. As a large-scale, publicly available, and behaviorally grounded benchmark, *TravelBehaviorQA* provides both a diagnostic lens and a foundation for future research on language-based interpretation of complex, long-horizon, non-textual data, with implications extending beyond mobility to a wide range of embodied and temporal reasoning domains.

Limitations

This work has several limitations. First, the benchmark evaluates state-of-the-art LLMs under prompting-based inference and does not examine fine-tuning or architectural modifications that may better integrate spatiotemporal information. Second, the trajectory representations used in the dataset are derived from a specific segmentation approach and a single source of place information, and alternative abstraction strategies may affect model performance. Finally, although *TravelBehaviorQA* covers diverse users and behaviors over multiple years, it is constructed from one underlying GPS dataset, and extending the benchmark to additional geographic contexts and mobility settings remains an important direction for future work.

Acknowledgements

This work was supported in part by the Federal Transit Administration (FTA) under Federal Award Identification Number MD-2023-002-00, Enhancing Mobility Innovation.

References

Timur Abbiasov, Cate Heine, Sadegh Sabouri, Arianna Salazar-Miranda, Paolo Santi, Edward Glaeser, and Carlo Ratti. 2024. The 15-minute city quantified using human mobility data. *Nature Human Behaviour*, 8(3):445–455.

Hikaru Asano, Hiroki Ouchi, Akira Kasuga, and Ryo Yonetani. 2025. Mobqa: A benchmark dataset for semantic understanding of human mobility data through question answering. *arXiv preprint arXiv:2508.11163*.

Derya Birant and Alp Kut. 2007. St-dbscan: An algorithm for clustering spatial–temporal data. *Data & knowledge engineering*, 60(1):208–221.

Yong Chen, Ben Chi, Chuanjia Li, Yuliang Zhang, Chenlei Liao, Xiqun Chen, and Na Xie. 2025. Toward interactive next location prediction driven by large language models. *IEEE Transactions on Computational Social Systems*.

Jiawei Cheng, Jingyuan Wang, Yichuan Zhang, Jiahao Ji, Yuanshao Zhu, Zhibo Zhang, and Xiangyu Zhao. 2025. Poi-enhancer: An llm-based semantic enhancement framework for poi representation learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(11):11509–11517.

Xiao Fu, Yi Zhang, Juan de Dios Ortúzar, and Guonian Lü. 2025. Activity-travel pattern inference based on multi-source big data. *Transport Reviews*, 45(1):26–48.

Shichao Ge, Peijun Ye, Renrui Zhang, Min Zhou, Hairong Dong, and Fei-Yue Wang. 2025. Llm-driven cognitive modeling for personalized travel generation. *IEEE Transactions on Computational Social Systems*.

Shreya Ghosh, Soumya K Ghosh, Sajal K Das, and Prasenjit Mitra. 2024. Mobilytics: Mobility analytics framework for transferring semantic knowledge. *IEEE Transactions on Mobile Computing*, 23(12):11588–11603.

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. *Advances in Neural Information Processing Systems*, 37:36185–36217.

Xiao Han, Dayan Pan, Xiangyu Zhao, Xuyuan Hu, Zhaolin Deng, Xiangjie Kong, and Guojiang Shen. 2025. A dataset for spatiotemporal-sensitive poi question answering. *arXiv preprint arXiv:2505.10928*.

Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Srinivas Sunkara, Victor Cărbune, Jason Lin, Maria Wang, Yun Zhu, and Jindong Chen. 2025. Screenqa: Large-scale question-answer pairs over mobile app screenshots. 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 9427–9452.

Tianhao Huang, Xuan Pan, Xiangrui Cai, Ying Zhang, and Xiaojie Yuan. 2024. Learning time slot preferences via mobility tree for next poi recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8):8535–8543.

Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming Jin, and Qingsong Wen. 2025. Time-mqa: Time series multi-task question answering with context enhancement. *arXiv preprint arXiv:2503.01875*.

Shengjie Lai, Alessandro Sorichetta, Jessica Steele, Corrine W Ruktanonchai, Alexander D Cunningham, Grant Rogers, Patrycja Koper, Dorothea Woods, Maksym Bondarenko, Nick W Ruktanonchai, and 1 others. 2022. Global holiday datasets for understanding seasonal human mobility and population dynamics. *Scientific Data*, 9(1):17.

Zhengxing Lan, Lingshan Liu, Bo Fan, Yisheng Lv, Yilong Ren, and Zhiyong Cui. 2024. Traj-llm: A new exploration for empowering trajectory prediction with pre-trained large language models. *IEEE Transactions on Intelligent Vehicles*.

Xuchuan Li, Fei Huang, Jianrong Lv, Zhixiong Xiao, Guolong Li, and Yang Yue. 2024. Be more real: Travel diary generation using llm agents and individual profiles. *arXiv preprint arXiv:2407.18932*.

Chenxi Liu, Zhu Xiao, Wangchen Long, Tong Li, Hongbo Jiang, and Keqin Li. 2025a. Vehicle trajectory data processing, analytics, and applications: A survey. *ACM Computing Surveys*, 57(9):1–36.

Tianming Liu, Jirong Yang, and Yafeng Yin. 2025b. Toward llm-agent-based modeling of transportation systems: A conceptual framework. *Artificial Intelligence for Transportation*, 1:100001.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. 2025c. Datasets for large language models: A comprehensive survey. *Artificial Intelligence Review*, 58(12):403.

Joseph Molloy, Alberto Castro, Thomas Götschi, Beaumont Schoeman, Christopher Tchervenkov, Uros Tomic, Beat Hintermann, and Kay W Axhausen. 2023. The mobis dataset: a large gps dataset of mobility behaviour in switzerland. *Transportation*, 50(5):1983–2007.

Hamed Nilforoshan, Wenli Looi, Emma Pierson, Blanca Villanueva, Nic Fishman, Yiling Chen, John Sholar, Beth Redbird, David Grusky, and Jure Leskovec. 2023. Human mobility networks reveal increased segregation in large cities. *Nature*, 624(7992):586–592.

Guoying Qiu, Guoming Tang, Chuandong Li, Deke Guo, Yulong Shen, and Yan Gan. 2023. Behavioral-semantic privacy protection for continual social mobility in mobile-internet services. *IEEE Internet of Things Journal*, 11(1):462–477.

Benjamin Reichman, Xiaofan Yu, Lanxiang Hu, Jack Truxal, Atishay Jain, Rushil Chandrupatla, Tajana S Rosing, and Larry Heck. 2025. Sensorqa: A question answering benchmark for daily-life monitoring. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, pages 282–289.

Zijian Shao, Jiancan Wu, Weijian Chen, and Xiang Wang. 2025. Personal travel solver: A preference-driven LLM-solver system for travel planning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27622–27642, Vienna, Austria. Association for Computational Linguistics.

Xuan Shen, Hangyu Zheng, Yifan Gong, Zhenglun Kong, Changdi Yang, Zheng Zhan, Yushu Wu, Xue Lin, Yanzhi Wang, Pu Zhao, and 1 others. 2025. Sparse learning for state space models on mobile. In *The Thirteenth International Conference on Learning Representations*.

Chris Stanford, Suman Adari, Xishun Liao, Yueshuai He, Qinhua Jiang, Chenchen Kuai, Jiaqi Ma, Emmanuel Tung, Yinlong Qian, Lingyi Zhao, and 1 others. 2024. Numosim: A synthetic mobility dataset with anomaly detection benchmarks. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Geospatial Anomaly Detection*, pages 68–78.

Jiaan Wang, Fandong Meng, Zengkui Sun, Yunlong Liang, Yuxuan Cao, Jiarong Xu, Haoxiang Shi, and Jie Zhou. 2025. An empirical study of many-to-many summarization with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11344, Vienna, Austria. Association for Computational Linguistics.

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. *Advances in Neural Information Processing Systems*, 37:124547–124574.

Xinglei Wang, Meng Fang, Zichao Zeng, and Tao Cheng. 2023. Where would i go next? large language models as human mobility predictors. *arXiv preprint arXiv:2308.15197*.

Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, and Yueming Jin. 2025. Agentic reasoning: A streamlined framework for enhancing LLM reasoning with agentic tools. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28489–28503, Vienna, Austria. Association for Computational Linguistics.

Zhiwen Xiao and Huagang Tong. 2025. Federated contrastive learning with feature-based distillation for human activity recognition. *IEEE Transactions on Computational Social Systems*.

Zhu Xiao, Linshan Wu, Hongbo Jiang, Zheng Qin, Chengxi Gao, You Li, Hongyang Chen, and Jiangchuan Liu. 2023. Exploring intercity mobility in urban agglomeration: Evidence from private car trajectory data. *IEEE Transactions on Computational Social Systems*, 11(2):2940–2954.

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.

Kairui Yang, Zihao Guo, Gengjie Lin, Haotian Dong, Zhao Huang, Yipeng Wu, Die Zuo, Jibin Peng, Ziyuan Zhong, Xin WANG, Qing Guo, Xiaosong Jia, Junchi Yan, and Di Lin. 2025a. Trajectory-LLM: A language-based data generator for trajectory prediction in autonomous driving. In *The Thirteenth International Conference on Learning Representations*.

Sean Bin Yang, Ying Sun, Yunyao Cheng, Yan Lin, Kristian Torp, and Jilin Hu. 2025b. Spatio-temporal trajectory foundation model-recent advances and future directions. *arXiv preprint arXiv:2511.20729*.

Zhiwei Yin, Bin Jia, Xiao-Yong Yan, Yitao Yang, Hao Ji, and Ziyou Gao. 2026. Classification of the freight trip purpose of heavy trucks using trajectory data and waybill data. *Transportation Research Part E: Logistics and Transportation Review*, 206:104584.

Ayda Zaroujtaghi, Omid Mansourihanis, Mohammad Tayarani, Fatemeh Mansouri, Moein Hemmati, and Ali Soltani. 2025. A systematic review of gis evolution in transportation planning: Towards ai integration. *Future Transportation*, 5(3):97.

Meijing Zhang and Ying Xu. 2025. Transmode-llm: Feature-informed natural language modeling with domain-enhanced prompting for travel behavior modeling. In *Large Language Models for Scientific and Societal Advances*.

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.

Yuanshao Zhu, Yongchao Ye, Ying Wu, Xiangyu Zhao, and James Yu. 2023. Synmob: Creating high-fidelity synthetic gps trajectory dataset for urban mobility analysis. *Advances in Neural Information Processing Systems*, 36:22961–22977.

A Dataset Construction Details

A.1 Preprocessing Pipeline

Prior to annotation and downstream modeling, raw GPS logs are processed through a standardized preprocessing pipeline designed to reduce sensor noise, normalize spatiotemporal resolution, and compute kinematic descriptors of movement. This preprocessing stage is intentionally limited to signal-level transformation and feature extraction; no trip or stop segmentation is performed in order to avoid introducing behavioral assumptions at this stage.

Each GPS record is represented as a tuple $p_i = (\ell_i, h_i, t_i)$, where $\ell_i = (\text{lat}_i, \text{lon}_i)$ denotes geographic coordinates, h_i denotes elevation, and t_i denotes the timestamp. Location measurements are first smoothed using a Kalman filter to mitigate high-frequency GPS noise. Coordinates are then rounded to a fixed spatial precision, and timestamps are discretized to minute-level granularity to ensure consistency across users and devices. Trajectory continuity is preserved except in cases of implausible spatial jumps or excessive temporal gaps, which are treated as discontinuities.

From the cleaned trajectories, we derive a set of kinematic features based on consecutive GPS observations. Let $\Delta t_i = t_i - t_{i-1}$ denote the elapsed time between observations. The following quantities are computed:

Displacement. The spatial displacement between consecutive points is computed using the haversine distance,

$$d_i = d(\ell_{i-1}, \ell_i), \tag{5}$$

where $d(\cdot)$ denotes the great-circle distance.

Speed. Instantaneous speed is defined as

$$v_i = \frac{d_i}{\Delta t_i}. \tag{6}$$

Acceleration. Instantaneous acceleration is computed as the first temporal derivative of speed,

$$a_i = \frac{v_i - v_{i-1}}{\Delta t_i}. \tag{7}$$

Elevation change. Vertical movement is quantified by the elevation difference,

$$\Delta h_i = h_i - h_{i-1}. \tag{8}$$

Bearing and bearing rate. Movement bearing is computed as

$$\theta_i = \text{bearing}(\ell_{i-1}, \ell_i), \tag{9}$$

and the bearing rate, which captures directional variability, is defined as

$$\omega_i = \frac{\theta_i - \theta_{i-1}}{\Delta t_i}. \tag{10}$$

Based on these point-level quantities, we compute trajectory-level summary statistics, including mean and median speed, acceleration variance, maximum jerk, cumulative elevation change, and bearing-rate variability. These kinematic descriptors provide a physically grounded representation of movement dynamics and are used as inputs for subsequent transportation mode inference and behavioral analysis.

A.2 Additional details on trip segmentation

In our implementation, we set $\varepsilon_{\text{space}} = 100$ m, $\varepsilon_{\text{time}} = 300$ s, and minPts $= 5$, and later examine the robustness of these choices through sensitivity analysis. Algorithm 1 summarizes the ST-DBSCAN-based trip segmentation procedure.

A.3 Additional details on travel mode prediction

Each *move* segment is classified into a transportation mode using a supervised gradient-boosted decision tree model implemented with XGBoost. For each segment, we input a set of aggregate kinematic and geometric features, including total travel distance, segment duration, mean speed, median acceleration, maximum jerk, number of GPS points, and bearing-rate variability. Ground-truth transportation mode labels provided in Geolife are aligned with segmented trajectories using overlapping temporal windows to accommodate partial mode transitions.

The gradient-boosting model is trained with 200 trees, a maximum tree depth of 5, a learning rate of 0.1, and a subsample ratio of 0.8, using a multiclass logistic loss objective. Early stopping is applied with a 20% validation split to mitigate overfitting. For benchmarking purposes, we additionally train support vector machines with an RBF kernel and random forest classifiers under comparable experimental settings. Table 9 reports accuracy, precision, recall, and F1-score across transportation modes.

Algorithm 1 Trip Segmentation with ST-DBSCAN

Require: Preprocessed trajectory $\mathcal{T} = \{(\ell_i, t_i)\}$, $\varepsilon_{\text{space}}$, $\varepsilon_{\text{time}}$, minPts
Ensure: Stop $\mathcal{S}_{\text{stop}}$ and move $\mathcal{S}_{\text{move}}$

```
Mark all points in T as unvisited
for each unvisited point p_i ∈ T do
    Mark p_i as visited
    Compute N(p_i)
    if |N(p_i)| < minPts then
        Mark p_i as noise
    else
        Initialize new cluster C
        Add p_i and N(p_i) to C
        for each point p_j ∈ C do
            if p_j is unvisited then
                Mark p_j as visited
                Compute N(p_j)
                if |N(p_j)| ≥ minPts then
                    Add N(p_j) to C
                end if
            end if
        end for
        Add C to S_stop
    end if
end for
Group points between consecutive stop clusters into move segments S_move
return S_stop, S_move
```

Table 9: Comparison of transportation mode detection models on the GeoLife dataset.

Model	Accuracy	Precision	Recall	F1-score
Gradient Boosting	**0.89**	**0.88**	**0.87**	**0.88**
Random Forest	0.85	0.84	0.83	0.83
SVM (RBF)	0.82	0.80	0.81	0.81

Gradient boosting achieves the highest and most stable performance across all evaluation metrics. Remaining classification errors primarily occur between bus and car modes, reflecting overlapping speed and acceleration characteristics in urban environments. In data-scarce scenarios, we further employ simple, interpretable heuristics as a fallback: segments with mean speed below 2 m/s and low acceleration variance are classified as walking, segments with speeds between 2 m/s and 6 m/s as cycling, and higher-speed segments as motorized travel.

B Additional Data Description

B.1 Transportation mode

The modal composition of the dataset reflects a diverse range of travel behaviors, with clear dominance of everyday surface transportation modes (Table 10). Motorized ground travel accounts for the majority of observed activity, indicating that the trajectories primarily capture routine urban mobility rather than sporadic long-distance travel. At the same time, substantial representation of active modes highlights the presence of fine-grained, short-range movements that are critical for understanding daily activity patterns and multimodal decision-making. The inclusion of multiple public transit modes further enriches the dataset by supporting reasoning over transfers, mode choice, and mixed-mode trips. This modal diversity ensures that the dataset supports behavioral inference across heterogeneous mobility contexts rather than being narrowly concentrated on a single travel mode.

Table 10: Distribution of trajectory segments and total duration by travel mode.

Mode	Segments	Total Duration (hours)
Airplane	45	130.22
Bike	2,867	4,463.04
Boat	8	3.57
Bus	6,727	5,674.26
Car	3,953	10,836.81
Motorcycle	2	0.16
Run	2	1.40
Subway	1,013	1,097.43
Taxi	490	575.05
Train	374	868.51
Walk	3,250	3,724.99

B.2 POI category

The POI distribution indicates that the dataset is strongly anchored in routine, everyday urban activities, with the majority of observations corresponding to food-related establishments, retail locations, and service-oriented destinations. This concentration reflects realistic daily mobility patterns driven by dining, shopping, and personal services, rather than sporadic or event-based travel. At the same time, the presence of healthcare facilities, educational institutions, cultural venues, parks, and transportation infrastructure demonstrates that the dataset captures a broad spectrum of functional urban spaces beyond consumption-oriented activities. The resulting POI composition balances frequency and diversity, ensuring that trajectory semantics are

grounded in common daily behaviors while still supporting reasoning over less frequent but behaviorally meaningful destinations (Table 11).

Table 11: POI categories count.

POI Category	Count
Fast Food & Quick Service Restaurants	31,775
Full-Service Restaurants (All Cuisines)	24,392
Coffee Shops & Cafés	9,748
Retail Stores (Big Box, Grocery, Convenience)	11,279
Shopping & Commercial Facilities	4,612
Healthcare Facilities (Hospitals, Clinics)	7,782
Education & Training Institutions	1,311
Cultural & Entertainment Venues	4,209
Parks, Scenic, & Outdoor Locations	2,252
Transportation & Infrastructure	3,696
Residential & Neighborhood Areas	536
Financial & Public Services	1,738
Sports & Recreation Facilities	1,121
Other / Miscellaneous POIs	3,045

B.3 Geographic coverage

The dataset provides geographically diverse mobility traces that span multiple metropolitan contexts and travel environments. At the global scale, the spatial footprint covers a wide set of cities distributed across several regions, which reduces the risk that model evaluation is driven by a single local mobility grammar. This breadth exposes models to heterogeneous spatial layouts and travel constraints, including differences in block structure, arterial networks, and the availability of public transport and mixed-use destinations. Thus, the benchmark reflects real-world variation in how daily mobility is organized, supporting generalization-oriented evaluation rather than place-specific memorization (Figure 7).

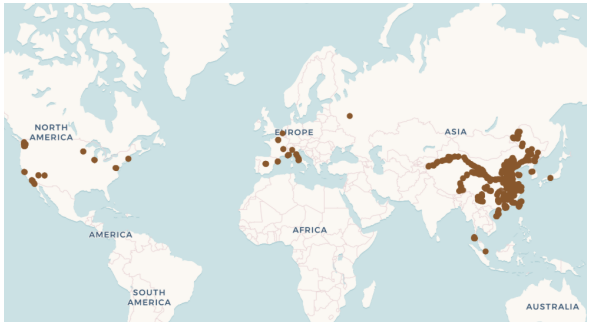

Figure 7: Global distribution of trajectory points across observed cities.

At the same time, the dataset contains a dense, high-resolution core region that enables fine-grained behavioral analysis at the street and neighborhood level. The Beijing visualization shows highly structured trace density that aligns with a realistic urban mobility backbone: concentrated activity clusters in central districts and corridor-like patterns along major connectors. This combination of dense urban cores and lower-intensity peripheral coverage is valuable for trajectory understanding tasks, because it forces models to operate across different spatial regimes (dense downtown grids versus sparser edges) while still retaining enough observations to support robust analysis within a primary city (Figure 8). Together, these properties yield a benchmark with both broad geographic diversity and a sufficiently dense anchor region to support detailed spatial-semantic interpretation, including comparisons across urban form, neighborhood contexts, and city-level mobility structure.

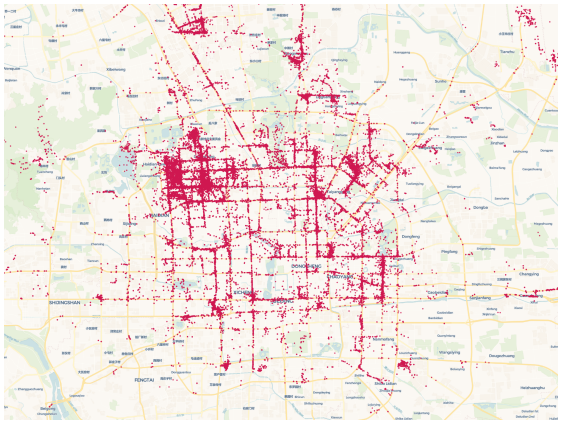

Figure 8: Spatial density of trajectory points within Beijing.

B.4 Temporal Coverage and Longitudinal Depth

TravelBehaviorQA spans multiple years of observed mobility (Figure 9), enabling evaluation across diverse temporal contexts rather than isolated snapshots of travel behavior. Core behavioral attributes remain consistently represented over time, while narrative questions are distributed throughout the dataset, ensuring that interpretive tasks reflect persistent and recurring mobility patterns. This longitudinal structure supports assessment of models' ability to generalize behavioral understanding across time, rather than memorizing period-specific trajectories

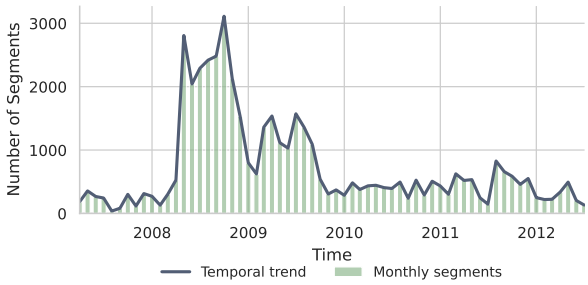

Figure 9: Dataset time spanning

B.5 User Coverage

TravelBehaviorQA is designed to ensure balanced user representation across behavioral categories. Each category includes questions contributed by the full set of users, and individual users are associated with multiple task types, including both deterministic and open-ended questions. This design prevents concentration of data in a small subset of users and mitigates user-specific bias.

Table 12: User-level coverage statistics.

Metric	Min	Mean	Max
QA instances	1,041	2,106	28,968
Days observed	46	93	1,273
Categories	11	11	11
Question types	30	30	30
Open question share (%)	21.9	24.9	27.3

Moreover, narrative and interpretive questions are broadly distributed rather than limited to highly active individuals, ensuring that higher-level reasoning tasks reflect general travel behavior. Overall, the dataset supports robust evaluation of cross-user generalization rather than memorization of individual mobility patterns.

C Additional Performance Results

Open ended questions consistently yield low BLEU and BERTScore values across all evaluated models, indicating that narrative level mobility reasoning remains challenging for current LLMs. While models can often extract isolated factual attributes, generating coherent behavioral explanations that integrate temporal structure, activities, and travel modes is substantially more difficult. This gap highlights the complexity of translating spatiotemporal traces into meaningful behavioral narratives rather than answering low entropy factual queries.

Table 13: Performance of LLMs on open-ended questions in TravelBehaviorQA. BLEU and BERTScore are reported, with higher values indicating better semantic alignment with reference answers.

Model	**BLEU ↑**	**BERTScore ↑**
gpt-5.2	9.73	0.39
gpt-5-mini	7.43	0.38
gpt-5-nano	6.62	0.33
claude-sonnet-4.5	9.86	0.40
qwen3-max	18.11	0.51
qwen-plus	20.45	0.54
qwen-flash	15.43	0.49
deepseek-chat	11.57	0.41
grok-4-1-fast	11.03	0.42
gemini-2.5-pro	10.02	0.38

Increasing model capacity provides only modest benefits for trajectory understanding. Larger models achieve incremental improvements in open ended generation quality, but they continue to struggle with organizing spatiotemporal information into coherent representations of activities, intent, and routine. Consequently, improvements in deterministic performance do not reliably translate into stronger open ended reasoning, suggesting that factual extraction and behavioral interpretation rely on fundamentally different capabilities.

Performance also varies across behavioral categories. Aggregate and structurally constrained questions are relatively easier, whereas tasks requiring temporal semantics, mode interpretation, and higher level behavioral insights remain substantially more difficult. Open ended narrative questions exhibit the greatest inter model variance, reflecting differences in models' ability to abstract salient patterns and maintain temporal coherence over extended contexts.

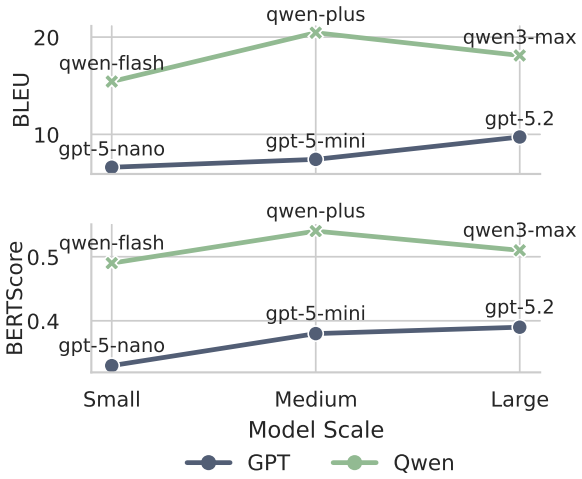

Figure 10: Performance comparison across model families and scales.

Table 14: Category-level performance on open-ended questions in *TravelBehaviorQA*. Each cell reports **BERTScore / BLEU**, where higher values indicate better semantic alignment and narrative quality.

Model	Daily Narrative	Trip Insights	Place Insights	Mode Insights	Weekly Narrative
gpt-5.2	0.57 / 21.41	0.37 / 6.26	0.32 / 5.20	0.39 / 12.12	0.36 / 5.21
gpt-5-mini	0.46 / 10.19	0.04 / 0.00	0.25 / 1.96	0.30 / 5.69	0.24 / 1.63
gpt-5-nano	0.33 / 9.60	0.26 / 4.40	0.09 / 1.30	0.27 / 3.80	0.18 / 0.94
claude-sonnet-4.5	0.47 / 14.56	0.39 / 9.07	0.37 / 8.44	0.38 / 9.69	0.40 / 7.74
qwen3-max	0.59 / 28.85	0.54 / 18.31	0.41 / 12.38	0.47 / 13.04	0.54 / 20.14
qwen-plus	0.62 / 31.20	0.60 / 22.95	0.45 / 12.61	0.48 / 13.41	0.59 / 24.77
qwen-flash	0.57 / 21.55	0.54 / 18.85	0.37 / 7.57	0.46 / 12.09	0.54 / 19.06
deepseek-chat	0.46 / 16.91	0.42 / 11.39	0.36 / 9.59	0.38 / 7.11	0.46 / 13.89
grok-4-1-fast	0.51 / 16.60	0.45 / 10.47	0.35 / 8.62	0.39 / 9.73	0.44 / 10.73
gemini-2.5-pro	0.46 / 17.70	0.39 / 9.71	0.30 / 4.23	0.34 / 4.83	0.45 / 15.19

D LLM Settings and Experimental Configuration

Decoding parameters are held constant across experiments. We use a temperature of 1.0 and a top-p value of 0.95 to balance response diversity and stability, reflecting standard practice for open-ended generation while avoiding overly deterministic outputs. For deterministic questions, responses are constrained to short factual outputs where applicable, whereas open-ended questions allow free-form generation without length truncation beyond model-imposed limits.

Table 15: LLMs and decoding parameters.

Model	Temp.	Top-p	Thinking
GPT-5.2	1.0	0.95	Minimal
GPT-5-Mini	1.0	0.95	Minimal
GPT-5-Nano	1.0	0.95	Minimal
Gemini-2.5-Pro	1.0	0.95	Low
grok-4-1-fast	1.0	0.95	Low
Qwen3-Max	1.0	0.95	Low
Qwen-Plus	1.0	0.95	Low
Qwen-Flash	1.0	0.95	Low
Claude-Sonnet-4.5	1.0	0.95	Low
DeepSeek-chat	1.0	0.90	Low

To assess robustness and generalization across model families and scales, we evaluate the benchmark on a diverse set of state-of-the-art LLMs, including OpenAI models (GPT-5.2, GPT-5-mini, GPT-5-nano), Anthropic Claude (Sonnet 4.5), Qwen models (Qwen3-Max, Qwen-Plus), DeepSeek (DeepSeek-V3.2), Grok models (Grok-4-1 Fast), and Gemini (Gemini-2.5-Pro). All models are tested under consistent prompting and decoding settings, enabling systematic comparison of trajectory understanding performance across architectures, providers, and model capacities. No chain-of-thought prompting or external tools are used. Each model is queried once per QA instance, and no response selection or reranking is applied. Evaluation is conducted separately for deterministic and open-ended questions using task-appropriate metrics defined in the main text. All experiments are executed using the same dataset splits and input representations to ensure consistency across model families and scales.

E Examples of *TravelBehaviorQA*

The *TravelBehaviorQA* benchmark distinguishes between deterministic questions, which require low-entropy, rule-based inference grounded in observable trajectory attributes, and open-ended questions, which demand higher-level abstraction, synthesis, and behavioral interpretation. This separation enables controlled analysis of factual grounding versus semantic understanding within a unified behavioral framework.

E.1 Deterministic Questions

Deterministic questions target well-defined behavioral attributes that admit a single correct answer derived directly from trajectory data. These questions evaluate factual grounding, numerical reasoning, and temporal–spatial consistency.

Travel Intensity. This category quantifies the overall magnitude of travel within a given period, including total distance, total travel time, and number of trips. It evaluates whether models can accurately aggregate movement statistics from raw trajectories.

Example of Travel Intensity Question

Question:
For user_id 010, what is the distribution of trip lengths on 2007-12-31?

Retrieved trajectory with POI:
[1] MOVE 06:09-16:47 mode=car
[2] STOP 16:48-16:53 poi=Autumn Moon over the Calm Lake(Scenic Lookout)
[3] MOVE 16:54-17:10 mode=bus

Answer:
short 0, medium 1, long 1

Daily Schedule. Daily Schedule questions characterize the temporal organization of mobility, such as first departure, last arrival, and major travel periods, emphasizing precise time identification.

Example of Daily Schedule Question

Question:
What is user 010's preferred time of day for activities on 2008-10-18?

Retrieved trajectory with POI:
[1] MOVE 12:06–12:10 mode=taxi
[2] STOP 12:22–12:24 poi=Century Clock (Monument)
[3] STOP 12:44–12:50 poi=Century Clock (Monument)
[4] MOVE 13:00–13:06 mode=taxi

Answer:
Daytime

Activity Patterns. This category focuses on stop behavior and activity structure, including stop durations, longest stops, and sequencing of activities throughout the day.

Example of Activity Patterns Question

Question:
What was user_id 010's dominant activity with the longest stay on 2007-10-20?

Retrieved trajectory with POI:
[1] MOVE 03:00–11:02 mode=car
[2] STOP 11:03–11:08 poi=WenTaoFang Ancient Weapon Museum (History Museum)
[3] STOP 11:08–11:23 poi=WenTaoFang Ancient Weapon Museum (History Museum)
[4] MOVE 11:25–11:42 mode=bike

Answer:
Visiting WenTaoFang Ancient Weapon Museum, with the longest stop duration of approximately 15 minutes (11:08–11:23).

Place-Time Semantics. Place–Time Semantics link spatial locations with temporal constraints, requiring models to identify where an individual was during a specified time interval.

Example of Place–Time Semantics Question

Question:
What place was user 010 at on 2008-10-01 at 05:50:47?

Retrieved trajectory with POI:
[1] MOVE 00:00–05:40 mode=train
[2] STOP 05:40–06:00 poi=Pasaraya Speed 99 (Board Store)
[3] MOVE 06:07–06:19 mode=taxi
[4] STOP 06:20–06:28 poi=Pasaraya Speed 99 (Board Store)

Answer:
Pasaraya Speed 99

Mode Behavior. Mode Behavior questions evaluate factual reasoning about inferred transportation modes, such as counts or dominance of specific modes.

Example of Mode Behavior Question

Question:
How many distinct MOVE modes did user 010 use during the week of 2008-W38 (2008-09-15 to 2008-09-21)?

Retrieved trajectory with POI (excerpt):
[1] MOVE 2008-09-15 23:17–23:59 mode=train
[2] MOVE 2008-09-16 00:00–00:01 mode=bus
[3] MOVE 2008-09-16 00:32–00:32 mode=walk
[4] MOVE 2008-09-16 00:59–01:06 mode=subway
[5] MOVE 2008-09-16 14:56–23:52 mode=taxi
...

Answer:
5

E.2 Open-Ended Questions

Open-ended questions require higher-entropy reasoning, asking models to synthesize, abstract, and interpret behavioral patterns across multiple trajectory elements. These questions evaluate narrative coherence and semantic understanding.

Daily Narratives. Daily Narratives prompt models to produce coherent summaries of a user's daily mobility, integrating trips, activities, and temporal structure.

Example of Daily Narratives Question

Question:
What did user 020 do on 2011-10-06 based on their daily trajectory?

Retrieved trajectory with POI:
[1] MOVE 04:06–04:11 mode=bike
[2] STOP 04:11–04:16 poi=Costa Coffee (Coffee Shop); Food Court; Najia Bistro (Beijing Restaurant)
[3] MOVE 04:47–04:49 mode=bike
[4] STOP 04:49–04:51 poi=Xi He Ya Yuan (Chinese Restaurant); SUBWAY (Sandwich Spot); Starbucks (Coffee Shop)
[5] MOVE 15:24–15:49 mode=bike

Answer:
On 2011-10-06, user 020 began the day early by biking shortly after 04:00 and stopping briefly at a coffee shop. They made another short bike trip later in the morning with a brief stop at a fast-food restaurant. In the afternoon, the user undertook an additional bike trip. No further activities are observed in the available trajectory data.

Trip Insights. Trip Insights focus on interpretive assessment of travel burden and structure, requiring justification grounded in observed mobility patterns.

Example of Trip Insights Question

Question:
What was the trip structure and intensity for user_id 010 on 2008-09-22 based on their daily trajectory?

Retrieved trajectory with POI:
[1] MOVE 00:00–13:05 mode=subway start=39.947721,116.346156 end=39.136204,117.204953
[2] STOP 13:05–13:07 poi=McDonald's (Fast Food Restaurant); Starbucks (Coffee Shop)
[3] STOP 13:48–13:59 poi=McDonald's (Fast Food Restaurant); Pizza Hut (Pizzeria)
[4] MOVE 23:26–23:53 mode=train start=39.142243,117.198226 end=39.865523,116.376616

Answer:
On 2008-09-22, user_id 010 exhibited a highly travel-intensive day characterized by two long-distance movement segments and only brief intermediate stops. The day began with an extended subway trip lasting over 13 hours, followed by two short food-related stops in the early afternoon. The day concluded with a late-night train journey. Overall, the trip structure reflects high mobility with minimal dwell time, indicating high trip intensity and limited engagement at destinations.

Place Insights. This category evaluates semantic interpretation of visited locations, asking models to abstract dominant place types or usage patterns.

Example of Place Insights Question

Question:
What places did user 010 visit and what activities can be inferred from their trajectory on 2008-10-13?

Retrieved trajectory with POI:
[1] MOVE 00:00–00:04 mode=taxi
[2] STOP 00:06–00:09 poi=New Century Grand Hotel Beijing (Hotel Bar); BBQ Joint; Fayuan Temple (Temple/Monument)
[3] MOVE 00:11–00:52 mode=taxi
[4] STOP 00:55–00:58 poi=Tonkatsu & Curry (Japanese Restaurant); KFC (Fast Food Restaurant); Pizza Hut (Pizzeria)
[5] MOVE 10:44–23:55 mode=subway start=39.981688,116.332566 end=39.863580,116.374368

Answer:
On 2008-10-13, user 010 visited several distinct place types. The early trajectory indicates a brief stop near Fayuan Temple, suggesting a short visit to a cultural or religious site. This was followed by travel to a food-related location, where the user stopped briefly at a restaurant, indicating a likely meal. Later in the day, the user undertook a prolonged subway trip spanning most of the afternoon and evening, suggesting extended travel rather than prolonged engagement at a specific destination. Overall, the trajectory reflects short, activity-focused stops embedded within longer-distance travel.

Mode Insights. Mode Insights require summarizing overall transportation preferences or reliance across the day, beyond simple counting.

Example of Mode Insights Question

Question:
What can be inferred about user_id 010's mode usage and switching patterns on 2008-04-04 based on their full day trajectory?

Retrieved trajectory with POI (excerpt):
[1] MOVE 00:00–04:13 mode=train
[2] STOP 04:27–04:31 poi=miskai (Outdoor Area)
[3] MOVE 04:31–10:19 mode=taxi
[4] STOP 10:19–10:26 poi=Liquor Store; Chinese Restaurant
...

Answer:
On 2008-04-04, user_id 010 relied predominantly on train travel, using it for nearly all long-distance movement throughout the day. A single extended taxi segment occurred in the early morning, following an initial train trip, suggesting a temporary mode switch likely due to connectivity or access constraints. After this taxi segment, the user consistently returned to train travel for the remainder of the day, with only brief stops between consecutive train journeys. Overall, the trajectory indicates a strong preference for rail-based transportation, with limited and situational use of alternative modes.

Routine Narratives. Routine Narratives assess whether recurring temporal or spatial patterns are evident, emphasizing cautious interpretation under uncertainty.

Example of Routine Narratives Question

Question:
What was the daily schedule and rhythm like for user_id 010 on 2009-01-14 based on their trajectory?

Retrieved trajectory with POI:
[1] MOVE 04:58–06:11 mode=bus start=40.099303,116.364085 end=40.055958,116.320455
[2] STOP 06:12–06:14 poi=Pizza Hut (Pizzeria); KFC (Fast Food Restaurant); (Structure)
[3] STOP 06:15–06:18 poi=Pizza Hut (Pizzeria); KFC (Fast Food Restaurant); (Structure)
[4] MOVE 06:18–06:44 mode=car start=40.051056,116.324221 end=39.992488,116.325473
...

Answer:
On 2009-01-14, user_id 010 followed an early-morning travel rhythm characterized by continuous movement and brief intermediate stops. The day began before 05:00 with a long bus trip, followed by two short stops at food service locations in quick succession, and concluded with a car journey shortly after 06:00. Overall, the trajectory reflects a tightly scheduled morning with minimal dwell time and a clear sequence of travel and short pauses.