Large Language Models for Spatial Trajectory Patterns Mining

Zheng Zhang, Hossein Amiri, Zhenke Liu, Andreas Züfle, Liang Zhao
Department of Computer Science
Emory University
Atlanta, GA 30322, USA
{zheng.zhang,hossein.amiri,zhenke.liu,azufle,liang.zhao}@emory.edu

Abstract

Identifying anomalous human spatial trajectory patterns can indicate dynamic changes in mobility behavior with applications in domains like infectious disease monitoring and elderly care. Recent advancements in large language models (LLMs) have demonstrated their ability to reason in a manner akin to humans. This presents significant potential for analyzing temporal patterns in human mobility. In this paper, we conduct empirical studies to assess the capabilities of leading LLMs like GPT-4 and Claude-2 in detecting anomalous behaviors from mobility data, by comparing to specialized methods. Our key findings demonstrate that LLMs can attain reasonable anomaly detection performance even without any specific cues. In addition, providing contextual clues about potential irregularities could further enhances their prediction efficacy. Moreover, LLMs can provide reasonable explanations for their judgments, thereby improving transparency. Our work provides insights on the strengths and limitations of LLMs for human spatial trajectory analysis.

Introduction

The widespread adoption of location-enabled mobile devices has led to a massive collection of human mobility data (Mokbel et al. 2023), comprising diverse trajectory types from individual app usages to public transportation systems. These mobility traces can be modeled as dynamic graphs, representing sequences of location visits with associated semantics (Parent et al. 2013). Analyzing these dynamic graphs enables valuable insights for applications like transportation mode classification and detecting spatiotemporal patterns (Lee, Han, and Whang 2007; Yao et al. 2017; Biljecki, Ledoux, and Van Oosterom 2013; Zheng et al. 2010a; Toohey and Duckham 2015; Zhang and Zhao 2021; Bai et al. 2023; Guo, Wang, and Zhao 2022). A particularly difficult task is identifying anomalous mobility patterns within an individual's semantic trajectories, where the trajectory significantly deviates from their historical patterns. Finding such anomalous patterns of individuals may indicate a change in behavior which has many important applications. For instance, infectious disease monitoring (Kohn, Amiri, and Züfle 2023; Rambhatla et al. 2022; Zhang and Zhao 2022) or tracking elderly behaviors (Stavropoulos et al. 2020).

In recent times, there has been a surge in progress with large language models (LLMs) (Qiu et al. 2020; Ling et al. 2023) like Transformers (Vaswani et al. 2017), BERT (Devlin et al. 2019), GPT (Brown et al. 2020), among others. These LLMs act as foundational models, which can be easily adapted for various downstream applications with minimal adjustments (Brown et al. 2020; Kojima et al. 2022; Ling et al. 2023; Zhao et al. 2023). Notably, breakthroughs in design and training techniques have enabled emerging abilities in LLMs, distinguishing cutting-edge models like GPT-3.5 (Brown et al. 2020), GPT-4 (OpenAI 2023), Claude-2 (Bai et al. 2022), BARD (Google AI 2022), LlaMA (Touvron et al. 2023a), and LlaMA-2 (Touvron et al. 2023b) from earlier versions. For example, features such as in-context training (Min et al. 2021) and zero-shot learning (Kojima et al. 2022; Wei et al. 2021) allow these models to adapt to tasks they were not explicitly trained for.

Despite the remarkable progress LLMs have made in diverse NLP tasks like question answering (QA) and machine translation, their potential in analyzing human mobility patterns remains largely unexplored. Human mobility data, unlike typical language sequences, presents with intricate spatial-temporal dynamics and rich topological connections between entities. Detecting anomalous behaviors are especially difficult due to the intrinsic property of unkown nature of anomalies. Existing methods typically rely on creating hand-crafted features such as the total traveled distance and use heuristic rule to determine outliers, which limits their capability to generalize effectively to detect unseen outlier patterns. In contrast, LLMs has natural advanatge since they can directly perceive natural langauge input. As LLMs have already shown powerful reasoning ability and generalization capabilities directly from the input prompt, it becomes intriguing to assess to what extent LLMs can detect diverse anomaly behaviors under the human mobility patterns.

To systematically study the capabilities of LLMs on detecting outliers (anomlies) in human mobility trajectories, we conduct a series of empirical experiments with leading LLMs on diverse datasets. By comparing their performance to specialized human mobility anomaly detection methods, we aim to assess the potential strengths and limitations of LLMs in this domain. Critically, by altering the input prompt formats, we aim to evaluate how effectively LLMs can extract and leverage the underlying structural information from the dynamic mobility patterns to enhance their performance in subsequent tasks. Moreover, we delve into both the effectiveness and interpretability of LLMs' predictions.

Simulation and Dataset Generation

This section describes both the simulated using the Patterns-of-Life Simulation (Amiri et al. 2023; Kim et al. 2020) and real-world dataset based on the GeoLife dataset (Zheng et al. 2010b). Specifications of the datasets, including details and key attributes, can be found in Table 1. The source code of the simulation and data processing of the GeoLife dataset is accessible through the GitHub repository: https://github.com/azufle/pol. In addition, the dataset is available for download at https://osf.io/rxnz7/.

Outlier Type	#Agents	Source	Period	#Outliers
hunger	1000	POL	450+14 days	90
work	1000	POL	450+14 days	30
social	1000	POL	450+14 days	30
combined	3000	POL	450+14 days	150
imposter	69	GeoLife	4 years	20

Table 1: Specification of the datasets utilized in this paper.

Simulation of Patterns of Life

The patterns of life simulation was designed to emulate human needs and behavior in an urban environment (Züfle et al. 2023). Within the simulated environment, virtual entities referred to as agents perform actions that mirror human activities. These include attending work, forming friendships, engaging in social gatherings, and more. The agents' existence is crafted to resemble human life in a real-world environment (roads, buildings) obtained from OpenStreetMap (Bennett 2010). Throughout their simulated lives, agents navigate to diverse locations, including restaurants, workplaces, residential apartments, and recreational venues. A salient feature of the simulation is the generation of comprehensive log files. These logs contain extensive data regarding the agents, including their location and current state information, thus allowing for in-depth analysis and research.

In our study, we generated data by running simulations over four distinct maps, namely Fairfax County, Virginia, USA (FVA); the French Quarter of New Orleans, Louisiana, USA (NOLA); Atlanta, Georgia, USA (ATL); and Beijing, China (BJNG). The simulations were conducted over a period of 450 days to replicate normal life, followed by an additional 14 days to incorporate abnormal behavior into the regular patterns. We introduced three specific types of abnormal behavior that define outliers trajectories:

- **Hunger outlier:** An agent under this category becomes hungry more quickly. Such agents have to go to restaurants or their homes much more often.
- **Social outlier:** This type of agent randomly selects recreational sites to visit when needed, rather than being guided by their attributes and social network.
- **Work outlier:** Agents in this category abstain from going to work on workdays.

We further divided these abnormalities into three intensity levels: red, orange , and yellow . Red outliers exhibit extremely abnormal behavior, orange outliers act moderately abnormal, and yellow outliers display abnormal behavior less frequently. For example, a work outlier will decide not to go to work 100%, 50%, or 20% of the time when classified as red, orange , or yellow , respectively. We divide the simulation into 450 simulation of days of normal behavior followed by 14 days of a small number of agents exhibiting outlier behavior. Details can be found in Table 1.

Real World Dataset (GeoLife)

The real-world dataset for this study was created using the Microsoft Research Asia's GPS Trajectory dataset (Zheng et al. 2010b). Since the original data did not conform to a check-in format, we employed the method outlined in (Zheng et al. 2008) to extract stay points, thereby transforming the data to fit the check-in pattern used in life simulation studies. Next, we utilized OpenStreetMap to categorize locations into four groups: apartments, workplaces, pubs, and restaurants. Given that OpenStreetMap encompasses a broad array of categories and types, we manually classified them into these four distinct groups. Upon preprocessing the data, we eliminated agents with fewer than 50 records, resulting in a final count of 69 agents with a total of 14,080 training trajectories and 3,552 test trajectories. Within the context of the GeoLife dataset, we introduced a specific outlier type called the "imposter outlier". An agent acting as an imposter outlier by switching the trajectories with another agent after a specific time point.The dataset was then divided into two segments: 80% of the stay points for training and introduced outliers into the remaining 20% for test.

Experiments

Experimental Settings

Datasets. We conducted the experiments on two human mobility benchmark datasets: GEOLIFE (Zheng et al. 2010b) and PATTERNS-OF-LIFE (Amiri et al. 2023; Kim et al. 2020). Brief descriptions of the datasets are as follows:

- GEOLIFE: This dataset was created using the Microsoft Research Asia's GPS Trajectory dataset (Zheng et al. 2010b). The GeoLife dataset, sourced from the GeoLife project by Microsoft Research Asia, captures the GPS trajectories of 182 users over a span of more than three years, from April 2007 to August 2012. Each trajectory in this dataset consists of time-stamped points detailing the latitude, and longitude. The dataset provides insights into leisure and sports activities, such as shopping, sightseeing, dining, hiking, and cycling, offering a comprehensive view of users' outdoor movements. We first eliminated agents with fewer than 50 records, resulting in a final count of 69 users. Due to the lack of ground truth anomalies, we introduced a specific outlier type called the "imposter outlier", by switching the trajectories with another agent after a specific time point. In this paper we choose 80% of the stay points of trajectories as the switching point.
- PATTERNS-OF-LIFE (POL): A simulated dataset, where agents emulate human activities such as working, socializing, and more, in a real-world-like setting sourced from

Task setting	Prompt to LLM
Separate	Task: You are a human mobility trajectory behavior anomaly detector. Given a historical human trajectory information, can you analyse the pattern behind the trajectory and give an anomaly score (from 0 to 1, where larger value indicates more abnormal) of this user's behavior? Here is the sequence of trajector: [Sequence]. Give your analysis and present your esimated anomaly score (from 0 to 1, where larger value indicates more abnormal) inside a pair of square brackets.
Separate With Hint	Task: (. . . Same as Above . . .) Hint: The anomaly users would suddenly change their mobility pattern starting from a time point, which means after a certain time, their mobility behavior would significantly deviate from their past behaviors. We would use "***[deviate-point]***" inside each trajectory to denote the time point as hint. Here is the sequence of trajector: [Sequence-before] ***[deviate-point]*** [Sequence-after]. Give your analysis and present your esimated anomaly score (from 0 to 1, where larger value indicates more abnormal) inside a pair of square brackets.
Combine	Task: You are a human mobility trajectory behavior anomaly detector. Given a set of N users' historical human trajectories information, can you analyse the pattern behind each user's trajectory and give an anomaly score (from 0 to 1, where larger value indicates more abnormal) of users' behavior? Here is the sequence of user 1: [Sequence-1] . . . Here is the sequence of user N: [Sequence-N] Give your analysis and present your esimated anomaly scores about all users (from 0 to 1, where larger value indicates more abnormal):

Table 2: Examples of different prompts used in anomaly detection experiments. We also have 'Combine With Hint' prompt which is the similar way of adding hints in 'Separate With Hint' prompt. A detailed prompt example is given in Appendix 6 due to space limitation.

OpenStreetMap (Bennett 2010). Throughout their simulated lives, agents navigate to diverse locations, including restaurants, workplaces, residential apartments, and recreational venues. We generate 1,000 users over a span of 4 weeks. We introduced 90 anomaly users with three specific types of abnormal behavior: (1) **Hunger outlier:** An agent under this category becomes hungry more quickly. Such agents have to go to restaurants or their homes much more often. (2) **Social outlier:** This type of agent randomly selects recreational sites to visit when needed, rather than being guided by their attributes and social network. (3) **Work outlier:** Agents in this category abstain from going to work on workdays.

Comparison Methods. We compare with several unsupervised trajectory outlier detection methods, including three non-deep learning methods **OMPAD** (Basharat, Gritai, and Shah 2008), **MoNav-TT** (Zhang 2012) and **TRAOD** (Lee, Han, and Li 2008), and two state-of-the-art deep learning methods **DSVDD** (Ruff et al. 2018) and **DAE** (Zhou and Paffenroth 2017; Dotti, Popa, and Asteriadis 2020). A detailed introduction of the comparison methods can be found in Appendix .

LLM Detection Results

Broadly, this paper focuses on studying the central question of investigating the capabilities of LLMs on identifying anomalous behaviors within human mobility patterns from three perspectives:

- **Can LLMs effectively detect anomalous behaviors within human mobility patterns without any indicative information?** It is intriguing to assess whether LLMs can attain substantial predictive performance on anomaly detection tasks, even in the absence of any clue about the anomalies, e.g. such as temporal occurrence or the nature of the anomaly.
- **Can providing indicative clues about the anomaly enhance the detection efficacy of LLMs?** Incorporating specific clues or hints about potential anomalies might bolster the LLM's ability to identify irregularities more accurately. By offering contextual information, it could guide the LLM to focus on certain aspects of the data and make more informed predictions.
- **Can LLMs provide reasonable explanation to their judgements?** Beyond mere classification, it is imperative to observe whether LLMs can elucidate the fundamental reasoning behind their determinations. Specifically, can these models articulate the underlying rationale when predicting human mobility patterns as anomalous or normal, thereby enhancing the transparency and trustworthiness of their judgments?

Prompt Settings. To systematically study the research questions above, we design two different dimensions to create

	Geolife				Patterns-of-Life			
Model	Top-10 Hits	Top-25 Hits*	AP score	AUC score	Top-10 Hits	Top-100 Hits	AP score	AUC score
OMPAD	1	4	0.1665	0.1697	0	0	0.0079	0.4512
MoNav-TT	0	7	0.2849	0.3989	0	0	0.0094	0.4798
TRAOD	4	7	0.1060	0.5498	0	1	0.0030	0.4390
DSVDD	7	15	0.6246	0.7714	1	2	0.0120	0.5398
DAE	5	12	0.4627	0.6234	0	1	0.0089	0.4649
GPT-3.5	5	8	0.4014	0.4979	0	6	0.0365	0.7572
GPT-3.5-with-hint	4	12	0.3741	0.5917	0	2	0.0176	0.6220
GPT-4	3	9	0.2732	0.4417	-	-	-	-
GPT-4-with-hint	5	8	0.3181	0.4818	-	-	-	-
Claude-2	4	13	0.4756	0.7474	-	-	-	-
Claude-2-with-hint	7	16	0.6879	0.8875	-	-	-	-

Table 3: Outlier detection performance for all datasets. The best performance for AP and AUC scores is highlighted for each dataset. *We report Top-25 Hits instead of Top-100 for Geolife dataset due to their size constraints on datasets. (-) denotes the absence of experiments due to the API cost issue.

input prompts for the LLMs. Specifically, (1) **With/Without Hint**: Given that an anomaly begins at a specific time point in the data, it is crucial to evaluate the performance of the LLMs whether this information is provided or not. Notably, for all comparative methods, this hint is used to divide the data into training and testing sets.; (2) **Separate vs Combine**: It would also be interesting to assess whether there is a significant difference in performance when presenting all the trajectories in a single prompt versus in separate prompts. This is because placing them all in one prompt might allow the model to consider interactions between different trajectories. In Table 2, we present examples to illustrate the details of prompt.

Choices of LLMs. We opted to utilize OpenAI's state-of-the-art models, GPT-3.5 and GPT-4, via their API system, and Claude-2 by Anthropic. Specifically, we use gpt-3.5-turbo-16k-0613 and gpt-4-0613 for 'Separate' prompt, and Claude-2 for 'Combine' prompt due to its capability to hold long input prompt up to 100K input context window size.

- **LLMs can effectively detect anomaly behaviors without any indicative information.** We observed that the LLM demonstrates commendable detection results on both datasets. For the Geolife dataset, Claude-2 surpasses all non-deep learning methods, achieving performance on par with the deep learning method. Both GPT-3.5 and GPT-4 also produce results that are comparable to those of other methods. This might suggest that presenting all mobility trajectories in a single prompt may lead to better performance than using separate prompts. As for the PoL dataset, the GPT-3.5 model significantly outperforms all the methods it was compared against.
- **Providing additional indicative information can further enhance the detection efficacy of LLMs.** We observed that by incorporating a 'hint' into the LLMs, detection performance consistently improved across all models when tested on the Geolife dataset. Notably, Claude-2-with-hint demonstrated a significantly superior detection rate, surpassing all other comparison methods. On the other hand, there was a slight dip in performance on the PoL dataset when adding the hint. This could arguably be attributed to the LLM's ability to manage longer input temporal trajectories, as evidenced by the average length of trajectories being 52.2 for Geolife and 182.0 for PoL.
- **LLMs is capable to provide reasonable explanation to their judgements.** Examples of generated explanations alongside predictions can be found in Appendix 4. Notably, we observed that the LLMs are capable of providing cogent explanations for their prediction results. Such clarity is pivotal for ensuring transparency in anomaly detection methods.

Conclusion and Future Works

In this work, we conduct empirical studies to provide insights on the strengths and limitations of large language models (LLMs) for detecting anomalous behaviors from mobility data, by comparing LLMs to specialized anomaly detection methods. Our key findings show that LLMs can achieve promising anomaly detection performance even without any specific cues about potential anomalies. Furthermore, providing contextual information about possible irregularities can enhance the prediction accuracy of LLMs. In addition, LLMs can provide explanations for their anomaly judgments, thereby improving model transparency. For future work, we plan to study the effectiveness of open source LLMs such as Llama-2 models to improve model transparency. We also aim to address the issue that LLMs have difficulty processing long mobility trajectories due to the limited context window size. Moreover, we intend to evaluate our approach on additional mobility datasets. This work represents an initial exploration of applying LLMs for the important and promising task of mobility anomaly detection. We hope it will inspire more research in this direction.

References

Amiri, H.; Ruan, S.; Kim, J.-S.; Jin, H.; Kavak, H.; Crooks, A.; Pfoser, D.; Wenk, C.; and Zufle, A. 2023. Massive Trajectory Data Based on Patterns of Life. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701689.

Bai, G.; Yu, Z.; Chai, Z.; Cheng, Y.; and Zhao, L. 2023. Staleness-Alleviated Distributed GNN Training via

Online Dynamic-Embedding Prediction. *arXiv preprint arXiv:2308.13466*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Basharat, A.; Gritai, A.; and Shah, M. 2008. Learning object motion patterns for anomaly detection and improved object detection. In *2008 IEEE conference on computer vision and pattern recognition*, 1–8. IEEE.

Bennett, J. 2010. *OpenStreetMap*. Packt Publishing Ltd.

Biljecki, F.; Ledoux, H.; and Van Oosterom, P. 2013. Transportation mode-based segmentation and classification of movement trajectories. *International Journal of Geographical Information Science*, 27(2): 385–407.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T. J.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. *ArXiv*, abs/2005.14165.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dotti, D.; Popa, M.; and Asteriadis, S. 2020. A hierarchical autoencoder learning model for path prediction and abnormality detection. *Pattern Recognition Letters*, 130: 216–224.

Google AI. 2022. BARD: Deepmind's Conversational AI Assistant. https://blog.google/products/search/introducing-bard-google-ai. Accessed: January 10, 2024.

Guo, X.; Wang, S.; and Zhao, L. 2022. Graph neural networks: Graph transformation. *Graph Neural Networks: Foundations, Frontiers, and Applications*, 251–275.

Kim, J.-S.; Jin, H.; Kavak, H.; Rouly, O. C.; Crooks, A.; Pfoser, D.; Wenk, C.; and Züfle, A. 2020. Location-based social network data generation based on patterns of life. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, 158–167. IEEE.

Kohn, W.; Amiri, H.; and Züfle, A. 2023. EPIPOL: An Epidemiological Patterns of Life Simulation (Demonstration Paper). In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on Spatial Computing for Epidemiology*, 13–16.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Lee, J.-G.; Han, J.; and Li, X. 2008. Trajectory outlier detection: A partition-and-detect framework. In *2008 IEEE 24th International Conference on Data Engineering*, 140–149. IEEE.

Lee, J.-G.; Han, J.; and Whang, K.-Y. 2007. Trajectory clustering: a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, 593–604.

Ling, C.; Zhao, X.; Lu, J.; Deng, C.; Zheng, C.; Wang, J.; Chowdhury, T.; Li, Y.; Cui, H.; et al. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*.

Min, S.; Lewis, M.; Zettlemoyer, L.; and Hajishirzi, H. 2021. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.

Mokbel, M.; Sakr, M.; Xiong, L.; Züfle, A.; Almeida, J.; Aref, W.; Andrienko, G.; Andrienko, N.; Cao, Y.; Chawla, S.; et al. 2023. Towards Mobility Data Science (Vision Paper). *arXiv preprint arXiv:2307.05717*.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*, abs/2303.08774.

Parent, C.; Spaccapietra, S.; Renso, C.; Andrienko, G.; Andrienko, N.; Bogorny, V.; Damiani, M. L.; Gkoulalas-Divanis, A.; Macedo, J.; Pelekis, N.; et al. 2013. Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)*, 45(4): 1–32.

Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; and Huang, X. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10): 1872–1897.

Rambhatla, S.; Zeighami, S.; Shahabi, K.; Shahabi, C.; and Liu, Y. 2022. Toward Accurate Spatiotemporal COVID-19 Risk Scores Using High-Resolution Real-World Mobility Data. *ACM Trans. Spatial Algorithms Syst.*, 8(2): 1–30.

Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S. A.; Binder, A.; Müller, E.; and Kloft, M. 2018. Deep one-class classification. In *International conference on machine learning*, 4393–4402. PMLR.

Stavropoulos, T. G.; Papastergiou, A.; Mpaltadoros, L.; Nikolopoulos, S.; and Kompatsiaris, I. 2020. IoT wearable sensors and devices in elderly care: A literature review. *Sensors*, 20(10): 2826.

Toohey, K.; and Duckham, M. 2015. Trajectory similarity measures. *Sigspatial Special*, 7(1): 43–50.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023a. LLaMA: Open and Efficient Foundation Language Models. *ArXiv*, abs/2302.13971.

Touvron, H.; Martin, L.; Stone, K. R.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D. M.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A. S.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I. M.; Korenev, A. V.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.;

Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv*, abs/2307.09288.

Vaswani, A.; Shazeer, N. M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Yao, D.; Zhang, C.; Zhu, Z.; Huang, J.; and Bi, J. 2017. Trajectory clustering via deep representation learning. In *2017 international joint conference on neural networks (IJCNN)*, 3880–3887. IEEE.

Zhang, J. 2012. Smarter outlier detection and deeper understanding of large-scale taxi trip records: a case study of NYC. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, 157–162.

Zhang, Z.; and Zhao, L. 2021. Representation learning on spatial networks. *Advances in Neural Information Processing Systems*, 34: 2303–2318.

Zhang, Z.; and Zhao, L. 2022. Unsupervised Deep Subgraph Anomaly Detection. In *2022 IEEE International Conference on Data Mining (ICDM)*, 753–762. IEEE.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zheng, Y.; Chen, Y.; Li, Q.; Xie, X.; and Ma, W.-Y. 2010a. Understanding transportation modes based on GPS data for web applications. *ACM Transactions on the Web (TWEB)*, 4(1): 1–36.

Zheng, Y.; Xie, X.; Li, Q.; and Ma, W.-Y. 2008. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL conference on Advance in Geographical Information Systems*.

Zheng, Y.; Xie, X.; Ma, W.-Y.; et al. 2010b. GeoLife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2): 32–39.

Zhou, C.; and Paffenroth, R. C. 2017. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 665–674.

Züfle, A.; Wenk, C.; Pfoser, D.; Crooks, A.; Kim, J.-S.; Kavak, H.; Manzoor, U.; and Jin, H. 2023. Urban life: a model of people and places. *Computational and Mathematical Organization Theory*, 29(1): 20–51.

Comparison Methods

OMPAD (Basharat, Gritai, and Shah 2008) is an outlier detection method that analyzes objects' movement patterns by counting the types of locations they visit. It identifies abnormal activities by measuring the deviations in moving trends compared to established normal patterns.

MoNav-TT (Zhang 2012) is an outlier detection algorithm tailored for urban human trajectory networks, where it detects outliers by measuring discrepancies in traffic distances. In particular, a user is identified as an outlier if the traveled distance significantly deviates from their previous behavior.

TRAOD (Lee, Han, and Li 2008) is a partition-and-detect framework for trajectory outlier detection, which partitions a trajectory into a set of line segments, and then, detects outlying line segments for trajectory outliers.

DSVDD (Ruff et al. 2018) is a deep one-class classification based outlier detection method. We generalize it to handle the task of semantic trajectory outlier detection in a most intuitive way. We map the weekly trajectories of each user to a high dimensional sphere by a deep neural network encoder. Then the distance of trajectories from the sphere's surface is quantified as an outlier score.

DAE (Zhou and Paffenroth 2017; Dotti, Popa, and Asteriadis 2020) is a widely-used outlier detection method that leverages a deep autoencoder. Utilizing an encoder-decoder model architecture, it reconstructs input trajectories, and the resulting reconstruction error is used as an outlier indicator, signifying deviations from the normal pattern.

For the data preprocessing, we transform the semantic types of locations into onehot encodings and treat the spatial coordinates as 3D vectors as the input to all the comparison methods.

Model	Output
GPT-3.5 No hint	Based on the given trajectory data, the anomaly score indicates the abnormal behavior of the user's mobility pattern. The analysis of the trajectory suggests that the user's movements appear highly irregular and inconsistent. The user tends to visit different types of locations (e.g., Pub, Workplace, Apartment) frequently and travels long distances within short time durations. Additionally, the user exhibits sudden changes in visited locations, such as back-and-forth movements between the same places.Furthermore, the user shows unusual patterns of movement, such as traveling long distances in short periods (e.g., a trip of 250.1 km followed by a trip of 17.4 km) or visiting distant locations in rapid succession (e.g., traveling 90.3 km to a Pub). These patterns deviate significantly from regular human mobility behavior.Considering all these factors, the estimated anomaly score for this user's behavior is [0.85]. This high score suggests that the user's mobility trajectory exhibits highly abnormal patterns and deviates significantly from typical human movements.
GPT-4 No hint	Analyzing the user's given mobility trajectory, several patterns and potential anomalies are noted. Most noticeably, the user frequents pubs substantially more than any other location type. This is somewhat unusual as typically individuals might frequently visited places like apartments or workplaces. Another significant observation is the abrupt and sporadic changes in distance to the next location - with some movements being in the order of hundreds of kilometers, indicating long-distance travel, and then returning immediately to local sites in a short span of time. The routine visits to the pubs often happen at atypical time slots, including times when a standard pub would potentially not be open or the individual would traditionally be at the workplace. The lack of visit to other essential places like shops or eateries also is a little abnormal. However, the data set does not provide any benchmark for what could be considered 'normal' behavior, making it challenging to give a precise anomaly score. Therefore, based on the noticeable inconsistencies in the temporal and spatial patterns, I would estimate the anomaly score to be moderately high, but not extreme.Estimated Anomaly Score: [0.65]

Table 4: Example LLMs' generated answers in geolife dataset without hint.

Model	Output
GPT-3.5 With hint	Analysis and Estimated Anomaly Score: Looking at the given trajectory data, we can observe the following patterns:1. The user frequently visits the Pub and Apartment locations.2. The distances traveled between locations are varying, with some long distances and some short distances. 3. There is a “***¡deviate-point¿***” denoting a time point where the user’s mobility behavior significantly changes.Based on these patterns, we can analyze the trajectory and estimate the anomaly score. Before the deviate point:- The user visits the Pub and Apartment locations alternately with varying distances.- The trajectory seems consistent and follows a pattern.After the deviate point:- The user visits the Apartment again, but the distance traveled is shorter compared to previous visits. The user then visits the Workplace, which is a new location not seen before in the trajectory.- The user visits the Pub again but travels a longer distance than before.This change in the mobility pattern, the addition of a new location, and the differences in distance traveled suggest a deviation from the user’s past behavior.Based on these observations, the estimated anomaly score for this user’s behavior would be relatively high. Given that larger values indicate more abnormal behavior, the estimated anomaly score can be represented as [0.9].
GPT-4 With hint	Analyzing the trajectory data before the “***¡deviate-point¿***”, the individual had a somewhat consistent pattern of movement. On weekends (Saturday), they frequently visited a Pub and their Apartment. Sometimes, they also visited a Restaurant. There was also consistent movement between their Apartment on Wednesdays and Thursdays.However, after the deviate point, there has been a significant change in behaviour. For example, they visit their Apartment on different days (Wednesday and Thursday), with a visit to the Workplace added into the mix. Moreover, they started going to the Pub also on Tuesday, which was not previously observed.The traveled distances also have increased after the deviate point showing that the user has significantly changed their mobility pattern, particularly regarding the type and average distance of trips. Overall, the sudden deviation from their previous behavioral pattern does indeed make the user suspicious, which we are considering as an anomaly. It clearly indicates that the user’s activity after the “deviate-point” is significantly different from their past behaviours. Anomaly score is subjectively judged based on the provided context and may vary based on specific models used, but based on my analysis, I’d put it at around [0.7] as there’s a clear and substantial change in behaviour, but this still has room for more drastic changes, such as the inclusion of entirely new location types or far greater distances. So it does pose a certain level of abnormality, but there’s potential for it to be higher.

Table 5: Example LLMs’ generated answers in geolife dataset with hint.

Task setting	Prompt to LLM
Separate	Task: You are a human mobility trajectory behavior anomaly detector. Given a historical human trajectory information, can you analyse the pattern behind the trajectory and give an anomaly score (from 0 to 1, where larger value indicates more abnormal) of this user's behavior? Description of input trajectory data: A temporal sequence of visited place points, each place is consisted of the visited timestamp and its type of location. Then the traveled distance to next location is given. Here is the sequence of trajector: Sat 10:36, Pub, 0.4 km →Sat 10:57, Apartment, 4.8 km →Sat 13:44, Pub, 3.4 km →Sat 04:45, Apartment, 4.9 km →Sat 07:39, Pub, 0.6 km →Sat 09:44, Apartment, 0.8 km →Sat 10:14, Apartment, 3.1 km → Wed 00:30, Apartment, 1.1 km →Wed 06:08, Apartment, 0.2 km →Wed 06:44, Apartment, 1.3 km → Wed 07:27, Apartment, 13.5 km →Sat 03:39, Apartment, 0.3 km →Sat 04:17, Apartment, 2.1 km →Sat 05:52, Apartment, 3.7 km →Sat 08:48, Pub, 0.6 km →Sat 09:36, Restaurant, 11.9 km →Thu 06:01, Apartment, 1.5 km →Wed 08:25, Apartment, 1.3 km →Thu 01:05, Workplace, 0.6 km →Sat 01:18, Pub, 10.4 km →Sat 05:33, Pub. Give your analysis and present your esimated anomaly score (from 0 to 1, where larger value indicates more abnormal) inside a pair of square brackets [] :
Separate With Hint	Task: You are a human mobility trajectory behavior anomaly detector. Given a historical human trajectory information, can you analyse the pattern behind the trajectory and give an anomaly score (from 0 to 1, where larger value indicates more abnormal) of this user's behavior? Hint: The anomaly users would suddenly change their mobility pattern starting from a time point, which means after a certain time, their mobility behavior would significantly deviate from their past behaviors. We would use "***deviate-point***" inside each trajectory to denote the time point as hint. Description of input trajectory data: A temporal sequence of visited place points, each place is consisted of the visited timestamp and its type of location. Then the traveled distance to next location is given. Here is the sequence of trajector: Sat 10:36, Pub, 0.4 km →Sat 10:57, Apartment, 4.8 km →Sat 13:44, Pub, 3.4 km →Sat 04:45, Apartment, 4.9 km →Sat 07:39, Pub, 0.6 km →Sat 09:44, Apartment, 0.8 km →Sat 10:14, Apartment, 3.1 km →Wed 00:30, Apartment, 1.1 km →Wed 06:08, Apartment, 0.2 km →Wed 06:44, Apartment, 1.3 km →Wed 07:27, Apartment, 13.5 km →Sat 03:39, Apartment, 0.3 km →Sat 04:17, Apartment, 2.1 km →Sat 05:52, Apartment, 3.7 km →Sat 08:48, Pub, 0.6 km →Sat 09:36, Restaurant, 11.9 km →Thu 06:01, Apartment ***deviate-point*** , 1.5 km →Wed 08:25, Apartment, 1.3 km →Thu 01:05, Workplace, 0.6 km →Sat 01:18, Pub, 10.4 km →Sat 05:33, Pub. Give your analysis and present your esimated anomaly score (from 0 to 1, where larger value indicates more abnormal) inside a pair of square brackets [] :

Table 6: A detailed 'Separate' prompt example in geolife dataset.