Extracting Location-Based Insights from Unstructured Urban and Mobility Data with Large Language Models

Yihong Tang
McGill University
yihong.tang@mail.mcgill.ca

Lijun Sun
McGill University
lijun.sun@mcgill.ca

Abstract

Rapid urbanization and technological advancements present an exciting opportunity to re-design urban transportation systems. Traditional models rely heavily on static, structured data, which fail to capture urban mobility's dynamic, human-centered nature. However, the emergence of unstructured data sources, such as web posts, social media content, and user-generated information, provides a wealth of location-specific insights in real-time that can make transportation systems more adaptive and user-centric. Large language models (LLMs), with their advanced natural language processing capabilities, are particularly well suited for handling unstructured data. They can interpret context, detect patterns, and extract valuable insights from vast and diverse data streams. This paper explores how LLMs can process these unstructured data streams to enhance urban mobility systems. By effectively extracting, transforming, and using these data sources with LLMs and the corresponding tools, we can develop more adaptive transportation systems that improve efficiency, accessibility, equity, and user experience. Our approach outlines a general workflow for handling unstructured urban and transportation data, with a specific focus on extracting meaningful location-based information relevant to urban and mobility-related challenges.

1 Motivation and Background

The limitations of traditional urban mobility data stem from its structured nature, which primarily captures pre-defined metrics such as scheduled transit operations, fixed infrastructure constraints, and aggregated traffic patterns. Although valuable for transportation planning and system optimization, these datasets do not reflect the fluidity of real-world mobility behaviors. Human movement patterns are shaped by numerous unpredictable factors, including spontaneous travel decisions [7], weather or event-induced disruptions [6]. As a result, transportation systems that rely solely on structured data often struggle to adapt to emerging urban challenges, such as sudden congestion surges [8], localized accessibility issues [3], or shifting commuter trends [9].

Unstructured data sources, including user-generated content on social networks, mobility-related online records, and crowd-sourced travel reviews, offer a unique lens into these real-time urban dynamics. Unlike structured data, these sources provide granular insights into user sentiment, service satisfaction, and evolving mobility trends. For instance, a surge of complaints about delays at a particular station on social media might indicate an unreported operational issue, while a spike in ride-sharing discussions in a specific neighborhood could reflect a growing demand for alternative transit options. However, the sheer volume, variability, and noise inherent in unstructured data make it difficult to systematically extract meaningful, actionable insights.

Conventional approaches to analyzing such data typically involve manual review or rudimentary keyword-based extraction techniques, both of which are impractical and time-consuming processes that fail to capture detailed context or infer dependencies between different data points. The emergence of large language models (LLMs) [1, 2] presents a transformative opportunity to bridge this gap. By leveraging their advanced natural language understanding and contextual reasoning capabilities, LLMs can process and extract large amounts of unstructured, location-specific data, automatically categorizing and structuring information in ways that are directly useful for urban mobility planning [4]. This capability enables the development of more adaptive real-time transportation models that integrate human-centric insights, allowing for smarter, more responsive urban mobility solutions.

2 Proposed Approach

Given the previously mentioned challenges and opportunities, we propose a conceptual framework that leverages LLMs to extract meaningful location-based insights from unstructured data across various urban, transportation, and mobility domains. The framework focuses on a generalized, scalable workflow for processing diverse data sources, which includes web posts, social media interactions, reviews, and other forms of user-generated content.

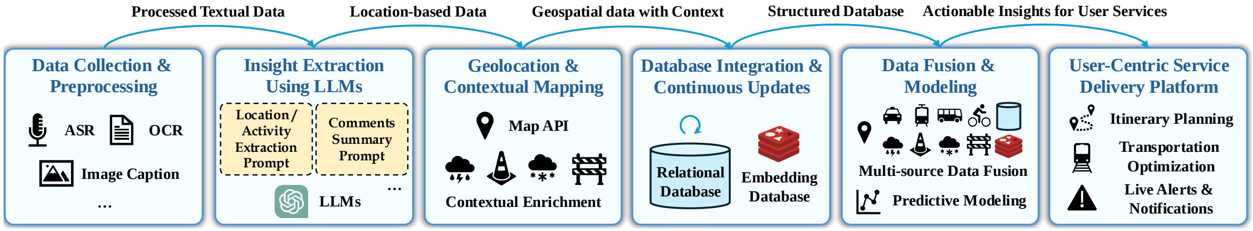
Processed Textual Data
Location-based Data
Geospatial data with Context
Structured Database
Actionable Insights for User Services
Data Collection & Preprocessing
ASR
OCR
Image Caption
...
Insight Extraction Using LLMs
Location / Activity Extraction Prompt
Comments Summary Prompt
...
LLMs
Geolocation & Contextual Mapping
Map API
Contextual Enrichment
Database Integration & Continuous Updates
Relational Database
Embedding Database
Data Fusion & Modeling
Multi-source Data Fusion
Predictive Modeling
User-Centric Service Delivery Platform
Itinerary Planning
Transportation Optimization
Live Alerts & Notifications

Figure 1: The proposed conceptual framework.

The overall workflow is shown in Figure 1 and is detailed as follows:

- **Data Collection and Preprocessing:** Unstructured data are scraped from diverse online platforms, including social media posts, travel blogs, forums, and multimedia content (images, videos). Techniques like Automatic Speech Recognition (ASR), Optical Character Recognition (OCR), and Image Captioning (IC) are applied to transcribe multimedia content and extract text from images, ensuring that multimedia sources are also included in the analysis.
- **Insight Extraction Using LLMs:** LLMs are prompted or fine-tuned to process location-based, transportation-related, and urban mobility content. These models identify location-specific references, such as places, events, transport modes, and routes, along with contextual details like user sentiment, activity patterns, and potential challenges. The extracted insights are categorized into relevant themes such as popular routes, congestion hotspots, points of interest, and accessibility challenges.
- **Geolocation and Contextual Mapping:** After identifying meaningful locations and activities, geolocation APIs are used to map these points to geographic coordinates (longitude, latitude). The data is enriched with contextual information, such as real-time activity patterns, environmental factors, or emerging transportation needs (e.g., temporary road closures, special events), to provide a more complete understanding of urban mobility dynamics.
- **Database Integration and Continuous Updates:** The insights are integrated into a dynamic database that is continuously updated through ongoing scraping of relevant unstructured data sources. This ensures that the data reflects real-time shifts in user behavior and urban conditions, helping to adapt transportation models to changing mobility needs.
- **Multi-source Data Fusion and Modeling:** The location-based insights extracted from unstructured data are combined with other structured and semi-structured data, such as traffic data, transit schedules, and environmental sensors. By merging these data streams, we generate adaptive, real-time urban mobility models that better reflect the complexities of the urban environment.
- **User-Centric Service Delivery Platform:** To effectively translate these insights into actionable services, a user-centric platform should be developed to provide real-time mobility solutions based on dynamic information. This platform, potentially an LLM-driven Mobility-as-a-Service (MaaS) system, can offer personalized travel recommendations, multimodal route optimization, demand-responsive transport options, and accessibility enhancements. By integrating live updates and predictive analytics, the platform ensures that users receive timely and relevant mobility support, improving efficiency, convenience, and inclusivity in urban transportation.

The proposed conceptual framework can leverage unstructured, location-based data to enhance various user-centric services, including urban itinerary planning, transportation optimization, and real-time alerts and notifications. Each module in this framework can be further augmented with LLM-assisted capabilities, such as automated data preprocessing, intelligent insight extraction, contextual enrichment, adaptive data fusion, and personalized service recommendations. Figure 2 illustrates an example framework proposed by [5], showcasing an on-demand, LLM-driven urban itinerary generator that dynamically responds to user requests.

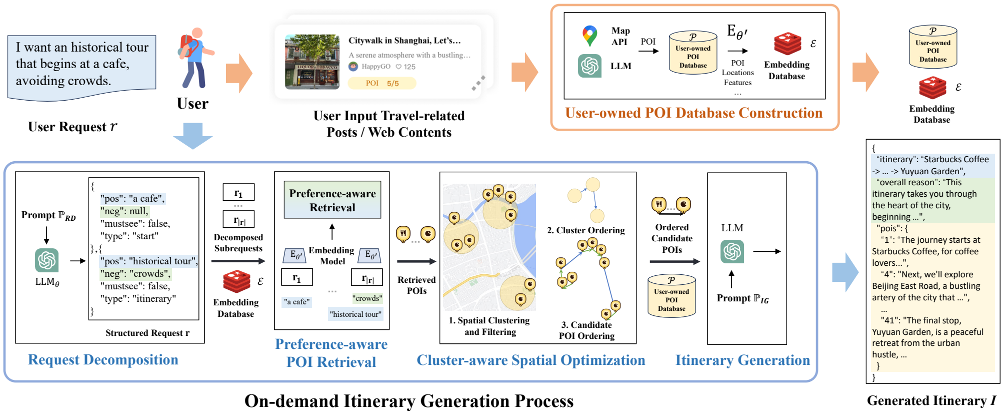
I want an historical tour that begins at a cafe, avoiding crowds.
User Request r
User
Citywalk in Shanghai, Let's...
A serene atmosphere with a bustling...
HappyGO ♡ 125
POI 5/5
User Input Travel-related Posts / Web Contents
Map API
LLM
POI
$\mathcal{P}$
User-owned POI Database
$E_{\theta'}$
POI Locations Features ...
Embedding Database
$\mathcal{E}$
User-owned POI Database Construction
Prompt $\mathbb{P}_{RD}$
LLM$_\theta$
{
"pos": "a cafe",
"neg": null,
"mustsee": false,
"type": "start"
},{
"pos": "historical tour",
"neg": "crowds",
"mustsee": false,
"type": "itinerary"
}
Structured Request r
Request Decomposition
r_1
$r_{|r|}$
Decomposed Subrequests
Embedding Database
Preference-aware Retrieval
Embedding Model
"a cafe"
"crowds"
"historical tour"
Preference-aware POI Retrieval
Retrieved POIs
1. Spatial Clustering and Filtering
2. Cluster Ordering
3. Candidate POI Ordering
Cluster-aware Spatial Optimization
Ordered Candidate POIs
LLM
Prompt $\mathbb{P}_{IG}$
Itinerary Generation
On-demand Itinerary Generation Process
{
"itinerary": "Starbucks Coffee -> ... -> Yuyuan Garden",
"overall reason": "This itinerary takes you through the heart of the city, beginning ...",
"pois": {
"1": "The journey starts at Starbucks Coffee, for coffee lovers...",
"4": "Next, we'll explore Beijing East Road, a bustling artery of the city that ...",
...
"41": "The final stop, Yuyuan Garden, is a peaceful retreat from the urban hustle, ...
}
}
Generated Itinerary I

Figure 2: The example urban itinerary generator design following the conceptual framework.

3 Significance and Value

3.1 Technical Value

One of the key technical advantages of the proposed conceptual framework is its ability to process heterogeneous, noisy, and high-volume data streams efficiently. Unlike structured databases, unstructured data varies in format, language, and context, making it difficult to analyze with traditional rule-based methods. LLMs excel in semantic understanding, contextual inference, and pattern recognition, enabling them to extract meaningful insights without relying on predefined keyword lists or rigid taxonomies. This flexibility allows the system to detect emerging trends, uncover hidden mobility patterns, and generate high-resolution urban insights that evolve dynamically.

Additionally, the multi-source data integration within this framework ensures that urban transportation models remain fluid, responsive, and continuously updated. By combining LLM-driven unstructured data extraction with structured datasets such as public transit schedules, traffic sensor readings, and geospatial databases, the system enables real-time multimodal transportation optimization. This adaptability is crucial for handling unexpected disruptions, peak hour congestion, and emergency response planning, where traditional static models often fall short.

From an implementation perspective, this framework is designed for scalability and interoperability. Cloud-based processing and edge AI capabilities can facilitate real-time data ingestion and analysis across diverse urban regions, ensuring seamless integration with existing smart city infrastructure. Moreover, the API-driven architecture allows it to be incorporated into Mobility as a Service (MaaS) platforms, urban planning dashboards, and transit management systems without requiring extensive modifications to legacy systems.

3.2 Social Value

The ability to collect real-time, user-generated data from unstructured sources represents a democratization of urban mobility data. Unlike traditional data collection methods, which are often centralized and controlled by government agencies or large corporations, this approach allows for a more inclusive and decentralized data ecosystem. The information is processed to be location-based, with user information properly anonymized to protect privacy. Individuals can contribute valuable insights without compromising their privacy, ensuring that the data remains user-owned and ethically sourced. By leveraging diverse and dynamic data streams, the system can cater to a broad spectrum of user needs, whether it is avoiding congestion, identifying safer routes, or discovering underutilized transportation options. This inclusivity fosters a more participatory urban mobility landscape where citizens actively shape their environments rather than passively adapting to them.

Moreover, this conceptual framework enhances equity and accessibility in urban transportation systems. Traditional planning models often overlook marginalized communities, including people with disabilities, lower-income populations, and those living in underserved areas. By integrating real-time, user-driven insights, transportation services can be adapted to accommodate diverse mobility needs, improving accessibility for all. For example, identifying and addressing gaps in public transit coverage or highlighting infrastructure challenges (such as inaccessible pedestrian paths) can lead to more inclusive urban development.

From a sustainability perspective, real-time mobility insights can play a crucial role in promoting efficient and eco-friendly transportation solutions. By dynamically adjusting services based on demand and congestion patterns, urban transit systems can reduce unnecessary emissions and energy consumption. For instance, ride-sharing and public transit optimization based on real-time user input can help reduce single-occupancy vehicle use, contributing to reduced traffic congestion and a lower carbon footprint.

4 Conclusion

This paper presents a conceptual framework that leverages LLMs to transform unstructured, location-based data into actionable insights for urban mobility. By integrating advanced natural language processing with multi-source data fusion, our approach enables real-time situational awareness and adaptive transportation planning that addresses the limitations of traditional, static models.

The proposed framework not only enhances operational efficiency through dynamic data integration but also promotes a more inclusive and responsive urban ecosystem. By democratizing data collection and enabling user-centric service delivery, it lays the groundwork for smarter, equitable, and sustainable urban transportation systems. Future work will focus on refining extraction algorithms, expanding integration capabilities, and validating the framework in real-world settings, ultimately contributing to the evolution of resilient and adaptive cities.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1

[3] Peng Chen, Wei Wang, Chong Qian, Mengqiu Cao, and Tianren Yang. Gravity-based models for evaluating urban park accessibility: Why does localized selection of attractiveness factors and travel modes matter? *Environment and Planning B: Urban Analytics and City Science*, 51(4):904–922, 2024. 1

[4] Yuebing Liang, Yichao Liu, Xiaohan Wang, and Zhan Zhao. Exploring large language models for human mobility prediction under public events. *Computers, Environment and Urban Systems*, 112:102153, 2024. 1

[5] Yihong Tang, Zhaokai Wang, Ao Qu, Yihao Yan, Zhaofeng Wu, Dingyi Zhuang, Jushi Kai, Kebing Hou, Xiaotong Guo, Jinhua Zhao, Zhan Zhao, and Wei Ma. ItiNera: Integrating spatial optimization with large language models for open-domain urban itinerary planning. In Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1413–1432, Miami, Florida, US, November 2024. Association for Computational Linguistics. 2

[6] Dimos Touloumidis, Michael Madas, Vasileios Zeimpekis, and Georgia Ayfantopoulou. Weather-related disruptions in transportation and logistics: A systematic literature review and a policy implementation roadmap. *Logistics*, 9(1):32, 2025. 1

[7] Xin Wang, Asad J Khattak, and Yingling Fan. Role of dynamic information in supporting changes in travel behavior: Two-stage process of travel decision. *Transportation research record*, 2138(1):85–93, 2009. 1

[8] Yulei Wang, Meng Li, Jian Zhou, and Hongyu Zheng. Sudden passenger flow characteristics and congestion control based on intelligent urban rail transit network. *Neural Computing and Applications*, pages 1–10, 2022. 1

[9] Chaoying Yin and Chunfu Shao. Revisiting commuting, built environment and happiness: New evidence on a nonlinear relationship. *Transportation Research Part D: Transport and Environment*, 100:103043, 2021. 1