Large Language Models for Traffic Forecasting: A Survey and Beyond

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have achieved prominent success in various domain-specific applications, driven by their robust foundational capabilities and strong generalization potential. While their impact on natural language processing and computer vision is well-established, recent works highlights their significant promise in other applications as well. One such area is traffic forecasting, where LLMs demonstrated the ability to generate powerful analytical insights, offering new opportunities for advancing Intelligent Transportation Systems (ITS). This survey provides a comprehensive review of the efforts to bridge the gap between textual LLMs and spatio-temporal traffic data, examining their design principles, and improvements in specific traffic scenarios. For deeper insights, we discuss current challenges and highlight future research directions. By providing a foundational understanding of LLMs for traffic forecasting, this survey aims to benefit not only the traffic mining community but also contribute to the broader advancement of Artificial General Intelligence (AGI).

1 Introduction

Large Language Models (LLMs) are a group of models trained on various forms of textual data to understand natural language and generate appropriate response in natural language conversations. They use deep learning techniques, particularly transformer (Vaswani et al., 2017) architectures, to learn the next-word prediction task in an autoregressive manner. A significant strength of LLMs is their great zero-shot and few-shot capabilities (Brown et al., 2020), allowing better generalizability. LLMs have demonstrated superior performance in many downstream tasks like language understanding (Hendrycks et al.), visual question answering (Shao et al., 2023), agents (Wang et al., 2024c), and robotics (Driess et al., 2023). Recently, LLMs have started to demonstrate their strengths in traffic modeling.

Traffic modeling is a crucial aspect of urban planning and management, offering significant benefits for improving daily transportation, economic efficiency, and life quality of urban residents. By predicting traffic conditions using historical data and real-time analytics, authorities can monitor and optimize traffic flow (Akhtar and Moridpour, 2021; Lee et al., 2020; Jiang et al., 2022), achieving better transportation efficiency. For commuters, accurate traffic forecasts can suggest the best times to travel and alternative routes (Lana et al., 2018; Park et al., 2011), enhancing convenience and reducing stress. Besides, advancements in machine learning and IoT allow traffic modeling to be integrated into smart city initiatives, providing a backbone for autonomous vehicle navigation and intelligent transportation systems (Neelakandan et al., 2021; Culita et al., 2020; Chen et al., 2020).

Given their ability to generalize to unseen scenarios and provide zero-shot predictions, LLMs have emerged as promising candidates for traffic modeling (Lai et al., 2023; Xu et al., 2023; Da et al., 2024). Unlike conventional machine learning methods, they do not need extensive fine-tuning and manual data labeling which is often costly. Moreover, LLMs can establish connections between textual descriptions and actual spatial-temporal traffic data, making it easy to identify complex traffic behavior patterns that are difficult to model explicitly.

However, integrating LLMs into traffic modeling poses unique challenges. Many LLMs are pretrained primarily on textual data, making it difficult to interpret non-textual traffic data such as sensor readings and GPS trajectories. Besides, LLMs often lack specific domain knowledge for traffic, making them prone to inaccuracies and hallucinations. Therefore, several works have created specialized LLM algorithms to understand and interact with real-time traffic dynamics. In particular, (Lai et al., 2023) introduced a specialized LLM customarily optimized for traffic signal control tasks. (Xu et al.,

2023) proposed a unified model foundation model for smart city management, which is trained from general models to incorporate specific urban knowledge. (Liu et al., 2024c) use LLMs to forecast time series, by embedding time series data to the latent space of the LLMs, hence avoiding the inconsistency issues of non-autoregressive forecasting.

Despite these efforts, a comprehensive understanding of the current progress and challenges in integrating LLMs with traffic modeling is still lacking. This survey aims to fill that gap by providing a systematic review of the research landscape. We highlight existing methods, explore key challenges, and identify promising directions for future research. By doing so, we aim to offer a clearer picture of how LLMs can be further developed and optimized for traffic applications.

The contributions of this work can be summarized as: (1) A ***pioneering survey on LLM-based traffic forecasting*** models. As illustrated in Figure 1, nearly all related papers were published in 2024, making this work a timely and pioneering survey in this area. ***(2) Systematic review and categorization.*** We thoroughly examine existing preliminary work and provide an in-depth review of efforts to bridge the gap between textual LLMs and spatio-temporal traffic data, covering key design principles, comparisons with related concepts, and specific modules tailored for traffic scenarios. ***(3) Comprehensive perspective.*** We identify and explore gaps in current research, offering insights into future directions for investigations into LLMs and traffic forecasting.

2 Related Surveys

Some surveys reviewing foundation LLMs and their applications on time series data, where LLMs for time series traffic data can be involved as part of their discussions (Jin et al., 2023b; Chen et al., 2024a; Liang et al., 2024b; Zhang et al., 2024c). In July 2024, Zhang et al. (Zhang et al., 2024a) and Wandeltet al. (Wandelt et al., 2024) published two surveys on LLMs and ITS. While they focus on specific applications in downstream scenarios, this paper emphasizes a more comprehensive summary and reflection on the methodology.

3 Design Principle

This section systematically summarizes the high-level design principles. At the data level, LLMs enhanced the sparse traffic data. To bridge the inherent gap between LLMs and spatio-temporal data, they developed feature extraction, prompting, embedding, and the alignment module. Then instruction tuning and reinforcement learning are employed to optimize the performance.

3.1 LLM-assisted Data Enhancer

Traffic data is typically collected through manually installed sensors. Compared to NLP corpora, traffic data is sparser and spans broader temporal and spatial dimensions, making it more challenging to align and analyze (Jin et al., 2024; Hess et al., 2015; Liang et al., 2024b). To address these challenges, some studies have leveraged large language models (LLMs) to enhance traffic data by generating detailed descriptions or task summaries. For instance, LM-MPE (Liang et al., 2024a) uses LLMs to interpret the semantics of textual mobility data, aiding in the understanding of human mobility demand during special events. Similarly, TrafficGPT (Zhang et al., 2024b) integrates traffic models with LLMs to improve reasoning and decision-making analysis. By generating task and dataset summaries, these models enable a deeper understanding of traffic patterns and anomalies, ultimately enhancing the decision-making capabilities and overall effectiveness of LLMs. This approach also makes Traffic-LLMs more interpretable and robust, offering stronger support for complex real-world tasks.

3.2 Bridging the Gap between LLMs and Spatio-Temporal Data

3.2.1 Traffic Feature Extraction

Due to the significant differences in modality and context between text corpora and traffic flow patterns, LLMs trained on text data are challenging to directly apply to traffic flow forecasting tasks. To facilitate effective adaptation, signal extraction from traffic flow data is a critical first step (Chen et al., 2024b; Liu et al., 2024b; Li et al., 2024a). Table 1 summarizes the features extracted by various studies. As shown in the table, 95% the listed Traffic-LLMs extracted time series features as input for the LLMs as reasoning hints, which proves the effectivenss and necessity of traffic feature extraction. These features can be broadly categorized into temporal and spatial types. Temporal features focus on historical trajectory patterns (Peng et al., 2024; Feng et al., 2024), while spatial features emphasize location information and sensor dependencies.

Additionally, some studies have incorporated

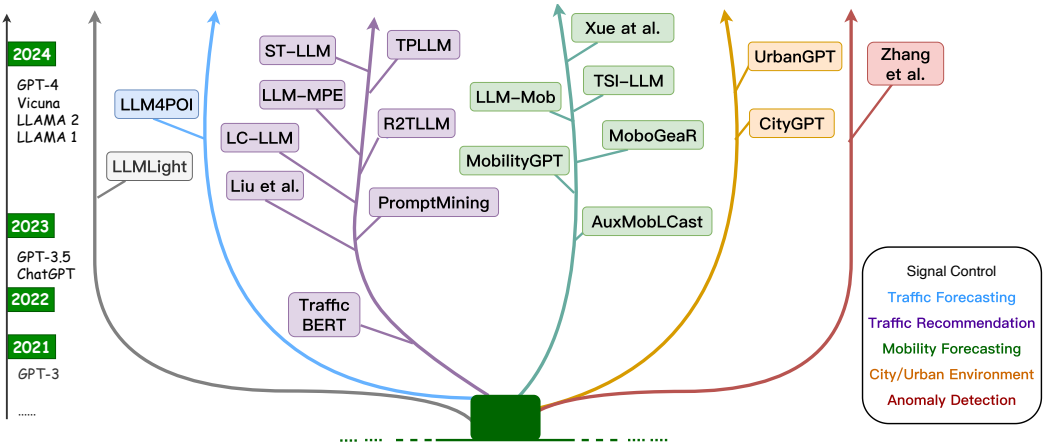

Figure 1: Timeline of Traffic-LLMs

Table 1: Overview of Traffic-LLMs.

Task	Model	Fndn. LLM	Time Series Feature	Prompting	CoT	Finetuned
Traffic Forecasting	Traj-LLM (Lan et al., 2024)	GPT-2	Agent and Lane Information (MLP, GRU)	×	×	LoRA
	LLM-MPE (Liang et al., 2024a)	GPT-4	Event Descriptions	✓	✓	×
	TPLLM (Ren et al., 2024)	GPT-4	Temporal Patterns (CNN), Spatial Dependencies (GCN)	×	×	✓
	R2TLLM (Guo et al., 2024)	Llama2-7B-Chat	Dates, Holidays, Weather, Temperature, Vehicle Visibility	✓	✓	LoRA
	TrafficBERT (Jin et al., 2021)	BERT	Weekday	×	×	Supervised Training
	GenFollower (Chen et al., 2024b)	GPT-4	Spacing, FV Speed, LV Speed, Relative Speed	✓	✓	Supervised Training
	PromptMining (Xue et al., 2024)	GPT-2	Region Information, Opening Information, Hourly Visit, Historical values	✓	✓	HuggingFace
	ST-LLM (Liu et al., 2024a)	GPT-2	Hour-of-Day, Day-of-Week	×	×	PFA
	Liu et al. (Liu et al., 2024b)	GPT-3.5, GPT-4	POI Observation, Category Description	✓	×	×
	LC-LLM (Peng et al., 2024)	Llama-2-7B-Chat	Map Information, Target Vehicle Info, Historical Trajectory, Surrounding Vehicle Information	✓	✓	LoRA
Traffic Recommendation	LLM4POI (Li et al., 2024a)	Llama2-7B	Check-in Records	✓	×	LoRA
	LLMmove (Feng et al., 2024)	GPT-3.5	Check-in Records, Long-term Preference, Current Preference Geospatial Distance, Sequential Transition	✓	×	×
Traffic Anomaly Detection	Zhang et al. (Zhang et al., 2023)	GPT-3.5, GPT-4, Claude-2	-	✓	✓	×
Mobility Forecasting	AuxMobLCast (Xue et al., 2022)	BERT, RoBERTa, GPT-2	Daily Visit Counting Records, Category Information	✓	×	HuggingFace
	Xue et al. (Xue et al., 2024)	GPT-2	POI Observation, Category Description	✓	✓	HuggingFace
	LLM-Mob (Wang et al., 2023b)	GPT-3.5	Historical Stays, Conext Stays	✓	✓	×
	MobiGeaR (Shao et al., 2024)	GPT-3.5	Physical Locations (Mechanistic Model)	✓	✓	×
	MobilityGPT (Haydari et al., 2024)	GPT-2	Historical Trajectory	✓	×	Reinforcement Learning
	TSI-LLM (Luo et al., 2024)	GPT-4	POI Observation, Category Description, Spatial Temporal Dependencies	✓	✓	×
City Environment	CityGPT (Xu et al., 2023)	ChatGPT	Road Database, POI Database, AOI Database, Public Information	✓	✓	DPO
	UrbanGPT (Li et al., 2024b)	Vicuna-7B	Spatial Temporal Dependencies, Time information	✓	×	Instruction Tuning
Signal Control	LLMLight (Lai et al., 2023)	Llama-2-13B, Llama-2-70B, ChatGPT-3.5, GPT-4	Vehicle Counts	✓	✓	LoRA

external knowledge to enhance LLMs' understanding and decision-making capabilities. For instance, CityGPT (Xu et al., 2023) integrates five distinct databases along with a domain-specific knowledge graph, UrbanKG, to aid LLM reasoning. These databases provide detailed information such as road geometry, number of lanes, speed limits, land use types, population, and road connectivity.

3.2.2 Spatio-Temporal Prompting

To further bridge the gap between time series data and language models, most of Traffic-LLMs adopt prompt learning techniques (Da et al., 2024; Xue et al., 2024; Yan et al., 2024; Ju et al., 2024b). The purpose of introducing prompting is to assist LLMs in better understanding the current task setup, dataset characteristics, observed or extracted time series features and even supplementary domain knowledge. The general prompting template of Traffoc-LLMs is summarized in Figure. 3. As shown in the figure, the prompting template of Traffic-LLMs differs from that of general LLMs, incorporating more specific traffic properties and detailed domain knowledge.

To enhance the logical reasoning capabilities of LLMs and achieves better results in complicated traffic series predictions, many works (Xue et al., 2024; Wang et al., 2023b; Xu et al., 2023) introduce the Chain-of-Thought (CoT) (Wei et al., 2022) and Plan-and-Solve (PS) (Wang et al., 2023a) techniques. CoT prompting have been proven to improve LLMs' reasoning capabilities in hard tasks, which inspire models to think from shallow to deep. For example, MobiGeaR (Shao et al., 2024) proposes the decomposed CoT, decomposes the whole process into several steps of individual predictions. As shown in Table 1, 86% of the models utilize prompting, with 63% of them employing the CoT, which facilitates a comprehensive utilization of the sparse traffic data.

3.2.3 Spatio-Temporal Embedding

Compared to general time series LLMs like Time-LLM (Jin et al., 2023a) and TEMPO (Cao et al., 2023), Traffic-LLMs adopt a simpler approach for embedding and tokenization, primarily utilizing one or more projection layers on the input spatio-temporal data X_t. Note that the concept of token embedding in Traffic-LLMs differs from its use in NLP models. In NLP models, each word (or token) is associated with a distinct embedding. In contrast, although many Traffic-LLM papers mentioned the concept of *tokenization*, they primarily refers to the process of embedding input features, rather than assigning unique embeddings to individual tokens as seen in NLP.

3.2.4 Text and Spatio-Temporal Features Alignment

LLMs are typically pre-trained on text corpora, making them less sensitive to spatial or temporal traffic patterns (Ju et al., 2024a; Wang et al., 2024a; Long et al., 2024). To enhance the understanding of such data, it is crucial to align textual information with spatio-temporal traffic patterns. By integrating contextual inputs from both textual and spatio-temporal domains, we can capture complementary information and achieve a higher level of semantic understanding (Ren et al., 2024; Ju et al., 2024a; Li et al., 2024b). To this end, TPLLM (Ren et al., 2024) introduces a 1-D convolutional layer with filters designed for processing temporal traffic data. Regarding spatio-temporal data, only two studies have been conducted: ST-LLM (Liu et al., 2024a) and UrbanGPT (Li et al., 2024b). Both studies employ a two-dimensional convolutional kernel to project spatio-temporal data into the dimensions required for the LLM.

After gaining the fused spatio-temporal embeddings, the special tokens are incorporated into the promptings. For example, UrbanGPT (Li et al., 2024b) proposes spatio-temporal tokens, and corresponding special identifiers, which can be formally defined as: "<start>, <placeholder>, ..., <placeholder>, <end>". <placeholder> denotes the spatio-temporal token, which corresponds to the hidden embedding layer. <start> and <end> serve as identifiers marking the beginning and end of spatial and temporal embeddings.

3.3 Spatio-Temporal Instruction-Tuning

After constructing the dataset, some works fine-tuned their foundation LLMs to better adapt the downstream tasks. As shown in Table 1, 68% of the listed models conduct apply the fine-tuning for adaptation, and employ various strategies, including Low-Rank Adaptation (LoRA) (Hu et al., 2021a), Direct Preference Optimization (DPO) (Rafailov et al., 2024), Partially Frozen Attention (PFA) (Liu et al., 2024a), supervised learning (Jin et al., 2021) and Reinforcement Learning (Haydari et al., 2024). The fine-tuning strategy of "HuggingFace" listed in Table 1 denotes the paper use standard Trainer provided by Hug-

gingFace[1], and do not mention specific fine-tuning method.

LLMs are typically pre-trained with a multi-class classification loss to predict vocabulary, producing a probability distribution of potential outcomes. However, this differs from the continuous value distribution required for regression tasks (Li et al., 2024b). To improve LLMs' understanding and performance on specific traffic tasks, domain-specific objectives are often incorporated. For instance, in traffic flow prediction tasks, a regression loss is used as the objective function to better align the model with the continuous nature of the data (Ren et al., 2024; Liu et al., 2024a), and the classification loss is employed for traffic recommendation tasks (Li et al., 2024a; Feng et al., 2024). The regression loss (absolute error) is a typical for traffic flow prediction tasks (Long et al., 2024; Fang et al., 2021), which allows LLMs to effectively tackle the unique challenges of each task, leading to more accurate predictions across various traffic scenarios.

3.4 Reinforcement Learning

LLMs have become essential tools for enhancing decision-making in transportation. LLMLight (Lai et al., 2023) integrates LLMs as agents for Traffic Signal Control (TSC), which treats TSC as a partially observable Markov game. Each agent manages traffic lights at intersections, controlling signals based on real-time data and scenario descriptions. LightGPT model is then employed to utilize reinforcement learning (RL) to optimize LLMLight. The subsequent feedback and reasoning trajectory is treated as an expected response, minimizing the Negative Log-Likelihood (NLL). For policy refinement, LightGPT employs a ranking feedback loss with a boundary-constrained term (RBC) to guide decision-making towards more effective control actions, enhancing the effectiveness of the model's policy.

Another model, MobilityGPT leverages these insights within RL from the Trajectory Feedback (RLTF) framework. It begins with a full trajectory, using the first segment as a prompt for predicting possible continuations, which are then evaluated and optimized through a Proximal Policy Optimization (PPO). This approach harnesses RL to synthesize and optimize trajectories without manual feedback, making it a robust tool for ITS and urban planning simulations.

[1]https://huggingface.co/

4 Applications

This section discusses the typical applications of LLMs in traffic mining, which includes traffic forecasting, human mobility modeling, city/urban management, and traffic signal control.

4.1 Traffic Forecasting and Mining

Large language models (LLMs) have extensive applications in traffic problems. Traffic forecasting is one of the most basic problems, which aims to forecast the future states of trajectories given extensive historical records. The future states include traffic flows, taxi demand, human mobility, etc (Chen et al., 2024a). There have been extensive works building on LLMs, which can be roughly categorized into tuning-free approaches and tuning-based approaches. Tuning-free approaches focus on designing effective prompts to generate predictions based on the strong zero-shot generalizability of LLMs. They usually keep the LLM frozen, which could enhance the training efficiency. They usually contain a trainable encoder to extract hidden features from time-series or convert time-series into texts based on the fixed template. One of the challenges is how to adapt time-series data to the input of LLMs, and mainstream approaches (Liang et al., 2024a; Wang et al., 2023b; Chen et al., 2024b; Xue et al., 2024) utilize feature engineering based on a learnable encoder and template-based sequence-to-text transformation. In contrast, tuning-based approaches usually train LLMs using fine-tuning and instruction tuning (Lan et al., 2024; Ren et al., 2024; Guo et al., 2024; Jin et al., 2021). Due to the complexity of LLMs, LoRA is typically adopted to enhance the tuning efficiency.

Traffic Flow Forecasting Besides traffic-specific methods, traffic flow forecasting is considered as a regular task in time-series predictions. Therefore, these time-series forecasting approaches can be seamlessly used for traffic predictions. For example, Time-LLM (Jin et al., 2023a) first aligns time series with the text space by prototypical learning, and then feeds both embedding and descriptions into a pre-trained LLM to generate the predictions. It also adopts the dataset context, instruction and data statistics as prompts to enhance the effectiveness of LLMs and achieve strong performance on traffic data. LLMTime (Gruver et al., 2024) proposes to divide each number into a series of digits and utilizes the strong zero-shot extrapolation to finish time-series predictions. It shows the strong

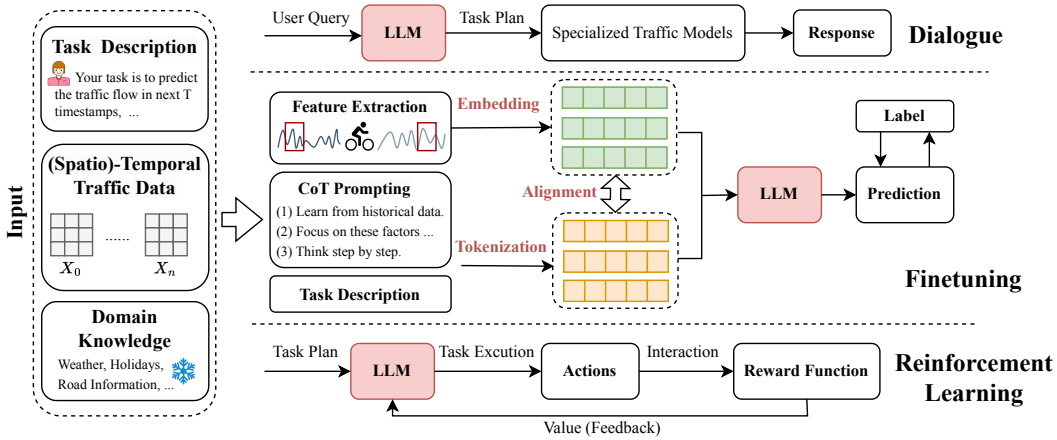

Figure 2: Design Frameworks of Traffic-LLMs

ability of LLMs to represent flexible distributions.

Traffic Recommendation In addition to traffic flow forecasting, there are more works focusing on more generalized traffic analysis including trajectory classification and POI recommendation (Wang et al., 2024b; Peng et al., 2024; Li et al., 2024a; Feng et al., 2024). Trajectory classification aims to predict the discrete future activity instead of continuous regression problems. For example, Liu et al. (Liu et al., 2024b) have made the first attempt to apply LLMs to trajectory classification. They first feed task description, category description as well as POI observation into LLMs to generate the potential activity types. Then, the Bayesian algorithm is adopted for activity inference.

Anomaly Detection Traffic anomaly detection identifies unusual mobility patterns or abnormal trajectories of different entities, which may indicate significant events, such as public gatherings, accidents, or emergencies, by detecting deviations from typical behavior (Smolyak et al., 2020; da Silva et al., 2021). LLMs excel in this area because of their ability to process complex sequential data, capturing patterns and subtle deviations that traditional methods might overlook. Their advanced understanding of context and ability to learn from multiple data sources improves their capability to identify and interpret anomalies, providing valuable insights into potential disturbances or emerging events (Liang et al., 2024a; Su et al., 2024).

4.2 Human Mobility Modeling

Human mobility modeling aims to predict and understand human movement patterns (Barbosa et al., 2018), which are vital for optimizing urban planning, traffic management, and public health (Wang et al., 2019a; Hu et al., 2021b; Zhang et al., 2024c). The integration of time series LLMs (Jin et al., 2023a,b) has significantly advanced this field by processing large datasets to reveal complex temporal patterns and interactions often ignored by traditional models (Gruver et al., 2024; Jin et al., 2024). These advancements offer deeper insights into how various factors influence human mobility (Liang et al., 2024a; Haydari et al., 2024). The application of LLMs in human mobility modeling can be summarized into following groups:

Predictive Modeling focuses on using historical data to predict future human mobility patterns (Sadilek and Krumm, 2012; Baumann et al., 2018). LLMs excel at processing complex sequential data and capturing long-term dependencies, making them highly effective in identifying patterns over time (Xue et al., 2022; Wang et al., 2023b). For instance, AuxMobLCast (Xue et al., 2022) leverages language foundation models to convert numerical and contextual data into natural language prompts for human mobility forecasting, improving POI visitor flow predictions while uncovering sequential mobility patterns through an auxiliary POI classification task. Building on the advancements in language models driven by prompts, Xue et al. (Xue et al., 2024) establish a system for extracting and utilizing prompts in mobility forecasting, focusing on exploring diverse prompt design strategies to improve the transformation of numerical data into actionable insights for human mobility prediction. To fully leverage the strengths of LLMs in capturing temporal dependence and reasoning abilities, LLM-Mob (Wang et al., 2023b) proposes a novel method to utilizes the reasoning capabilities of LLMs to predict human movement trajectories, which addresses the challenges posed by data sparsity and stochastic

daily activities. By incorporating the ideas of historical and contextual stays, LLM-Mob captures long-term and short-term dependencies related to human movement, enabling time-aware and context-inclusive predictions.

Behavioral Analysis in human mobility modeling aims to uncover factors that influence movement patterns, such as demographics, social behavior, and environmental conditions (Chen et al., 2016; Benita, 2021). LLMs enhance this process by processing and synthesizing extensive textual and temporal data, revealing insights into the drivers behind mobility trends and providing a deeper understanding of how various elements impact movement patterns (Shao et al., 2024; Haydari et al., 2024). For instance, MobiGeaR (Shao et al., 2024) introduces a novel framework for mobility generation by treating it as a commonsense reasoning problem, utilizing a context-sensitive chain-of-thoughts prompting strategy to integrate LLMs with behavioral insights, and a divide-and-coordinate technique to combine LLM reasoning with a mechanistic gravity model, thereby generating temporally coherent and contextually relevant mobility patterns. Expanding on the generative capabilities of LLMs, MobilityGPT (Haydari et al., 2024) develops a geospatially-aware generative model that reformulates human mobility modeling as an autoregressive task, leveraging the GPT architecture and a gravity-based sampling method to ensure generated trajectories are semantically realistic. TSI-LLM (Luo et al., 2024) introduces a framework for trajectory semantic inference by leveraging LLMs to interpret and enhance the depth of human mobility data, addressing limitations in understanding complex behaviors and user characteristics through spatio-temporal attributes and context-inclusive prompts.

4.3 City/Urban Management

Effective urban management strategies are essential for improving overall city efficiency and its residents' quality of life. However, the urban environment is a constantly evolving dynamic system with different urban scenarios often presenting unique characteristics and risks. Due to data sparsity and privacy concerns, traditional methods often struggle to adapt to new urban scenarios (e.g., new cities), despite performing well on specific datasets. With their extensive parameters and pre-trained datasets, LLMs offer strong generalization and transfer capabilities, positioning them to address these challenges. Leveraging LLMs in urban management can enhance the general applicability of smart cities, business decision-making, smart marketing, and further building a "city brain".

City/Urban Foundation Models integrate diverse urban data sources to capture the dynamic changes in city life. They provide decision-making support for urban operations, planning, and management, etc. CityGPT(Xu et al., 2023) is an urban foundational model trained on diverse data sources, developed to create embodied agents capable of addressing a wide range of urban challenges. This effort introduces the Urban Generative Intelligence (UGI) platform, which integrates LLMs with multi-source urban data sources to address the complex challenges of modern cities. A suite of generative agents is then developed to focus on areas such as urban mobility, economics, community dynamics, and social issues. These agents are equipped to make high-quality decisions and demonstrate their reasoning and behaviors within realistic simulation environments. UrbanGPT(Li et al., 2024b) is another foundation model designed to predict various urban phenomena. By integrating a spatio-temporal dependency encoder, UrbanGPT effectively aligns LLMs with spatio-temporal contexts. UrbanGPT demonstrated its zero-shot transferability on multiple datasets, including taxi travel, bike trajectories, and crime incident, showing great potential in various urban issues.

Urban planning properly zoning different functional areas in a city ensures the efficient allocation and use of land resources. However, organizing such discussions requires substantial human resources and time, and often results in case-by-case solutions. A multi-agent collaboration framework based on LLMs offers a more efficient and intelligent solution. Zhou et.(Zhou et al., 2024) created thousands of LLM agents to simulate the roles of planners and residents in participatory urban planning discussions. The agents engage in iterative interactions, repeatedly revising the final plan, and using a fishbowl discussion mechanism to accelerate the process. Their multi-agent framework has been deployed in two areas of Beijing, outperforming human experts in some key metrics.

4.4 Traffic Signal Control

Recent advancements in Traffic Signal Control (TSC) leverage reinforcement learning (RL)

methodologies. While effective, these methods face challenges such as limited generalization across diverse traffic conditions and lack interpretability, posing hurdles in understanding their decision-making processes under varying tasks.

LLMLight (Lai et al., 2023) integrates the prompting within an RL context to train the LightGPT model, framing TSC as a partially observable Markov game. It utilizes knowledge-based prompts that guide the LLM in decision-making without the need for parameter adjustments. LLMLight strategically enhances interpretability by providing insights into the reasoning behind traffic signal decisions, thereby optimizing traffic flow through real-time analysis and control. These methodologies not only broaden the scope of LLMs in practical applications but also enrich the TSC domain by combining deep learning insights with intuitive, context-aware reasoning, thus enhancing both transparency of traffic management systems.

5 Undisolved Problem and Future Work

Deep Utilization of Multi-source and Multimodal Traffic Data Although most Traffic-LLMs incorporated external knowledge through prompts (Xue et al., 2024; Wang et al., 2023b; Lai et al., 2023), they still fall short of achieving deep integration of multimodal and multi-source data, leading to insufficient predictive accuracy (Tabatabaie et al., 2023; El-Agroudy et al., 2022). Future research should prioritize the integration of multisource data, such as combining trajectory data with environmental, economic, and health-related indicators, to develop more comprehensive and reliable models.

Interpretability and Theoretical Analysis. Many traffic LLMs are envisioned as components of a "city brain" (Xu et al., 2023; Li et al., 2024b; Haydari et al., 2024). Unlike NLP tasks such as dialogue systems, LLMs in traffic scenarios, such as urban planning and traffic flow forecasting, can have direct and significant impacts on entire cities (Xue et al., 2022; Liu et al., 2024b). These models influence critical factors like population mobility and traffic congestion, affecting the daily lives of residents. As a result, LLM-based intelligent traffic systems must adhere to higher standards of interpretability, robust theoretical foundations, and prediction accuracy compared to other application domains. However, none of the papers listed in Table. 1 explore the interpretability or theoretical analysis.

Privacy Protection. During the fine-tuning stage, the input residential trajectory data can reveal important insights into residents' lifestyles, travel patterns, and consumption behaviors. This data also poses significant risks of misuse, potentially leading to breaches of residents' privacy (Feng et al., 2020). Furthermore, analyzing trajectory data may enable the inference of personally identifiable information, including sensitive details about specific individuals, such as their health or financial status(Ezequiel et al., 2022). Therefore, future work should prioritize the de-identification of resident information to ensure that the data cannot be traced back to specific individuals. Additionally, privacy protection algorithms may be considered to prevent the unintentional disclosure of residents' behavioral trajectories in the predictions of LLMs.

6 Conclusion

This survey offers a comprehensive review of LLM-based traffic forecasting models, drawing from over 100 papers published in the past two years, positioning it as both a pioneering and timely contribution to this field. We provide a comprehensive review of the efforts to bridge the gap between textual LLMs and spatio-temporal traffic data, examining their design principles, improvements in specific traffic scenarios, and benchmark datasets. We further discuss the challenges and highlight future research directions. This survey aims to benefit not only the traffic analysis community but also to the broader advancement of Artificial General Intelligence (AGI).

Limitations

Traffic forecasting is an application-oriented field. This survey focuses on summarizing existing published papers. Since we did not run their code or systems, this survey does not provide an analysis or evaluation from an engineering perspective. At the same time, despite our best efforts to collect relevant works, some studies may be unintentionally overlooked.

Ethics Statement

This paper strictly adhere to the ACL Ethics Policy. The findings and conclusions of our survey are original. We do not propose ethical concerns in this research work.

References

Mahmuda Akhtar and Sara Moridpour. 2021. A review of traffic congestion prediction using artificial intelligence. *Journal of Advanced Transportation*, 2021(1):8878011.

Ishteaque Alam, Dewan Md Farid, and Rosaldo JF Rossetti. 2019. The prediction of traffic flow with regression analysis. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2*, pages 661–671. Springer.

Juan S Angarita-Zapata, Antonio D Masegosa, and Isaac Triguero. 2019. A taxonomy of traffic forecasting regression problems from a supervised learning perspective. *IEEE Access*, 7:68185–68205.

Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. 2018. Human mobility: Models and applications. *Physics Reports*, 734:1–74.

Paul Baumann, Christian Koehler, Anind K Dey, and Silvia Santini. 2018. Selecting individual and population models for predicting human mobility. *IEEE Transactions on Mobile Computing*, 17(10):2408–2422.

Francisco Benita. 2021. Human mobility behavior in covid-19: A systematic literature review and bibliometric analysis. *Sustainable Cities and Society*, 70:102916.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. 2023. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. *arXiv preprint arXiv:2310.04948*.

Jie Cao, Zhiyi Fang, Guannan Qu, Hongyu Sun, and Dan Zhang. 2017. An accurate traffic classification model based on support vector machines. *International Journal of Network Management*, 27(1):e1962.

Bokui Chen, Duo Sun, Jun Zhou, Wengfai Wong, and Zhongjun Ding. 2020. A future intelligent traffic system with mixed autonomous vehicles and human-driven vehicles. *Information Sciences*, 529:59–72.

Cynthia Chen, Jingtao Ma, Yusak Susilo, Yu Liu, and Menglin Wang. 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation research part C: emerging technologies*, 68:285–299.

Wei Chen, Yuxuan Liang, Yuanshao Zhu, Yanchuan Chang, Kang Luo, Haomin Wen, Lei Li, Yanwei Yu, Qingsong Wen, Chao Chen, et al. 2024a. Deep learning for trajectory data management and mining: A survey and beyond. *arXiv preprint arXiv:2403.14151*.

Xianda Chen, Mingxing Peng, PakHin Tiu, Yuanfei Wu, Junjie Chen, Meixin Zhu, and Xinhu Zheng. 2024b. Genfollower: Enhancing car-following prediction with large language models. *arXiv preprint arXiv:2407.05611*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Janetta Culita, Simona Iuliana Caramihai, Ioan Dumitrache, Mihnea Alexandru Moisescu, and Ioan Stefan Sacala. 2020. An hybrid approach for urban traffic prediction and control in smart cities. *Sensors*, 20(24):7209.

Longchao Da, Minquan Gao, Hao Mei, and Hua Wei. 2024. Prompt to transfer: Sim-to-real transfer for traffic signal control with prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 82–90.

Tiago Tiburcio da Silva, Rodrigo Francisquini, and Mariá CV Nascimento. 2021. Meteorological and human mobility data on predicting covid-19 cases by a novel hybrid decomposition method with anomaly detection analysis: a case study in the capitals of brazil. *Expert Systems with Applications*, 182:115190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Mohamed El-Agroudy, Hatem Abou-Senna, and Essam Radwan. 2022. Mobility-as-a-service: Simulation of multi-modal operations in low-density cities. *Transportation research record*, 2676(2):235–246.

Castro Elizondo Jose Ezequiel, Martin Gjoreski, and Marc Langheinrich. 2022. Federated learning for privacy-aware human mobility modeling. *Frontiers in Artificial Intelligence*, 5:867046.

Zheng Fang, Qingqing Long, Guojie Song, and Kunqing Xie. 2021. Spatial-temporal graph ode networks for traffic flow forecasting. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 364–373.

Jie Feng, Can Rong, Funing Sun, Diansheng Guo, and Yong Li. 2020. Pmf: A privacy-preserving human mobility prediction framework via federated learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–21.

Shanshan Feng, Haoming Lyu, Caishun Chen, and Yew-Soon Ong. 2024. Where to move next: Zero-shot generalization of llms for next poi recommendation. *arXiv preprint arXiv:2404.01855*.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2024. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36.

Xusen Guo, Qiming Zhang, Junyue Jiang, Mingxing Peng, Hao Frank Yang, and Meixin Zhu. 2024. Towards responsible and reliable traffic flow prediction with large language models. *Available at SSRN 4805901*.

Ammar Haydari, Dongjie Chen, Zhengfeng Lai, and Chen-Nee Chuah. 2024. Mobilitygpt: Enhanced human mobility modeling with a gpt model. *arXiv preprint arXiv:2402.03264*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Andrea Hess, Karin Anna Hummel, Wilfried N Gansterer, and Günter Haring. 2015. Data-driven human mobility modeling: a survey and engineering guidance for mobile networking. *ACM Computing Surveys (CSUR)*, 48(3):1–39.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021a. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Tao Hu, Siqin Wang, Bing She, Mengxi Zhang, Xiao Huang, Yunhe Cui, Jacob Khuri, Yaxin Hu, Xiaokang Fu, Xiaoyue Wang, et al. 2021b. Human mobility data in the covid-19 pandemic: characteristics, applications, and challenges. *International Journal of Digital Earth*, 14(9):1126–1147.

Ping Jiang, Zhenkun Liu, Lifang Zhang, and Jianzhou Wang. 2022. Advanced traffic congestion early warning system based on traffic flow forecasting and extenics evaluation. *Applied Soft Computing*, 118:108544.

Yue Jiang, Xiucheng Li, Yile Chen, Shuai Liu, Weilong Kong, Antonis F Lentzakis, and Gao Cong. 2024. Sagdfn: A scalable adaptive graph diffusion forecasting network for multivariate time series forecasting. *arXiv preprint arXiv:2406.12282*.

KyoHoon Jin, JeongA Wi, EunJu Lee, ShinJin Kang, SooKyun Kim, and YoungBin Kim. 2021. Trafficbert: Pre-trained model with large-scale data for long-range traffic flow forecasting. *Expert Systems with Applications*, 186:115738.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023a. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Ming Jin, Qingsong Wen, Yuxuan Liang, Chaoli Zhang, Siqiao Xue, Xue Wang, James Zhang, Yi Wang, Haifeng Chen, Xiaoli Li, et al. 2023b. Large models for time series and spatio-temporal data: A survey and outlook. *arXiv preprint arXiv:2310.10196*.

Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. 2024. Position: What can large language models tell us about time series analysis. In *Forty-first International Conference on Machine Learning*.

Wei Ju, Yifan Wang, Yifang Qin, Zhengyang Mao, Zhiping Xiao, Junyu Luo, Junwei Yang, Yiyang Gu, Dongjie Wang, Qingqing Long, et al. 2024a. Towards graph contrastive learning: A survey and beyond. *arXiv preprint arXiv:2405.11868*.

Wei Ju, Siyu Yi, Yifan Wang, Qingqing Long, Junyu Luo, Zhiping Xiao, and Ming Zhang. 2024b. A survey of data-efficient graph learning. *arXiv preprint arXiv:2402.00447*.

Siqi Lai, Zhao Xu, Weijia Zhang, Hao Liu, and Hui Xiong. 2023. Large language models as traffic signal control agents: Capacity and opportunity. *arXiv preprint arXiv:2312.16044*.

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*.

Ibai Lana, Javier Del Ser, Manuel Velez, and Eleni I. Vlahogianni. 2018. Road traffic forecasting: Recent advances and new challenges. *IEEE Intelligent Transportation Systems Magazine*, 10(2):93–109.

Chunggi Lee, Yeonjun Kim, Seungmin Jin, Dongmin Kim, Ross Maciejewski, David Ebert, and Sungahn Ko. 2020. A visual analytics system for exploring,

monitoring, and forecasting road traffic congestion. *IEEE Transactions on Visualization and Computer Graphics*, 26(11):3133–3146.

Peibo Li, Maarten de Rijke, Hao Xue, Shuang Ao, Yang Song, and Flora D Salim. 2024a. Large language models for next point-of-interest recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1463–1472.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*.

Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. 2024b. Urbangpt: Spatio-temporal large language models. *arXiv preprint arXiv:2403.00813*.

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

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. 2024b. Foundation models for time series analysis: A tutorial and survey. *arXiv preprint arXiv:2403.14735*.

Thomas Liebig, Nico Piatkowski, Christian Bockermann, and Katharina Morik. 2017. Dynamic route planning with real-time traffic predictions. *Information Systems*, 64:258–265.

Xinjie Lin, Gang Xiong, Gaopeng Gou, Zhen Li, Junzheng Shi, and Jing Yu. 2022. Et-bert: A contextualized datagram representation with pre-training transformers for encrypted traffic classification. In *Proceedings of the ACM Web Conference 2022*, pages 633–642.

Chenxi Liu, Sun Yang, Qianxiong Xu, Zhishuai Li, Cheng Long, Ziyue Li, and Rui Zhao. 2024a. Spatial-temporal large language model for traffic prediction. *25th IEEE International Conference on Mobile Data Management*.

Shuai Liu, Xiucheng Li, Gao Cong, Yile Chen, and Yue Jiang. Multivariate time-series imputation with disentangled temporal representations. In *The Eleventh international conference on learning representations*.

Shuai Liu, Guojie Song, and Wenhao Huang. 2020. Real-time transportation prediction correction using reconstruction error in deep learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(2):1–20.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. 2024c. Autotimes: Autoregressive time series forecasters via large language models. *arXiv preprint arXiv:2402.02370*.

Qingqing Long, Zheng Fang, Chen Fang, Chong Chen, Pengfei Wang, and Yuanchun Zhou. 2024. Unveiling delay effects in traffic forecasting: A perspective from spatial-temporal delay differential equations. In *Proceedings of the ACM on Web Conference 2024*, pages 1035–1044.

Yuxiao Luo, Zhongcai Cao, Xin Jin, Kang Liu, and Ling Yin. 2024. Deciphering human mobility: Inferring semantics of trajectories with large language models. *arXiv preprint arXiv:2405.19850*.

Zhang Mingheng, Zhen Yaobao, Hui Ganglong, and Chen Gang. 2013. Accurate multisteps traffic flow prediction based on svm. *Mathematical Problems in Engineering*, 2013(1):418303.

SBMATSDVBBI Neelakandan, MA Berlin, Sandesh Tripathi, V Brindha Devi, Indu Bhardwaj, and N Arulkumar. 2021. Iot-based traffic prediction and traffic signal control system for smart city. *Soft Computing*, 25(18):12241–12248.

Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2018. Spatial as deep: Spatial cnn for traffic scene understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Jungme Park, Dai Li, Yi L. Murphey, Johannes Kristinsson, Ryan McGee, Ming Kuang, and Tony Phillips. 2011. Real time vehicle speed prediction using a neural network traffic model. In *The 2011 International Joint Conference on Neural Networks*, pages 2991–2996.

Mingxing Peng, Xusen Guo, Xianda Chen, Meixin Zhu, Kehua Chen, Xuesong Wang, Yinhai Wang, et al. 2024. Lc-llm: Explainable lane-change intention and trajectory predictions with large language models. *arXiv preprint arXiv:2403.18344*.

Lilian Pun, Pengxiang Zhao, and Xintao Liu. 2019. A multiple regression approach for traffic flow estimation. *IEEE access*, 7:35998–36009.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Navin Ranjan, Sovit Bhandari, Hong Ping Zhao, Hoon Kim, and Pervez Khan. 2020. City-wide traffic congestion prediction based on cnn, lstm and transpose cnn. *Ieee Access*, 8:81606–81620.

Yilong Ren, Yue Chen, Shuai Liu, Boyue Wang, Haiyang Yu, and Zhiyong Cui. 2024. Tpllm: A traffic prediction framework based on pretrained large language models. *arXiv preprint arXiv:2403.02221*.

Adam Sadilek and John Krumm. 2012. Far out: Predicting long-term human mobility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 814–820.

Chenyang Shao, Fengli Xu, Bingbing Fan, Jingtao Ding, Yuan Yuan, Meng Wang, and Yong Li. 2024. Beyond imitation: Generating human mobility from context-aware reasoning with large language models. *arXiv preprint arXiv:2402.09836*.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 14974–14983.

Daniel Smolyak, Kathryn Gray, Sarkhan Badirli, and George Mohler. 2020. Coupled igmm-gans with applications to anomaly detection in human mobility data. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, 6(4):1–14.

Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. 2020. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 914–921.

Jing Su, Chufeng Jiang, Xin Jin, Yuxin Qiao, Tingsong Xiao, Hongda Ma, Rong Wei, Zhi Jing, Jiajun Xu, and Junhong Lin. 2024. Large language models for forecasting and anomaly detection: A systematic literature review. *arXiv preprint arXiv:2402.10350*.

Yanfeng Sun, Xiangheng Jiang, Yongli Hu, Fuqing Duan, Kan Guo, Boyue Wang, Junbin Gao, and Baocai Yin. 2022. Dual dynamic spatial-temporal graph convolution network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):23680–23693.

Mahan Tabatabaie, Suining He, and Kang G Shin. 2023. Cross-modality graph-based language and sensor data co-learning of human-mobility interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(3):1–25.

Thammasak Thianniwet, Satidchoke Phosaard, and Wasan Pattara-Atikom. 2009. Classification of road traffic congestion levels from gps data using a decision tree algorithm and sliding windows. In *Proceedings of the world congress on engineering*, volume 1, pages 1–3.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sebastian Wandelt, Changhong Zheng, Shuang Wang, Yucheng Liu, and Xiaoqian Sun. 2024. Large language models for intelligent transportation: A review of the state of the art and challenges. *Applied Sciences*, 14(17):7455.

Chengrui Wang, Qingqing Long, Xiao Meng, Xunxin Cai, Chengjun Wu, Zhen Meng, Xuezhi Wang, and Yuanchun Zhou. 2024a. Biorag: A rag-llm framework for biological question reasoning. *arXiv preprint arXiv:2408.01107*.

Jiawei Wang, Renhe Jiang, Chuang Yang, Zengqing Wu, Makoto Onizuka, Ryosuke Shibasaki, Noboru Koshizuka, and Chuan Xiao. 2024b. Large language models as urban residents: An llm agent framework for personal mobility generation.

Jinzhong Wang, Xiangjie Kong, Feng Xia, and Lijun Sun. 2019a. Urban human mobility: Data-driven modeling and prediction. *ACM SIGKDD explorations newsletter*, 21(1):1–19.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024c. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Shuaian Wang, Ran Yan, and Xiaobo Qu. 2019b. Development of a non-parametric classifier: Effective identification, algorithm, and applications in port state control for maritime transportation. *Transportation Research Part B: Methodological*, 128:129–157.

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

Zhumei Wang, Xing Su, and Zhiming Ding. 2020. Long-term traffic prediction based on lstm encoder-decoder architecture. *IEEE Transactions on Intelligent Transportation Systems*, 22(10):6561–6571.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Qiong Wu, Kaiwen He, Xu Chen, Shuai Yu, and Junshan Zhang. 2021. Deep transfer learning across cities for mobile traffic prediction. *IEEE/ACM Transactions on Networking*, 30(3):1255–1267.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. 2020. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 753–763.

Yingjie Xia, Jinlong Chen, Xindai Lu, Chunhui Wang, and Chao Xu. 2016. Big traffic data processing framework for intelligent monitoring and recording systems. *Neurocomputing*, 181:139–146.

Fengli Xu, Jun Zhang, Chen Gao, Jie Feng, and Yong Li. 2023. Urban generative intelligence (ugi): A foundational platform for agents in embodied city environment. *arXiv preprint arXiv:2312.11813*.

Hao Xue and Flora D Salim. 2023. Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*.

Hao Xue, Tianye Tang, Ali Payani, and Flora D Salim. 2024. Prompt mining for language-based human mobility forecasting. *arXiv preprint arXiv:2403.03544*.

Hao Xue, Bhanu Prakash Voutharoja, and Flora D Salim. 2022. Leveraging language foundation models for human mobility forecasting. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–9.

Yuchen Yan, Peiyan Zhang, Zheng Fang, and Qingqing Long. 2024. Inductive graph alignment prompt: Bridging the gap between graph pre-training and inductive fine-tuning from spectral perspective. In *Proceedings of the ACM on Web Conference 2024*, pages 4328–4339.

Chuanting Zhang, Haixia Zhang, Jingping Qiao, Dongfeng Yuan, and Minggao Zhang. 2019. Deep transfer learning for intelligent cellular traffic prediction based on cross-domain big data. *IEEE Journal on Selected Areas in Communications*, 37(6):1389–1401.

Dingkai Zhang, Huanran Zheng, Wenjing Yue, and Xiaoling Wang. 2024a. Advancing its applications with llms: A survey on traffic management, transportation safety, and autonomous driving. In *International Joint Conference on Rough Sets*, pages 295–309. Springer.

Hanyuan Zhang, Hao Wu, Weiwei Sun, and Baihua Zheng. 2018. Deeptravel: a neural network based travel time estimation model with auxiliary supervision. *arXiv preprint arXiv:1802.02147*.

Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Siyao Zhang, Daocheng Fu, Wenzhe Liang, Zhao Zhang, Bin Yu, Pinlong Cai, and Baozhen Yao. 2024b. Trafficgpt: Viewing, processing and interacting with traffic foundation models. *Transport Policy*, 150:95–105.

Zheng Zhang, Hossein Amiri, Zhenke Liu, Andreas Züfle, and Liang Zhao. 2023. Large language models for spatial trajectory patterns mining. *arXiv preprint arXiv:2310.04942*.

Zijian Zhang, Yujie Sun, Zepu Wang, Yuqi Nie, Xiaobo Ma, Peng Sun, and Ruolin Li. 2024c. Large language models for mobility in transportation systems: A survey on forecasting tasks. *arXiv preprint arXiv:2405.02357*.

Zheng Zhao, Weihai Chen, Xingming Wu, Peter CY Chen, and Jingmeng Liu. 2017. Lstm network: a deep learning approach for short-term traffic forecast. *IET intelligent transport systems*, 11(2):68–75.

Fan Zhou, Qiang Gao, Goce Trajcevski, Kunpeng Zhang, Ting Zhong, and Fengli Zhang. 2018. Trajectory-user linking via variational autoencoder. In *IJCAI*, pages 3212–3218.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pages 27268–27286. PMLR.

Zhilun Zhou, Yuming Lin, Depeng Jin, and Yong Li. 2024. Large language model for participatory urban planning.

In the appendix, we provide supplementary discussions on the following topics: (A) Technical comparisons of traffic LLMs and their related techniques, (B) Background knowledge of traffic LLMs, and (C) Benchmark datasets and evaluation methods.

A Comparision with Related Techniquies

This section presents a comparative analysis of LLMs for traffic forecasting against relevant existing technologies to elucidate their distinctive features and advantages. We conduct three primary comparisons: **(A.1)** Traffic-LLMs versus traditional supervised learning; **(A.2)** Zero-shot Traffic-LLMs in relation to transfer learning; **(A.3)** Traffic-LLMs in relation to pretraining and fine-tuning traffic models. These comparisons aim to provide a comprehensive understanding of traffic-LLM's position within the current technological landscape.

A.1 Comparison with Supervised Learning

Traditional supervised learning methods are extensively applied in urban and transportation domains, relying primarily on large volumes of labeled data, with each training sample comprising input and corresponding accurate ground truth outputs. The main advantages of these methods are their mature theoretical foundations and straightforward implementations. They have been effectively utilized in various transportation applications such as traffic flow prediction (Liu et al., 2020; Jiang et al., 2024), travel time estimation (Zhang et al., 2018), traffic state imputation (Liu et al.), transportation state control (Wang et al., 2019b), and dynamic route analysis (Liebig et al., 2017). Initially, techniques like regression analysis, decision trees, and support vector machines (SVM) were commonly employed. In transportation, regression analysis is typically used to predict continuous outputs, such as traffic flow or travel speed, based on various input features like time of day, day of the week, weather conditions, and historical traffic data (Alam et al., 2019; Pun et al., 2019). Decision trees are employed for making sequential, rule-based decisions on traffic data, aiding in the classification of traffic conditions (e.g., congestion levels, traffic efficiency) or predicting the likelihood of traffic events (Thianniwet et al., 2009). SVMs operate by identifying the hyperplane that best separates different categories, commonly used for classification tasks in transportation such as recognizing traffic signal indications or different driving patterns (Cao et al., 2017). SVMs can also be extended to regression problems (known as Support Vector Regression, SVR) to predict traffic flow or speed (Mingheng et al., 2013).

With the rise and development of deep learning in recent years, complex architectures have been developed and integrated into various downstream tasks in the transportation sector. Encoder-decoder models are highly effective in sequence generation and prediction tasks, such as generating subsequent user trajectories (Zhou et al., 2018) or estimating future traffic volumes based on past data (Liu et al., 2020). The encoder part of the model processes input sequences (such as historical traffic data) to produce hidden representations that encapsulate input sequences, while the decoder uses these representations to generate future trajectories or predicted traffic attributes. Recurrent Neural Networks (RNNs), such as LSTM and GRU, are extensively used for traffic state prediction and trajectory-related tasks due to their excellent capability to capture temporal dependencies, like modeling historical traffic dynamics and capturing past traffic patterns (Zhao et al., 2017; Wang et al., 2020). Convolutional Neural Networks (CNNs), which have shone in computer vision tasks, are equally effective in transportation tasks such as raster-based traffic speed prediction and gridded city traffic modeling (Pan et al., 2018; Ranjan et al., 2020). CNNs can efficiently handle spatial relationships in data, better extracting spatial dependencies in traffic-related data. However, not all traffic-related tasks can be gridded, with urban road networks being a typical example where CNNs struggle to model spatial dependencies effectively. Consequently, modeling the spatial attributes of traffic data from a graph perspective has emerged and excelled. Transportation networks are naturally represented as graphs, with intersections and roadways forming nodes and edges. Graph Convolutional Networks (GCNs) leverage this structure to predict traffic conditions by learning from the network topology, integrating the influence of neighboring nodes to enhance prediction accuracy (Li et al., 2017; Song et al., 2020). More advanced models have further improved GCN performance through hypergraphs and dynamic graphs (Wu et al., 2020; Sun et al., 2022). Transformers process sequences using self-attention mechanisms, assessing the importance

of different parts of the input data. Compared to RNNs, they offer significant improvements in modeling long-term dependencies and great efficiency in parallel computing capabilities, making them increasingly favored for transportation-related tasks in recent years. Tasks such as long-term traffic speed prediction and long-term trajectory pattern modeling have seen significant contributions from Transformer models (Zhou et al., 2021, 2022).

Although traditional methods have achieved certain successes in transportation-related tasks, they exhibit a strong dependency on training data. To achieve high prediction accuracy, extensive data for training is required. This includes the collection, cleaning, normalization, and intricate processes of feature selection and extraction (Xia et al., 2016). Moreover, a great proportion of traffic related supervised models need to consider various influencing factors such as cyclical time patterns, weather conditions, holiday impacts, spatial and semantic distributions (Zhang et al., 2017), all of which demand significant time and effort for precise feature construction. At the same time, the transportation system is a task-specific and dataset-specific environment. Supervised learning traffic-related methods show significant limitations when dealing with dynamic changes and biases caused by different tasks and datasets (Angarita-Zapata et al., 2019). These methods generally assume a static data distribution, but in practical applications, traffic patterns change with various factors such as time, weather, and spatial reasons, making it difficult for models to adapt to new or unseen traffic conditions. Furthermore, these models typically perform well only on specific tasks and datasets, but may yield poor results on different datasets, and even within the same city across different times, they cannot guarantee good performance.

Compared to traditional supervised learning methods, modern Large Language Models (LLMs) exhibit outstanding transfer capabilities in handling downstream transportation tasks due to the vast world knowledge embedded during the pre-training phase. These models, through extensive text data acquired during pre-training, encompass knowledge on a wide range of topics and scenarios, including geographic and environmental factors related to transportation. This rich background knowledge enables LLMs to better understand and predict complex traffic situations and dynamics. When employing LLMs for transportation tasks, there are two main approaches: fine-tuning and direct inference (i.e., through prompting). Through fine-tuning, the model can focus on the specific requirements of a particular task, such as traffic flow prediction, while retaining the broad knowledge learned during the pre-training phase On the other hand, direct inference or using LLMs through prompting does not require additional training steps. In this approach, users can guide the model to produce the required information or predictions by constructing specific questions or commands (i.e., prompts). For example, one might use the model to predict traffic conditions by submitting a natural language question like, "What is the traffic situation in the city center at 8 AM tomorrow?". The advantages of these LLM-based traffice-related methods include rapid deployment and flexibility, making it particularly suitable for tasks that require quick responses to new situations or where only limited data is available. Moreover, the use of LLMs also reveals their potential contributions to traffic management and planning. The model can assist decision-makers by providing insights and recommendations based on complex data analysis of historical and real-time traffic data. This not only improves the quality and speed of decision-making but also reduces reliance on specialized traffic analysts, thereby lowering costs and enhancing operational efficiency.

A.2 Comparison with Transfer Learning

In the field of urban and transportation, traditional methods typically require training models from scratch for each specific predictive task. This process not only consumes considerable time and resources but also poses significant challenges in data collection, cleaning, and engineering. Transfer learning, as an emerging training method, effectively addresses these issues and enhances the performance of traffic prediction models. This approach involves transferring the knowledge of a well-trained model from one task to another related task. By leveraging existing experience and knowledge, models can tackle new problems, thereby reducing the required training time and data volume. For instance, traffic models trained in data-rich major cities like New York and London can be used to predict traffic conditions in smaller or emerging cities where traffic data is scarce. Additionally, traffic patterns from one time period can be transferred to another, such as from weekdays to weekends or

from summer to winter, thus improving prediction accuracy across different times (Zhang et al., 2019; Wu et al., 2021).

However, despite its effectiveness in utilizing existing models under resource constraints, transfer learning faces several challenges in the transportation domain. Primarily, it relies on a high degree of similarity between the source and target tasks, but traffic patterns can vary significantly across different regions. If the source data is too dissimilar from the target environment, the effectiveness of transfer learning may be substantially reduced. Secondly, transfer learning models are typically optimized for specific datasets. When changes occur in the traffic environment, such as the opening of new roads or changes in traffic regulations, these models may require retraining or significant adjustments to remain effective.

In contrast, approaches based on large pre-trained models (LLMs) can effectively circumvent these issues. LLMs, having been exposed to diverse data during the pre-training phase, understand how various factors from weather changes to city activities can influence traffic flows. This capability allows LLMs to not only rely on historical experiences of a single city for traffic prediction and management but also utilize information from other cities and time periods in the pre-trained corpus to make reasonable guesses and responses to unseen situations. Moreover, unlike transfer learning models that are trained for a single task, LLMs are designed with versatility, allowing them to handle multiple types of traffic tasks simultaneously. From route planning to traffic flow prediction, LLMs can operate within a unified framework, reducing the need to develop and adjust separate models for each specific task.

A.3 Comparison with Pretraining and Finetuning Methods

In recent years, with the rapid advancement of natural language processing (NLP) technology, especially the successful application of pre-trained models such as BERT and GPT, researchers have begun exploring the potential of applying these advanced model technologies to the transportation field. These pre-trained models are trained on vast urban databases, acquiring extensive traffic knowledge, thus providing a solid foundation for a variety of tasks. The significant advantage of this approach lies in its robust multi-task processing capability and adaptability to different data types. For instance, a pre-trained model can be used simultaneously to predict urban traffic flow, analyze accident hotspots, and forecast road conditions, all by appropriately adjusting the same model. This capability greatly enhances the model's application flexibility and economic efficiency, as the same model can be applied to various scenarios without the need for training from scratch. Specifically, this technology allows models to understand and predict traffic conditions in one city by learning from the traffic data of another city. By training models on traffic data from different cities like New York, Beijing, or London, the models can capture the unique traffic patterns of each city while using the general knowledge acquired during pre-training to identify commonalities between different cities. This enables models to be quickly deployed in new cities, making accurate predictions even with relatively scarce traffic data from those cities (Lin et al., 2022; Jin et al., 2021).

Despite the immense success of pre-trained models in NLP, there are some notable challenges and limitations when applying them to the transportation field:

- Dataset Size: Compared to the NLP field, transportation datasets are typically smaller in scale. Large language models like BERT or GPT are trained on massive text datasets containing rich language structures and semantic information. However, high-quality transportation data is often harder to obtain and relatively limited in quantity, which may restrict the model's ability.

- Training Time and Resources: Pre-trained models, especially large ones, require substantial computational resources and time for training. In the transportation field, this demand can lead to high costs and long training cycles.

- Specificity Issues: Transportation data has high geographic and temporal specificity. Cities with abundant traffic data are mostly developed cities, whose traffic characteristics and travel patterns may not apply to less developed or sparsely populated areas.

Given these challenges, fine-tuning or direct inference through designed prompts of pre-trained large language models (LLMs) brings significant

advantages to the transportation field. Fine-tuning or inferring these pre-trained models allows them to quickly adapt to specific traffic scenarios and needs. This approach enables the model to retain its broad language and knowledge processing capabilities while optimizing for specific tasks, significantly shortening training and deployment times. Pre-trained models have already learned a vast amount of general knowledge during training, which means they can effectively improve performance even in data-scarce environments. This reduces reliance on large-scale specialized traffic data. Furthermore, since LLMs are already pretrained, fine-tuning or direct inference will not require as much memory as pretrained traffic models need.

B Background

This section provides an overview of the foundational LLMs employed for traffic forecasting **(B.1)**. We then introduce the most straightforward approach **(B.2)**, in dialogue mode, for integrating LLMs into traffic analysis, offering it as a basic design pipeline.

B.1 Foundation LLMs

Foundation language models are the core components of Traffic-LLMs. Since 2021, researchers have experimented with a range of mainstream LLMs, from BERT to ChatGPT (see Table 1). BERT(Devlin et al., 2019) is a bidirectional encoder model excelling in understanding context and language representation; RoBERTa(Liu et al., 2019) is an optimized version of BERT; GPT-2(Radford et al., 2019) is a powerful generative model with impressive text generation capabilities; GPT-3.5 and GPT-4 are more advanced in dialogue generation and complex task handling, particularly in multimodal capabilities; Claude, developed by Anthropic, focuses on safety and bias reduction; LLaMA(Touvron et al., 2023) is a lightweight model optimized for computational efficiency, ideal for dialogue tasks; and Vicuna(Chiang et al., 2023) is a fine-tuned dialogue model based on LLaMA.

Due to ChatGPT's strong performance and influence, most Traffic-LLMs studies (17 out of 22 papers) now predominantly utilize OpenAI's GPT series models as foundation, while the application of BERT models has primarily been limited to research before 2022. Additionally, since ChatGPT is not open-source, researchers often use prompts or chain-of-thought techniques to leverage GPT-3.5 and GPT-4. For methods requiring fine-tuning of foundation LLMs, researchers tend to prefer community-friendly open-source models like LLaMA and Vicuna (5 papers) or GPT-2, the open-source model from the GPT series (6 papers).

B.2 Forecasting in Dialogue Mode

The most straightforward way to apply LLMs in traffic time series analysis is by interacting with the models through dialogue. This method fully leverages the text generation and conversational capabilities of LLMs, offering a user-friendly experience. However, a significant challenge lies in the fact that the pre-training process does not equip LLMs with the ability to understand traffic-related complexities. Specifically, LLMs struggle to interpret numerical inputs, which are common in traffic flow prediction, and providing specialized advice and accurate predictions remains a key consideration when using general-purpose language models for domain-specific problems.

PromptCast(Xue and Salim, 2023) offers a "code-less" approach by directly feeding the problem description and input data to the LLMs in a natural language manner, without altering the model's architecture to achieve the desired response. They fine-tune the models in a supervised learning way, aiming to equip them with the ability to understand and solve practical problems. Experiments conducted on tasks such as city temperature, electricity consumption load, and human mobility prediction yielded results comparable to state-of-the-art (SOTA) models, demonstrating that LLMs can grasp traffic-related issues and perform simple predictive tasks. However, PromptCast's prompts rely on fixed templates and lack the ability to produce multi-turn dialogue. The effectiveness of the fine-tuned models may largely come from the LLM's strong fitting and memorization abilities, rather than a deep exploration of the model's valuable understanding and reasoning capabilities.

TrafficGPT(Zhang et al., 2024b) presents a more refined approach to utilizing LLMs. While it still employs natural language interaction, it addresses specialized traffic problems by incorporating dedicated traffic foundation models to provide predictions and decision-making. In this setup, LLMs function as a "account manager", acting as a bridge between the user and these specialized models. TrafficGPT attempts to understand the user's needs, breaks down complex traffic problems into smaller,

Instruction | Task Description
You are an expert traffic ... model, that can predict future ... values according to spatial and temporal information ...

Data | Describe the Input Data
You will be provided with <history>, which is a list containing ... The dataset contains ... timestamps, the interval is ...

Instruction | Domain Knowledge
[Traffic Volume]: the number of vehicles . . .
[Traffic Period]: The recording time . . .
[Traffic Pattern]: Traffic flow patterns in a city are influenced by various area attributes . . .
[Spatial-temporal Features]: Traffic flow in an area will be affected by its nearby infrastructures, during specific periods for different areas. . . This region is located. . .
[Traffic Tokens]: Given the historical data and the corresponding tokens <...> for taxi flow. . . Please generate the predictive tokens for . . .
[External Factors]: The weather is ..., the date is

Instruction | Chain of Thought (CoT) for guiding LLMs to "think"
Your answer must considering the following rules:
(1) Learn from similar historical data.
(2) Focus on these factors: ...
(3) Think step by step before making the final prediction.

Instruction | Format the Outputs and Ask for Explanations
Please format your answer in a JSON object containing the following keys: ...

Figure 3: The general prompting template of Traffic-LLMs to bridge the gap between LLMs and spatio-temporal traffic data.

more manageable tasks, assigns them to the appropriate traffic models for processing, and then compiles and presents the results back to the user. Additionally, it can further adjust the output based on user feedback.

Interacting with LLMs in chat mode offers a flexible way to utilize these models. However, converting traffic data into text inevitably leads to some loss of information, making it necessary to fine-tune the models or integrate specialized models to ensure effective outputs.

C Benchmark Datasets and Evaluation Methods

This section summarizes the benchmark datasets **(C.1)** and evaluation standards **(C.2)**.

C.1 Benchmark Datasets

We surveyed datasets used in LLMs for various traffic flow forecasting problems. Table 2 provides a comprehensive overview of widely-used datasets for traffic forecasting, mobility prediction, and related tasks. These datasets span multiple domains, including mobility, traffic, trajectory prediction, and signal control, offering a diverse foundation for the application of large language models (LLMs) in urban planning and transportation research.

In the mobility domain, datasets such as NYC, Dallas, and GeoLife capture patterns of human movement across different urban environments. These datasets are commonly used to model individual and collective mobility behaviors, allowing researchers to predict future movement based on past trajectories. Traffic datasets, like NYC-Taxi, CHI-Taxi, and PeMS04, provide detailed traffic flow information, including taxi inflow-outflow data and traffic sensor data, which are critical for building models that predict congestion, optimize traffic management, and improve urban transportation systems. In the trajectory prediction domain, datasets such as nuScene, GeoLife, and SafeGraph Human Mobility Data are central to tasks like route prediction, where understanding an individual's future movements based on past behavior is essential. These datasets include detailed GPS traces and allow for the modeling of spatial-temporal patterns. The signal control domain encompasses datasets from cities such as Jinan, Hangzhou, and New York, where traffic signal data is collected to study and optimize the flow of vehicles through urban intersections. Finally, other datasets like Waymo

Table 2: Detailed information of datasets

Domain	Dataset	Source
Traffic	NYC-Taxi	ST-LLM (Liu et al., 2024a), UrbanGPT (Li et al., 2024b),LLM-MPE (Liang et al., 2024a)
	CHI-Taxi	UrbanGPT (Li et al., 2024b)
	NYC-Bike	UrbanGPT (Li et al., 2024b)
	PeMS04	TPLLM (Ren et al., 2024)
	PeMS08	TPLLM (Ren et al., 2024)
	CATraffic	R2TLLM (Guo et al., 2024)
	Metr-LA	TrafficBERT (Jin et al., 2021)
	CHBike	ST-LLM (Liu et al., 2024a)
Mobility	NYC	AuxMobLCast (Xue et al., 2022)
	Dallas	AuxMobLCast (Xue et al., 2022)
	Miami	AuxMobLCast (Xue et al., 2022), PromptMining (Xue et al., 2024)
	Tencent	LLM-Mob (Wang et al., 2023b), MobiGeaR (Shao et al., 2024)
	ChinaMobile	LLM-Mob (Wang et al., 2023b), MobiGeaR (Shao et al., 2024)
	GeoLife	LLM-Mob (Wang et al., 2023b)
	NYC-FSQ	LLM-Mob (Wang et al., 2023b)
	Porto	MobilityGPT (Haydari et al., 2024)
	Beijing	MobilityGPT (Haydari et al., 2024)
City / Urban Management	Huilongguan Region Planning	CityGPT (Zhou et al., 2024)
	Dahongmen Region Planning	CityGPT (Zhou et al., 2024)
	City Temperature	PromptCast (Xue and Salim, 2023)
	Electricity Consumption Load	PromptCast (Xue and Salim, 2023)
Trajectory	nuScene	Traj-LLM (Lan et al., 2024)
	Geolife	LLM-Mob (Wang et al., 2023b)
	POL	LLM-Mob (Wang et al., 2023b)
	HighD	LC-LLM (Peng et al., 2024)
	Sampled from OSM	(Liu et al., 2024b)
	SafeGraph Human Mobility Data	PromptCast (Xue and Salim, 2023)
	Tokyo Personal Activity	LLMob (Wang et al., 2024b)
Signal Control	Jinan	LLMLight (Lai et al., 2023)
	HangZhou	LLMLight (Lai et al., 2023)
	New York	LLMLight (Lai et al., 2023)
	High-traffic	LLMLight (Lai et al., 2023)

expand the scope of LLM-based forecasting into domains beyond traffic, showing the versatility of these models in handling various types of urban data. Together, these datasets provide a robust and diverse benchmark for evaluating the effectiveness of large language models in traffic and mobility forecasting, enabling researchers to tackle a wide range of urban challenges.

C.2 Evaluation Criterion

Different methods use different ways to evaluation their effectiveness. In this section, we collect the metrics each method uses for evaluation, as shown in Table 3 This table provides a summary of the evaluation metrics used across different models in the context of mobility prediction, traffic forecasting, trajectory prediction, and signal control. These metrics serve to evaluate LLMs' and other approaches' performance in addressing various urban challenges.

In the mobility domain, models like GenFollower, PromptMining, and LLM-Mob evaluate their results using common metrics including Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and accuracy measures (Acc@1, Acc@5, Acc@10). Other advanced metrics, such as Weighted F1, nDCG, and JSD, are also used to capture more nuanced aspects of prediction, including ranking accuracy and distributional similarity. MobiGear and MobilityGPT, on the other hand, introduce domain-specific metrics like intent distance, LocFreq, and ODSim, highlighting the need for tailored metrics in mobility-based models. For the traffic domain, models like UrbanGPT and TrafficBERT commonly rely on metrics such as MAE, RMSE, and MAPE (Mean Absolute Percentage Error) to measure the accuracy of traffic flow predictions. More complex metrics like Macro-F1 and R^2 are also applied to capture the robustness of model predictions across different traffic patterns and scenarios. In the trajectory prediction domain, models such as Traj-LLM and LC-LLM focus on metrics like minADE, minFDE (minimum Average and Final Displacement Error), and MR (Miss Rate), which are crucial for evaluating the accuracy and reliability of predicted trajectories. Additional metrics like Precision, Recall, and F1 Score offer further insights into the model's performance in trajectory-based tasks. Lastly, for signal con-

Table 3: Metrics used in different models

Domain	Source	Metric
Traffic	UrbanGPT (Li et al., 2024b)	MAE, RMSE, Recall, Macro-F1
	LLM-MPE (Liang et al., 2024a)	RMSE, MAE, MAPE, R^2
	ST-LLM (Liu et al., 2024a),TrafficBERT (Jin et al., 2021),TPLLM (Ren et al., 2024)	RMSE, MAE, MASE, MAPE
Mobility	GenFollower (Chen et al., 2024b)	MSE, Collision Rate, TTC
	PromptMining (Xue et al., 2024)	RMSE, MAE
	LLM-Mob (Wang et al., 2023b)	Acc@1, Acc@5, Acc@10, Weighted F1, nDCG@10
	MobiGeaR (Shao et al., 2024)	Radius, DailyLoc, IntentDist, G-rank, LocFreq, ODSim
	MobilityGPT (Haydari et al., 2024)	JSD
Trajectory	Traj-LLM (Lan et al., 2024)	minADE, minFDE, MR
	LC-LLM (Peng et al., 2024)	Precision, Recall, F1 Score, RMSE
Signal Control	LLMLight (Lai et al., 2023)	ATT, AQL, AWT

trol systems, LLMLight utilizes metrics such as ATT (Average Travel Time), AQL (Average Queue Length), and AWT (Average Waiting Time) to assess the effectiveness of traffic signal optimization models. These metrics are pivotal for ensuring efficient vehicle flow through urban intersections, contributing to overall traffic management Together, these metrics provide a comprehensive framework for evaluating model performance across a diverse range of urban and mobility prediction tasks, ensuring that models are assessed from both accuracy and functionality perspectives.