Traffic Flow Prediction Based on Large Language Models and Future Development Directions

Muhua Zhang[1*] and *Wenzheng* Zhao[2]

[1]School of Computing, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom
[2]College of Computer Science, Chongqing University, Chongqing 400044, China

Abstract. As the application of deep learning in intelligent transportation systems becomes increasingly widespread, the accuracy and reliability of traffic flow prediction have become crucial. However, existing deep learning methods are often complex in model design and lack intuitiveness, making it challenging to provide responsible explanations for traffic predictions. This paper references a responsible and reliable traffic flow prediction model (R2T-LLM) based on large language models (LLMs). This model captures complex spatiotemporal patterns and external factors by converting multimodal traffic data into natural language descriptions. By leveraging LLMs' advanced understanding capabilities, R2T-LLM provides more transparent and interpretable predictions. It bridges the gap between technical performance and real-world application needs. While maintaining accuracy comparable to deep learning baselines, R2T-LLM offers intuitive and reliable prediction explanations. This paper also explores the spatiotemporal and input dependencies of conditional future traffic predictions and compares the model with other different approaches and types. Furthermore, it evaluates the potential of R2T-LLM in addressing challenges in large-scale urban traffic systems, highlighting its developmental and application prospects.

1 Introduction

Traffic network forecasting is a key component of traffic management systems, aiming to predict future traffic conditions such as congestion, traffic volume, and travel time. Accurate forecasting is crucial for decision-makers to make informed choices. However, achieving reliability and precision in predictions is challenging due to the inherent nonlinear dynamics of traffic, spatiotemporal complexity, and the impact of dynamic factors such as accidents and weather. Compared to mature deep learning models, large language models (LLMs) for spatiotemporal learning have advantages in adapting domain knowledge to urban multimodal data and generating reasonable explanations.

* Corresponding author: m.zhang@ldy.edu.rs

At the same time, LLMs exhibit some notable characteristics in traffic flow forecasting compared to traditional neural network models. By processing and analyzing large-scale historical traffic data, they have the potential to learn complex traffic patterns, which may help improve the accuracy of predictions. Although the interpretability of LLMs remains an active research area, they can generate a certain degree of explanation for input data, which may help in understanding the results of predictions. In addition, the generalization ability of LLMs is noteworthy, especially in zero-shot prediction and adaptability. In most scenarios, even without specific training for particular city data, LLMs still show some predictive ability. This generalization ability can be further enhanced through fine-tuning techniques, which can quickly adapt to new prediction tasks, reducing the need for retraining and potentially lowering the cost and time of model deployment.

In this direction, this paper aims to explore the application advantages of LLMs in traffic flow forecasting and compare them with traditional neural network models. By analyzing their performance in processing traffic flow data, predictive accuracy, model interpretability, and generalization ability, it attempts to assess whether LLMs have significant advantages in practical applications, and finally, through comparative analysis, reveals the potential of LLMs in the field of traffic flow forecasting to provide references for future research and applications.

2 Research status analysis

2.1 Techniques for Traffic Flow Forecasting

Existing traffic flow forecasting methods are primarily based on deep learning, and these methods have made significant progress by learning the underlying patterns of traffic data. However, these methods often require specific network structure designs to cope with the multimodal dynamic nature and spatiotemporal complexity of traffic data. Although these designs can help the model improve prediction accuracy, as shown in Fig. 1, they are usually artificially constructed, and their abstract representation further obscures the generalization ability.

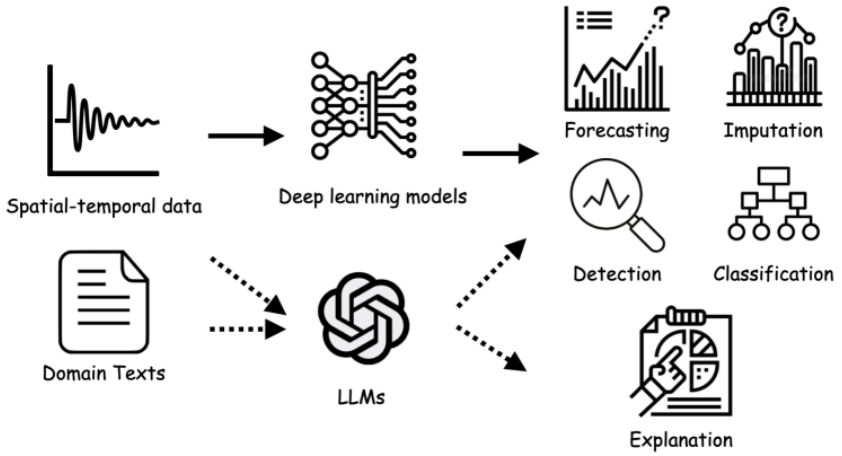

Fig. 1. Simple comparison between LLMs for spatiotemporal learning and deep learning models [1].

2.2 Application of LLMs

To address the aforementioned issues, a paper by Guo, Zhang, and others proposed a

Responsible and Reliable Traffic flow forecasting model with LLMs (R2T-LLM) [1]. This is a model that uses LLMs to generate reliable traffic predictions. R2T-LLM converts multimodal traffic data (such as traffic volume, speed, accident information, etc.) into natural language descriptions, fine-tunes based on language instructions, and aligns with spatiotemporal traffic flow data.

This model can understand and process complex traffic data and generate intuitive prediction results. In empirical studies, R2T-LLM has demonstrated accuracy comparable to deep learning baseline models, which is particularly important for traffic management decision-makers as it helps to better understand the logic behind the prediction results. Of course, even though R2T-LLM shows potential in traffic prediction, it still needs to be optimized in handling large-scale datasets and real-time forecasting, and future research can explore how to optimize LLMs to improve their forecasting efficiency and real-time performance.

2.3 Hybrid Transformer and spatiotemporal self-supervised learning

Long-term traffic prediction faces challenges of dynamic temporal dependencies and complex spatial dependencies, and the feasibility of models that combine hybrid Transformers and spatiotemporal self-supervised learning. Mainly composed of stacked spatiotemporal blocks and spatiotemporal heterogeneity measurement blocks, the spatiotemporal blocks use the self-attention mechanism of Transformers and graph convolution to capture long-term spatiotemporal dependencies in traffic data [2]. The spatiotemporal heterogeneity measurement block simulates the heterogeneity of time and space through self-supervised learning tasks, used for long-term traffic prediction.

Existing models are mostly limited to short-term predictions and lack the ability for medium and long-term forecasting capabilities. Deep learning models have made significant progress in traffic prediction, able to model complex spatiotemporal relationships to extract hidden features for accurate predictions. For example, graph convolutional networks (GCNs) and gated recurrent units (GRUs) are used to capture the topological structure of road networks and the dynamic changes in traffic data simultaneously. Self-supervised learning improves the quality of representation by utilizing the potential information in the data, without relying on a large amount of manually annotated label data [3]. This method has been successful in fields such as computer vision and natural language processing.

2.4 Meta-learning applications in traffic prediction

Meta-learning or "learning to learn" methods can quickly adapt to new traffic patterns or environmental changes. By training models to learn how to quickly adapt from a small amount of data, the model's generalization ability in different traffic scenarios can be improved [4].

This type of model can quickly adapt to new traffic patterns by learning efficiently from a small amount of data. It is suitable for cities with frequent changes in traffic conditions, such as sudden increases in holiday traffic volume or congestion caused by emergencies. In current research, meta-learning models generally demonstrate rapid adaptation capabilities in different scenarios. Compared with traditional deep-learning models, meta-learning models show better generalization when dealing with new scenarios. However, there is still room for optimization in handling large-scale datasets and real-time forecasting.

2.5 Application of multimodal fusion in traffic forecasting

Multimodal fusion models combine various data sources, such as traffic volume, weather

conditions, social media data, etc., to provide a more comprehensive traffic forecast. These methods can use complementary information between different data sources to improve the accuracy and robustness of predictions.

In actual applications based on this approach, multimodal fusion models usually include an input layer that integrates multiple data sources and a deep network structure for extracting and fusing features [5]. This method can handle different types and formats of data, extracting useful features for traffic forecasting. In empirical research, multimodal fusion models have demonstrated superior performance in different traffic scenarios, but the issues of handling large-scale multi-source data and real-time forecasting cannot be ignored.

3 LLMs in different domains

In the research and application of intelligent transportation systems, LLMs are gradually showing their unique potential and value. The following is a general analysis of the integration points of LLMs in the field of traffic flow forecasting with other methods and fields.

3.1 Traffic flow forecasting and traffic accident severity prediction

Traffic flow forecasting and the prediction of traffic accident severity are key tasks in the field of traffic management. They directly affect urban traffic safety, efficiency, and the quality of life for residents. They usually rely on similar data sources, such as traffic volume, speed, and accident reports. Using deep learning methods, such as Convolutional Neural Networks (CNNs), can learn traffic patterns and temporal dependencies, thus predicting the severity of traffic accidents [6]. However, deep learning models may face overfitting and insufficient generalization when dealing with large-scale, high-dimensional data. LLMs can provide more comprehensive forecasts by integrating contextual information.

By combining LLMs with these deep learning methods, insights into accident risks can be improved, learning accident patterns from historical data, and combining real-time traffic data to predict potential accident risks, providing more reliable support for traffic management and decision-making.

3.2 Internet of Things (IoT) and traffic flow forecasting

IoT technology provides a wealth of real-time data for traffic flow monitoring. Model-based techniques can efficiently select sensors and optimize the data collection process [7]. This data includes vehicle location, speed, traffic light status, etc. The real-time and diversity of the data make traffic flow forecasting more complex. Traditional traffic flow forecasting models often rely on fixed data sources and limited data processing capabilities, which make it difficult to cope with high-dimensional and highly dynamic data.

LLMs have advantages in processing high-dimensional data. They can not only handle structured data, such as vehicle location and speed, but also unstructured data, such as sensor logs and social media text. By integrating real-time data from different sensors, LLMs can provide more refined and real-time traffic flow forecasts. At the same time, the prediction accuracy can be improved by optimizing the data collection process, reducing the consumption of resources for data collection and processing.

3.3 Points of Interest (POI) data and traffic flow forecasting

POI data provides spatial distribution information of traffic generation and attraction points,

which is crucial for understanding traffic flow patterns. POI data usually includes the location and attribute information of places such as shopping malls and schools. By combining remote sensing data, the spatial and semantic enhancement of POI data can improve the accuracy of traffic flow forecasting [8].

LLMs can integrate real-time data about POIs and related social media comments, capturing the relationship between traffic flow and urban activities. For example, by analyzing comment text to identify traffic volume and patterns at specific locations to predict congestion and accidents. In addition, machine learning algorithms such as clustering and classification can further analyze the association between traffic flow patterns and specific types of POIs, providing more targeted optimization solutions.

3.4 Application of graph data and algorithms in traffic flow forecasting

Graph Neural Networks (GNNs) are a type of deep learning model specifically designed to handle graph-structured data. They predict traffic flow by considering the topological structure of the traffic network. GNNs can capture complex relationships and patterns in the traffic network to improve the accuracy of predictions. The introduction of attention mechanisms further enhances the model's ability to capture traffic patterns, making GNNs more flexible in handling dynamic changes in traffic networks [9].

LLMs can be combined with graph data algorithms, using their powerful language processing capabilities for more in-depth graph structure analysis and forecasting, that is, not only understanding the topological structure of the traffic network but also understanding the semantic information of traffic events.

Furthermore, LLMs can predict abnormal patterns in the traffic network through graph data algorithms, providing early warnings for congestion and accidents.

3.5 Application of reinforcement learning in traffic flow forecasting

Reinforcement learning is a learning method that optimizes the decision-making process through interaction with the environment. In traffic flow forecasting, it can be used to adjust prediction models to adapt to real-time data and dynamic changes [10]. Reinforcement learning models continuously try and learn to optimize their prediction strategies, thereby improving the accuracy and reliability of their final predictions.

LLMs can be combined with reinforcement learning to provide more flexible and adaptable prediction models. By learning and adapting to the ever-changing traffic conditions, they can understand and learn the semantic information and patterns of traffic events, providing more accurate predictions and optimization solutions [11].

Specifically, LLMs can analyze the descriptions of traffic events and traffic data to identify trends in changes in traffic flow patterns. At the same time, they can adjust their prediction strategies according to these trends to adapt to the ever-changing traffic conditions.

3.6 Application of event data in traffic flow forecasting

Event data plays a crucial role in traffic flow forecasting. These data usually include accidents, construction, special weather, large-scale events, and other sudden events that affect traffic volume. These events often lead to sudden changes in traffic patterns, affecting the accuracy of predictions. Traditional forecasting models may struggle to quickly respond to the changes brought about by these sudden events. Therefore, integrating event data can significantly improve the performance of forecasting models, adjusting node weights dynamically to improve the accuracy and timeliness of predictions [12].

LLMs have unique advantages in processing and integrating event data, capable of

extracting relevant information from various data sources, analyzing complex event relationships, and integrating this information into traffic flow forecasting. They show great potential in capturing subtle changes in traffic flow and responding to sudden events. With their powerful natural language processing and integration capabilities, they provide new ideas and methods, bringing more possibilities to traffic flow forecasting.

3.7 Application of transfer learning in traffic flow forecasting

Transfer learning allows models to transfer knowledge learned in one domain to another, which helps improve the model's adaptability and prediction accuracy in new environments. In traffic flow forecasting, transfer learning can help models quickly adapt to new traffic environments. For example, a traffic flow forecasting model trained in one city can quickly adapt to the traffic flow patterns of another city through transfer learning, reducing training time and data requirements [13].

The generality of LLMs makes them ideal candidates for transfer learning, capable of transferring knowledge across cities or scenarios. Through transfer learning, LLMs can quickly adapt to new traffic flow data, providing accurate prediction results. In addition, they can further improve the accuracy and reliability of traffic flow forecasting by combining data from different sources.

3.8 Application of few-shot learning in traffic flow forecasting

Few-shot learning addresses the issue of data scarcity, allowing models to learn effectively with limited data. For example, when a new road or traffic facility is put into use, the relevant data may be very limited, but the model still needs to adapt quickly and make accurate predictions. The generalization and powerful learning capabilities of LLMs give them potential application value in few-shot learning scenarios [14].

Compared to traditional neural models, which can also have sufficient generalization, the advantage of LLMs is that they can provide multiple better solutions in the same situation, rather than just a short-term optimal solution, and are more flexible in complex scenarios. At the same time, the combination of transfer learning and few-shot learning can further enhance the predictive ability under conditions of scarce data, thus providing accurate and timely predictions in a dynamic traffic environment. The advantages and disadvantages of LLMs compared to other traditional neural network models are shown in Table 1.

Table 1. Comparison of Advantages and Limitations of LLMs with Other Traditional Neural Network Models

	Advantages	Limitations
LLMs	Can consider complex factors during inference and have zero-shot generalization ability and fine-tuning capabilities. By fine-tuning for specific tasks, they can further optimize their performance to adapt to different application scenarios.	Strong dependence on data, high demand for computing resources, and higher complexity of algorithms and models. If the training data is biased or insufficient, the model may exhibit bias or lack of generalization, increasing the difficulty of model development and maintenance.
RNN (Recurrent Neural Networks)	Good at processing sequence data, such as time series or text, capable of capturing temporal dependencies in sequences, very useful for tasks that require remembering historical data.	Prone to the problem of vanishing or exploding gradients, which can affect the training effect of the model. Compared to other types of neural networks, RNNs usually require longer training times, especially when dealing with long sequence data.

LSTM (Long Short-Term Memory Networks)	Provide a certain degree of interpretability through its gating mechanism (forget gate, input gate, output gate), making the model's decision-making process easier to understand and analyze. Better at handling long-term dependencies in long sequence data, avoiding the vanishing gradient problem of traditional RNNs.	Long training time, due to its complex structure and parameters, LSTM is more prone to overfitting, especially when the amount of data is small. And parameter tuning is relatively difficult, requiring careful design and adjustment to avoid performance degradation or overfitting.
Transformer	Catches long-distance dependencies in sequences through self-attention mechanisms, good at processing long text data. Parallel design makes it more efficient during training and inference, especially when using modern hardware acceleration.	Sensitive to noise in input data, which may affect its performance and prediction results. The internal self-attention mechanism makes the model's decision-making process hard to explain, which may be a problem for applications that require transparency.
CNN	Reduces the number of model parameters through local connections and weight sharing, reducing the risk of overfitting. Good at extracting spatial features in image or sequence data, very effective for image recognition and natural language processing tasks.	Focuses mainly on local features, with limited ability to capture global information, which may be a problem for tasks that require a global perspective. Its performance largely depends on the quality and preprocessing of the input data, if the input does not meet the model's expectations, it may lead to a decline in performance.

In summary, LLMs have a wider application prospect in the field of traffic flow prediction and can be combined with a variety of data sources and methods to provide more accurate, reliable, and interpretable prediction results. With the continuous progress of technology, it is expected that LLMs will play a more important role in the future development of ITS.

4 Challenges and issues

Despite the impressive performance of R2T-LLM in traffic flow prediction, there are still significant challenges in capturing subtle changes in traffic flow, especially under the influence of dynamic human activities and complex transportation systems. Unlike the financial sector, traffic flow prediction needs to consider a variety of spatiotemporal factors, such as the structure of the road network and the impact of traffic events. In the field of education, the application of LLMs is more focused on language processing and recommending personalized learning paths, involving less complex physical environments. In the legal field, LLM applications are mainly concentrated on document analysis and reasoning of legal provisions, without involving dynamic environments and the integration of multi-sensor data.

Additionally, how to more effectively utilize spatial information and consider the spatial relationships between different sensors, and how to integrate city-level multimodal data into LLMs to handle downstream tasks such as urban planning, traffic management, and pollution control, are all issues that future research needs to address. Compared with other fields, traffic flow prediction not only requires efficient data processing capabilities but also a deep understanding and parsing of real-time traffic dynamics and complex spatiotemporal dependencies. These characteristics make the field of traffic flow face unique challenges and opportunities when applying LLMs.

5 Future research directions

Future research will delve into how to make LLMs more effective in utilizing spatial information to further enhance the accuracy of traffic flow prediction. In terms of specific implementation directions, research will focus on how to better integrate spatial data from different sensors. By optimizing the spatial relationships between these sensors, it is possible to more accurately capture the complex dynamics of real-time traffic flow.

At the same time, more external factors need to be included in the model's consideration. For example, selecting the frequency and location of traffic accidents, the distribution of activities in time and place, and the impact of major events are all crucial factors for traffic flow prediction [15]. By introducing these external variables, the prediction accuracy can be significantly improved to provide more reliable traffic flow estimates.

Another promising research direction is the development of LLM systems for "urban brain" concepts. This concept aims to integrate city-level multidimensional data (including traffic, environment, public safety, energy use, etc.) to handle and optimize various urban management tasks. Specifically, these systems will be able to process the city's real-time data streams, providing decision support for intelligent traffic management, environmental monitoring, emergency response, and more.

Moreover, interdisciplinary collaboration is also an important direction for future research. By working with experts in urban planning, public policy, computer science, and other fields, problems in traffic flow prediction can be examined and solved from different perspectives, leading to the development of more comprehensive and efficient solutions. Through multidimensional research, future LLMs will play a more important role in the field of smart transportation, aiding in urban management and traffic optimization.

6 Conclusion

R2T-LLM, as an innovative traffic forecasting model, has demonstrated its outstanding performance and broad application potential in many aspects. It is on par with, and even superior to, existing advanced classical model technologies in terms of predictive accuracy and performance. By integrating multimodal inputs, it can process information from different data sources, thus providing more comprehensive and precise traffic forecasts. This paper, through comparative analysis, further explores its unique advantages in accuracy, interpretability, and generalization capabilities.

More importantly, R2T-LLM is not just a powerful forecasting tool; it also provides in-depth insights into the results of its predictions. By converting multimodal traffic data into natural language descriptions and fine-tuning with language instructions, it demonstrates comparable accuracy to deep learning baseline models. This model design not only enhances the reliability of predictions but also strengthens adaptability in complex traffic environments. The language-based representation allows the model to generate explanatory and understandable prediction results, which are significant for informed decision-making in urban traffic planning and management.

R2T-LLM's predictive results are intuitively reliable and interpretable. Its language-based representation enables the model to provide explanations of input dependencies, helping decision-makers better understand the logic behind the predicted outcomes.

It is worth noting that R2T-LLM's multimodal capabilities and language processing abilities also perform exceptionally well in dealing with complex urban traffic environments. It can not only process and analyze a large amount of traffic data in real-time but also generate explanations and suggestions through natural language. With rapid response and strategy adjustment, highly integrated and intelligent features, it has important application value in the construction of modern smart cities.

Compared to traditional deep learning models, R2T-LLM's natural language processing capabilities enable it to generate easily understandable predictive explanations; through fine-tuning techniques, it can quickly adapt to new traffic prediction tasks without the need for retraining. Traditional models often rely on complex network structures and abstract representations, making it difficult to provide intuitive explanations. Its predictive capabilities on datasets it has not processed before demonstrate strong generalization and flexible problem-solving abilities, while traditional models may lack stability and a variety of choices in such new environments.

Overall, R2T-LLM provides an effective and responsible method for urban traffic planning and management. Its innovation and practicality are reflected not only in technical performance but also in its ability to provide decision-makers with profound observations and guidance. This new type of traffic forecasting model will bring more possibilities and higher efficiency to future urban traffic management.

Authors Contribution

All the authors contributed equally and their names are listed in alphabetical order.

References

1. X. Guo, Q. Zhang, M. Peng, et al. Explainable traffic flow prediction with large language models. arXiv preprint arXiv:2404.02937 (2024).
2. C. Zhou, P. Lin. A traffic volume forecasting method based on multi-channel transformer. Appl. Res. Comput./Jisuanji Yingyong Yanjiu **40**, 2 (2023).
3. H. Yuan, Z. Chen. A short-term traffic flow forecasting algorithm based on time convolutional neural network. J. South China Univ. Technol. (Nat. Sci. Ed.) **48**, 11 (2020).
4. S. H. Luo, Y. Yang. A travel time prediction method based on deep learning and meta-learning. J. Nanjing Univ. (Nat. Sci.) **58**, 4: 561-569 (2022).
5. W. C. Peng, S. N. Guo, H. Y. Wan, et al. Spatiotemporal multimodal point process for traffic accident prediction. Appl. Res. Comput./Jisuanji Yingyong Yanjiu **40**, 8 (2023).
6. F. Alhaek, W. Liang, T. M. Rajeh, et al. Learning spatial patterns and temporal dependencies for traffic accident severity prediction: A deep learning approach. Knowl.-Based Syst. **286**, 111406 (2024).
7. M. Fabris, R. Ceccato, A. Zanella. Efficient sensors selection for traffic flow monitoring: An overview of model-based techniques leveraging network observability. arXiv preprint arXiv:2404.08588 (2024).
8. N. Jiang, H. Yuan, J. Si, et al. Towards effective next POI prediction: Spatial and semantic augmentation with remote sensing data. arXiv preprint arXiv:2404.04271 (2024).
9. L. Sun, M. Liu, G. Liu, et al. FD-TGCN: Fast and dynamic temporal graph convolution network for traffic flow prediction. Inf. Fusion **106**, 102291 (2024).
10. S. Lee, C. Park. Continual traffic forecasting via mixture of experts. arXiv preprint arXiv:2406.03140 (2024).
11. S. C. Rajkumar, V. P. Optimized traffic flow prediction based on cluster formation and reinforcement learning. Int. J. Commun. Syst. **36**, 12: e4178 (2023).
12. Z. Qiu, T. Zhu, Y. Jin, et al. A graph attention fusion network for event-driven traffic speed prediction. Inf. Sci. **622**, 405-423 (2023).

13. Y. Zhang, Q. Cheng, Y. Liu, et al. Full-scale spatio-temporal traffic flow estimation for city-wide networks: A transfer learning based approach. Transportmetrica B: Transp. Dyn. **11**, 1: 869-895 (2023).
14. Z. Liu, G. Zheng, Y. Yu. Cross-city few-shot traffic forecasting via traffic pattern bank. In Proc. 32nd ACM Int. Conf. Inf. Knowl. Manag. 1451-1460 (2023).
15. Q. Shi, W. Zheng. A comparison of short-term traffic flow forecasting methods for road networks. J. Transp. Eng. **4**, 4: 68-71 (2004).