Systematic Review of Large Language Model Applications in Transport Engineering

Sneharsh Belsare[1][0009-0009-3120-4106], Shubham Sharma[1][0000-0001-6750-0753], Simon Denman[2][0000-0002-0983-5480] and Ashish Bhaskar[1] [0000-0001-9679-5706]

[1] School of Civil & Environmental Engineering, Queensland University of Technology, Brisbane, Australia

[2] School of Electrical Engineering & Robotics, Queensland University of Technology, Brisbane, Australia

Abstract

Large Language Models (LLMs) have recently emerged as a powerful tool which are being actively exploited to develop better solutions for transport engineering challenges. Recent studies demonstrated that LLM-based methods match or outperform traditional approaches for tasks like human mobility prediction and visual traffic safety analysis. While recent reviews have begun to explore LLM applications in transportation, a comprehensive systematic review covering public transport operations, traffic signal control, traffic forecasting, traffic safety and autonomous driving is still lacking. This is especially important as the field is rapidly evolving, with increasingly capable models emerging, particularly in vision-based tasks. This review paper systematically examines how LLM-based approaches are applied to public transport, traffic signal control, traffic forecasting, traffic safety, and autonomous driving; to identify strengths, weaknesses, trends, and future research needs. We searched Scopus using LLM and transport-domain keywords (2017–2025) from which 90 peer-reviewed articles met our inclusion criteria and were critically reviewed for their use of LLMs, novelty, contribution, and strengths and weaknesses. Despite the promise of LLMs, they face issues like high computational complexity, hallucinations, poor physical understanding of the real world and lack of standardised benchmarks, which hinder their real-world adoption. We advocate the future research focus on improving LLMs for transport specific tasks, developing better benchmarks and use Agent AI powered by LLM to develop more capable solutions for transport engineering.

Keywords: Large Language Models (LLMs), Transport Engineering, Systematic Review.

1 Introduction

Transport engineering is a discipline that designs, analyses and optimises the mobility of people and goods across road, rail, air and maritime networks. It is responsible for ensuring safe, efficient and sustainable mobility, and directly influences land-use patterns, environment quality and public health. Over the years, the field has sought to address many complex challenges such as reducing traffic congestion, coordinating different transit modes into a single, seamless travel system, and building robust and resilient infrastructure.

To address these issues, transport engineers have increasingly relied on data-driven solutions. The introduction of Intelligent Transport Systems (ITS) has led to the generation of massive and heterogeneous datasets that demand sophisticated analytical tools. Traditional statistical models and rule-based optimisation techniques have yielded important insights, yet they often struggle with the high dimensionality, non-linearity, and contextual nuance inherent in big transport data. Consequently, researchers are now harnessing advancements in artificial intelligence (AI) and machine learning (ML) to better capture complex patterns in mobility and system behaviour.

The recent emergence of Large Language Models (LLMs), a family of advanced deep learning models that excel in understanding and generating human language, offers a new computational paradigm. Beyond conversational agents, LLMs have demonstrated proficiency in code synthesis, data extraction, scenario generation, and multimodal reasoning, all capabilities that align closely with the analytical and operational needs of transport engineering. LLMs can be used to mine large volumes of unstructured data and generate insights, to predict human mobility, optimise traffic signal control, and more. Consequently, a growing body of research is exploring the applications of LLMs to the domain of transport engineering.

Despite this growth, the literature around the use of LLMs in the transport domain is largely fragmented. Contributions span disparate sub-domains and employ a variety of methodologies. Existing reviews on AI in transport are either too broad (general AI/ML capabilities like vision (Pérez et al., 2025)) or too narrow (focusing on one subdomain like traffic safety). LLMs have unique properties (reasoning, multimodality, and tool use) that distinguish them from conventional ML models, requiring a dedicated synthesis. Without a consolidated review, practitioners lack a roadmap for applying LLMs responsibly in transport engineering. Therefore, a systematic review is needed to map the current state-of-the-art, identify the current methodologies and best practices, critically assess approaches for strengths and weaknesses, and suggest directions for future improvement. While there have been attempts at a systematic review like Yan et al., 2025, as of the writing of this paper, there does not exist any peer-reviewed systematic review which addresses the applications of LLMs in the transport domain as a whole. This is especially important as the field is rapidly evolving, with increasingly capable models emerging, particularly in vision-based tasks.

The aim of this systematic review is to critically assess the applications of LLMs in key transport engineering domains: public transport, traffic signal control, traffic forecasting, traffic safety and autonomous driving. These domains were specifically chosen as they offer rich, structured data and clear performance metrics. This review catalogues, evaluates and assesses these studies, assesses their strengths and weaknesses, and identifies existing gaps. By systematically addressing these objectives, the review will provide a reference point for future research. The contributions of this review include:

- A 6-class taxonomy of the role that LLMs play in transport related applications.
- Identification of gaps like lack of benchmarks, ethical challenges, etc.
- Suggestions for future directions of research.

The organisation of the paper is as follows: Section 2 introduces LLMs, discusses how they work and highlights some key techniques. Section 3 discusses the search methodology. Section 4 discusses the application of LLMs in all but the autonomous driving domain, which is discussed in detail in Section 5. Section 6 provides discussion about the strengths, weaknesses and future scope of LLMs in transport engineering. Finally, Section 7 concludes the review.

2 Introduction to LLMs

Language is a rich and complex system of human expression, shaped by the rules of grammar. Language Modelling (LM) refers to methods that enable machines to understand and generate language. Early efforts in LM used statistical methods to predict the next word in a sequence. With the rise of machine learning, these methods evolved into neural network–based models, eventually giving rise to large language models (LLMs). Unlike earlier models, LLMs go beyond simple word prediction, they exhibit a level of reasoning and adaptability that often resembles common sense. In this section, we explore how LLMs emerged, how they work, and some key techniques.

2.1 History and Evolution

Figure 1 outlines the history and evolution of Language Modelling, which has led to the development of LLMs (Zhao et al., 2025). We discuss these methods in the following subsections.

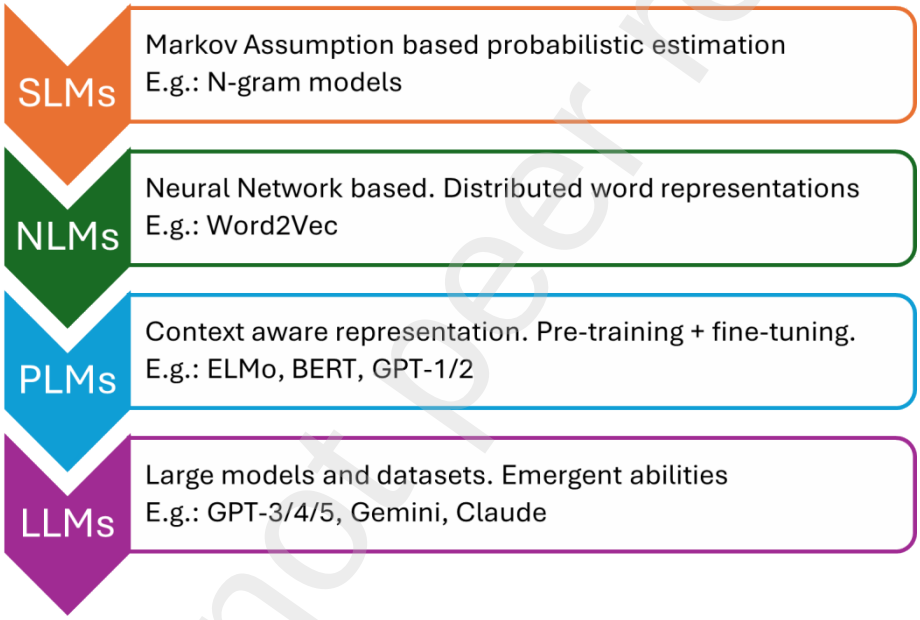

Figure 1 Evolution of Language Models

2.1.1 Statistical Language Models (SLMs)

Early attempts (1990s) at LMs relied on statistical methods, typically using the Markov Assumption, to model word prediction with a fixed length context (n-grams). For example, a bi-gram model would consider only the last two words to predict the next word. These models demonstrated some capability at NLP tasks, however, their capabilities in modelling language were limited given the complex nature of language.

2.1.2 Neural Language Models (NLMs)

In the 2000s, researchers turned to neural networks (NNs) which led to the development of Neural Language Models, which leveraged Multi-Layered Perceptrons (MLPs) and Recurrent Neural Networks (RNNs) to learn distributed word representations and contextual features. Notable works like word2vec (Mikolov et al., 2013) allowed NNs to better capture the semantics of text and automate feature generation, moving beyond mere next word prediction.

2.1.3 Pre-trained Language Models (PLMs)

Next came PLMs, starting with models like ELMo (Peters et al., 2018), BERT (Devlin et al., 2019) and GPT-1 (Openai et al., 2018). These models were trained (or more accurately, pre-trained) on massive corpora in an unsupervised manner, and then further trained (fine-tuned) for downstream tasks. Given language is a sequential construct, parallelising the training proved to be a significant challenge. This problem was mitigated with the introduction of the Transformer architecture in the 2017 seminal paper by Google, “Attention Is All You Need” (Vaswani et al., 2017). This led to an improvement in pre-training efficiency, generalising the learnt representation for a broad set of tasks and establishing the pre-training and fine-tuning paradigm.

2.1.4 Large Language Models (LLMs)

Researchers identified that scaling up PLMs, both in terms of number of parameters and data size, boosted their task solving abilities. As the models became bigger, they started to demonstrate surprising abilities (termed emergent abilities) such as in-context learning and few-shot problem solving, which were largely absent in smaller models. Owing to these emergent abilities, which the smaller PLMs didn’t have, these models were termed Large Language Models, and they are large with respect to in both training corpus and model size). Examples of recent LLMs include GPT-3 (175 billion parameters) (Brown et al., 2020), PaLM (Chowdhery et al., 2022) and LLaMA (Touvron et al., 2023).

2.2 The Inner Workings of LLMs

2.2.1 Architecture

At their core, virtually all LLMs use the Transformer architecture and follow a similar structure. In this section will we examine this common LLM pipeline. Figure 2 outlines the key components of this architecture: tokenisation, embedding mapping (with positional encoding), followed by stacked Transformer blocks (attention followed by a feed-forward layer), ending

with linear and SoftMax layers for output. It visualises the sequential flow and key components of the model architecture.

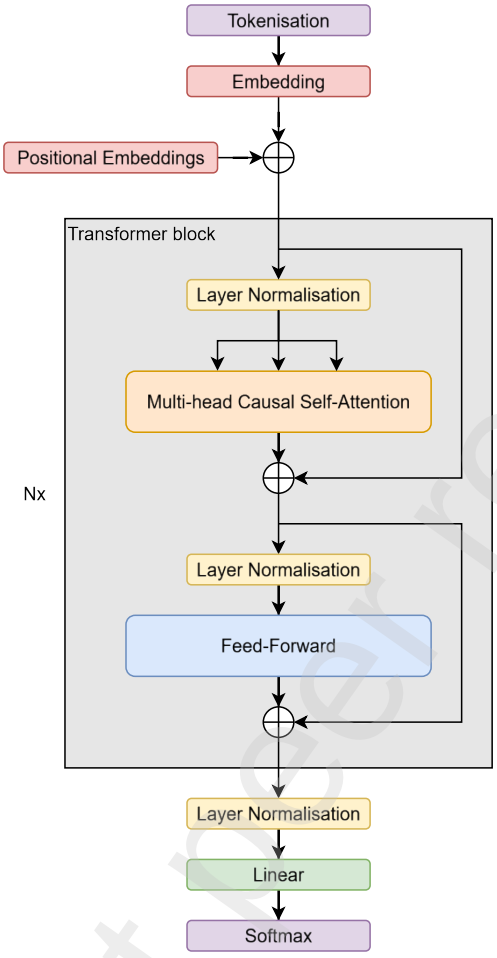

Figure 2 LLM Architecture

Tokenisation and Embedding

The first step in processing text is tokenization, where the raw text (sentences or paragraphs) is broken down into smaller discrete units called tokens. A token could be a word, part of a word, or even a single character. This breakdown is based on a pre-defined vocabulary, and each token has a corresponding integer ID. Since neural networks cannot directly work with text, this conversion into numeric form is essential.

The next step is transforming these tokens to embeddings. Here, each token ID is mapped to a high-dimensional, fixed-size vector, known as an embedding. These vectors capture the semantic meaning of tokens. The process is straightforward: the model maintains an embedding matrix with one row per token in the vocabulary, and each row corresponds to the embedding vector for that token. When given a token ID, the model simply looks up the matching row. Over training, this matrix comes to represent meaningful properties of tokens, allowing the model to reason about the similarity between tokens, and their relationships in continuous space.

Transformers lack an inherent mechanism to handle the order of tokens in a sequence. To solve this problem, a positional embedding is used. Additional information about the relative or absolute position of the tokens in the sequences is injected into the embeddings. In Vaswani

et al., 2017, the authors proposed two variants, i.e. sinusoidal and learned position embeddings.

The Transformer block

The Transformer architecture is what allows for the impressive performance of the LLM. Its function can be further broken down into two components:

Attention: This mechanism lets every token "consult" every other token in the same input sequence. Transformers enable each token in the input to weigh the relevance of every other token in the sequence, capturing long range dependencies and contextual relationships more effectively. This process is defined mathematically below.

For each embedding e_i,$e_i \in E$, where E is the embedding space, three vectors are computed: query (q), key (k) and value (v), using learned weight matrices W_Q, W_K and W_V as follows:

$$q_i = e_i \times W_Q$$

$$k_i = e_i \times W_K$$

$$v_i = e_i \times W_V$$

The attention score between the i[th] and the j[th] embedding (token) is calculated using the scaled dot product of their queries and keys.

$$\text{Attention score}_{ij} = \frac{q_i \cdot k_j}{\sqrt{d_k}}$$

Here, d_k is the dimensionality of the key vectors. The division by $\sqrt{d_k}$ scales the dot product to prevent large values which could result in small gradients and subsequently lead to the vanishing gradient problem (Hochreiter, 2011).

Next, attention weights are obtained by applying the SoftMax function:

$$\alpha_{ij} = \text{SoftMax}(\text{Attention score}_{ij})$$

Finally, the output for each embedding is obtained by calculating the weighted sum of the values of vector v_v of all tokens using the attention weights as the weights.

$$z_i = \sum_{j=1}^{n} \alpha_{ij} \cdot v_j$$

This entire process can be represented in a compact form as shown below

$$\text{Attention}(Q,K,V) = SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Here, $Q = E \times W_Q$, $K = E \times W_K$ and $V = V \times W_V$. This is the form seen in the original Transformer paper (Vaswani et al., 2017).

LLMs often use Multi-head Attention which allows the model to evaluate different parts of the input token sequence in parallel, capturing multiple aspects of the relationship between tokens. In multi-head attention, h different attention heads are used, each with its own set of learned projections: W_i^Q, K_i^Q and V_i^Q. The output from all heads is concatenated and projected to form the final output.

Feed-Forward Network: The Feed-Forward Network (commonly abbreviated to FFN) is a small neural network that is applied independently to each embedding after the attention layer to introduce non-linearity and to transform the representations (embeddings) into higher-level abstractions, allowing the model to develop a deep semantic understanding. This consists of two linear transformations with a non-linear activation, like a ReLU, in between.

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2$$

To facilitate training and improve gradient flow, the transformer architecture employs residual connections with layer normalisation. These techniques help prevent issues like vanishing gradients.

$$Residual(x) = x + Sublayer(x)$$

$$Output = LayerNorm(Residual(x))$$

An LLM consist of several dozen of such transformer blocks applied sequentially. This allows the model to learn more abstract and high-level representations of the input, and to discover deeper connections between the tokens and the concepts.

2.2.2 Training and Inference

Training Objective: Given a sequence of tokens $t_1, t_2, ..., t_n$, an LLM is trained to estimate the probability distribution over possible candidates for t_{n+1}. Instead of producing a single token, the LLM outputs a distribution, and training aims to assign higher probability to the correct continuation.

Loss Function: The discrepancy between the predicted probability distribution and the true next token is measured using cross-entropy loss, which is minimized during training.

Optimizer: Gradient descent algorithms like Adam (Kingma and Ba, 2014) and its variants are used.

Dataset: LLMs are trained on vast amounts of data, which is sourced from sources like webpages, books & articles, news, scientific articles, code from GitHub. LLMs which intend to specialize in certain tasks (like translation) might include data from other sources (or in this case, other languages). For example, GPT-3 (Brown et al., 2020) uses an augmented version of the CommonCrawl database.

Inference: At inference time, the model again produces a probability distribution over the vocabulary for each step. A token is then selected either deterministically (choosing the most probable token) or stochastically (sampling according to probabilities). Sampling can be controlled with parameters such as temperature, which adjusts the balance between determinism and diversity in the generated text.

2.3 Key LLM Techniques

2.3.1 Parameter Efficient Fine-Tuning (PEFT)

LLMs are pre-trained on a large and diverse corpora, which makes them effective for a wide range of tasks. However, sometimes the abilities of LLMs fall short of the requirements for downstream tasks. In such cases, we can fine-tune the LLM with domain specific knowledge. However, given the large size of the model, fine-tuning can be a resource intensive task. This has led to the development of parameter efficient fine-tuning techniques such as LoRA.

Low Rank Adaptation (Hu et al., 2021), or LoRA for short, works by freezing most of the LLM's weights, and injecting small, trainable low-rank matrices (called adapters) within specific layers (often the attention block). During fine-tuning, only these low rank adapters are updated, allowing for effective adaptation to a new task while requiring significantly less computational resources than if the entire model was fine-tuned. The adapters rely on matrix decomposition, where a high-dimensional weight matrix is factorised into two smaller low-rank matrices. For a weight matrix $W_{d\times d}$, it can be decomposed as $W_{d\times d} = A_{d\times r}B_{r\times d}, r \ll d$.

Quantised LoRA (Dettmers et al., 2023)or QLoRA for short, is another PEFT techniques which combines LoRA with aggressive weight quantisation. It is very similar to LoRA, with the key difference that the base model's weights are converted from a high precision (16 or 32 bit) to 4-bit precision using methods like NormalFloat4 (NF4) quantisation. This drastically reduces the memory footprint while maintaining strong predictive performance and accuracy.

2.3.2 Retrieval Augmented Generation (RAG)

LLMs answer questions based solely on their training data. This presents certain challenges. First, the data on which the LLM was trained on can become stale and outdated. Second, there's no way to trace the original source of the data in a response generated by an LLM. To mitigate these issues, and to allow for domain-specific responses, we can use Retrieval Augmented Generation (RAG) approaches.

Retrieval Augmented Generation (Lewis et al., 2020), commonly abbreviated to RAG, augments an LLM with a retrieval system. Instead of relying solely on static, pre-trained data, the model first searches a database of sources for information relevant to the user's query. It then incorporates this retrieved data into its response, yielding informed and contextually grounded outputs. This not only improves the accuracy and relevance of the LLM's output but also helps mitigate hallucinations and avoid outdated outputs by grounding answers in current and reliable sources. RAG also allows for domain adaptation, by providing the LLM with specialised and highly domain-relevant documents, without the need for fine-tuning.

2.3.3 Prompt Engineering

Prompt Engineering, also referred to as prompting, is the process of meticulously crafting the natural language input to an LLM in order to get specific, high-quality and relevant output.

2.3.3.1 Zero-shot / Few-shot prompting

Zero-shot prompting is an approach where we direct instructions to the LLM to perform certain tasks, without any example / demonstration in the prompt. The LLM provides a generalised solution solely based on the patterns it has seen in the training data. However, to improve the relevance of the response, we can use **Few-shot prompting**, where we supplement the instruction with additional details or examples / demonstrations, helping the model infer task structure, desired format and edge cases. Few-shot prompting is one of the easiest ways to get desirable responses from LLM without fine-tuning or using RAG.

2.3.3.2 Chain-of-Thought Prompting

Chain-of-Thought (CoT) prompting (Wei et al., 2022) seeks to encourage LLMs to work in a similar manner to human brains, which uses two systems of thinking: System 1, which is fast, intuitive, and gives quick answers; and System 2, which is slower, more careful, and reasons step by step. Normally, an AI (like a person using System 1 thinking alone) might jump straight to an answer, which can be fast but error-prone. CoT prompting encourages the AI to switch to a System 2–style process, where it explains its reasoning step by step before giving the final answer. This makes the reasoning clearer and often leads to more accurate results. Its main advantage is better accuracy and transparency, while its drawback is that it can be

slower and may sometimes produce unnecessarily long reasoning. Figure 3 shows an example of CoT Prompting in actions, highlighting how it improves the accuracy of an LLM.

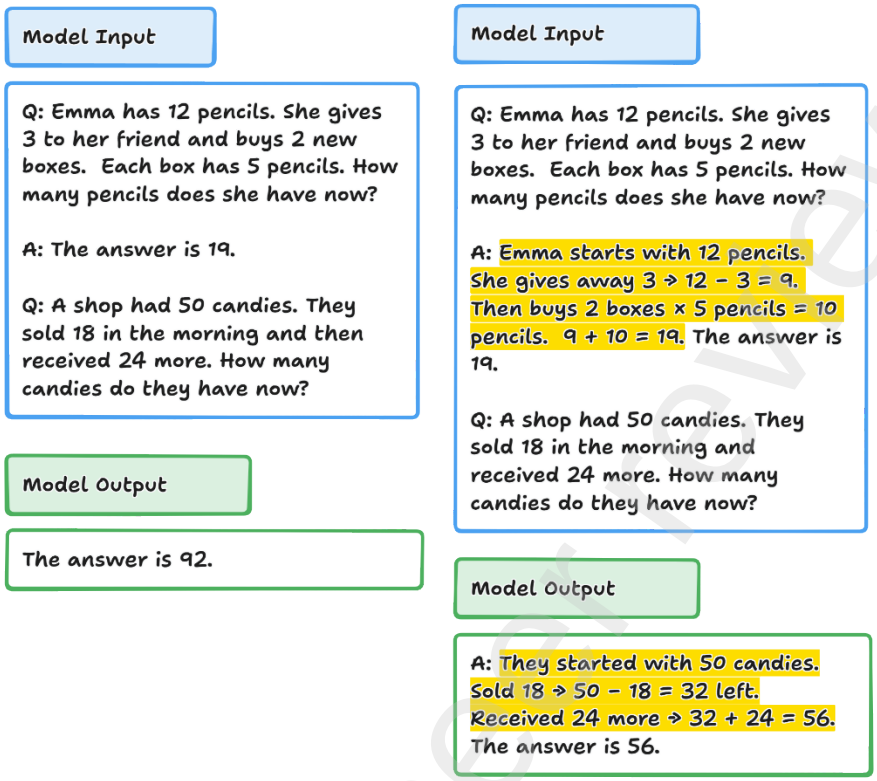

Figure 3 CoT example

2.3.4 Knowledge Distillation

Knowledge Distillation is the process of transferring knowledge from a large and complex model (called the “teacher”) to a smaller and simpler model (called the “student”). In context of LLMs, this is usually done to train a smaller, less powerful student LLM from a large and powerful teacher LLM for a specific downstream task. This allows the student LLM to achieve performance similar to the teacher LLM, while being less computationally expensive, mitigating one of the common issues of LLMs pertaining to high computational costs.

2.3.5 Multi-modal Large Language Models (MLLMs)

Originally, AI models were siloed: LLMs specialised in NLP (text modality) and Convolutional Neural Networks (CNNs) focused on visual perception tasks such as classification and object detection. Earlier attempts to fuse these modalities for context-rich AI resulted in models like OpenAI’s Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021), which allowed models to link images with text. This marked the beginning of Vision Language Models (VLMs). Recent advances have integrated such multimodal capabilities into LLMs themselves by allowing these models to ingest, interpret and reason across both text and images in one unified architecture, resulting in what we call Multimodal Large Language Models (MLLMs). Popular examples include OpenAI GPT-4o and Google Gemini-2.5-Pro.

2.3.6 Tool Use

LLMs excel at generating detailed plans and reasoning through complex scenarios, but they are fundamentally restricted to producing text only and thus cannot execute actions or interact

with the external world on their own. Tool use bridges this gap by enabling LLMs to interact with external functions, APIs, or plugins, enabling them to retrieve up-to-date information, perform calculations and carry out real-world tasks. Tools also help overcome specific weaknesses of LLMs. For example, since MLLMs struggle with counting objects in video frames, (Yao et al., 2025) integrates a YOLO tool that is far more accurate for this purpose.

3 Search Methodology

The scientific database used for searching relevant literature is Scopus. The search was limited to papers published from 2017 to the present (2025). The reason behind choosing 2017 was that this is when the Transformer architecture was introduced. The search was further restricted to peer-reviewed articles (journal and conference) published in English.

Given that transport engineering is a vast field, the search was performed to find the use of LLMs in five domains of transport engineering, namely: public transport, traffic signal control, traffic forecasting, traffic safety and autonomous driving. Out of these five domains, autonomous driving is the most mature and has a substantial number of publications ((Cui et al., 2024c; Li et al., 2024; Zhou et al., 2024; Ashqar et al., 2025; J. Li et al., 2025; Sathyam and Li, 2025; Zhu et al., 2025)). Therefore, instead of conducting a Scopus search for this domain, we rely on existing peer-reviewed review papers. In addition, Autonomous Driving will be discussed in a dedicated section to evaluate the applications of LLMs within this domain (section 5). The rationale for choosing these domains is that they offer rich, structured data and clear performance metrics where LLMs have already shown promising results. For this review, we restricted ourselves to the five transport domains mentioned earlier, as other domains (like transport planning and freight management) are largely nascent for LLM applications. Nonetheless, future research could include them as well.

The Scopus search query consisted of LLM specific keywords: “Large Language Model*” OR “LLM*” OR “ChatGPT” OR “GPT” OR “Gemini” OR “LLaMA”; and domain specific keywords, which can be found in **Table 1**. Papers which were not related to the domain were removed from the list. A table discussing all papers can be found in the Appendix A.

Table 1 Domain Specific Keywords

Domain	**Keywords**
Public Transport	“Public Transport” OR “Public Transit” OR “GTFS”
Traffic Signal Control	“Traffic Signal Control” OR “Traffic Signal Management”
Traffic Forecasting	“Traffic forecasting” OR “mobility forecasting” OR “demand forecasting” OR “mobility prediction” OR “human mobility” OR “traffic prediction” OR “demand prediction” OR “mobility prediction” OR “human mobility prediction”
Traffic Safety	“Traffic Safety” OR “Road Safety” OR “Transport Safety” OR “Pedestrian Safety”

3.1 Inclusion & Exclusion Criteria

The papers were filtered manually based on the following inclusion and exclusion criteria.

Inclusion

1. The papers must focus on the application of LLMs in the relevant domain.
2. Papers must be published after 2017, as the Transformer architecture itself was published in 2017.

3. Papers should be peer reviewed.
4. Studies must provide sufficient methodology detail, results, and domain relevance to transport engineering.
5. Papers must demonstrate novelty; those merely posing questions to large language models (e.g., ChatGPT) about the transport domain are not considered contributions, unless they critically evaluate the model's capabilities for a specific task and provide evidence of its usefulness.
6. The papers must provide a framework or discussion about the application of LLMs.

Exclusion

1. Papers pertaining to tourism, supply chain and aviation are excluded.
2. Non-English studies.

Any paper failing the inclusion criteria is excluded. To speed-up screening, ChatGPT (GPT-4 and GPT-5) were also used to extract information from papers (like which LLM was used, what were the baselines, etc.), and all outputs were verified by a human before considering them.

It is worth noting that given the rapid pace of development in this domain, many recent studies have not yet appeared in peer-reviewed journals and are therefore not included in this review. However, a significant portion of this emerging research is available as preprints on platforms such as arXiv, reflecting the fast-moving nature of advancements in LLM applications transport engineering.

4 Taxonomy of LLM Roles in Transport Engineering (excluding Autonomous Driving)

To understand how LLMs are applied in transport engineering, we propose a taxonomy based on the specific role the model plays within different applications. The taxonomy, developed through an iterative review of the literature, groups studies into six distinct roles.

4.1 Classification and Prediction (forecasting)

Classification and prediction (forecasting) are among the most prevalent applications of machine learning models. Owing to their exposure to vast amounts of data during training, LLMs develop the ability to recognize intricate patterns and trends. Leveraging these learned representations, LLMs can effectively perform classification and prediction tasks. Their strong performance in such tasks is largely attributed to their extensive parameterization and sophisticated architectures, which enable a deeper understanding of complex relationships within data.

Considering classification tasks, LLMs are primarily used for text-based classification and MLLMs are primarily used for visual inference-based classification. Recent studies (Zhen et al., 2024; Zhen and Yang, 2025) have tested and demonstrated LLMs' ability to infer crash severity from a textual accident description. Using the CrashStats dataset[1] from Victoria, Australia, the authors tasked the LLM with classifying the severity of a given accident into three categories: minor with no-injuries, serious injury, or fatal. They also employed a number of prompt engineering techniques like zero-shot/few-shot prompting and CoT prompting to

[1] https://discover.data.vic.gov.au/dataset/victoria-road-crash-data

improve the performance of the LLMs. The study's contribution lies primarily in demonstrating the feasibility of this approach rather than proposing any novel methodology.

Another study, DDC-Chat (Liao and Lin, 2025), proposes a novel framework for distracted driver classification (DDC) that leverages a fine-tuned VLM augmented with auxiliary tools, referred to as *skills* in the paper. The task involves classifying driver behaviour into categories such as yawning, sleeping, talking on cell phone, texting, drinking/eating and hands off the steering wheel. To address the complexity of the problems, the authors enhanced the VLM with a repository of skills, including understanding skills (e.g. pose detection and segmentation), external knowledge skills (via retrieval), quality enhancement skills (such as pixel enhancement, exposure correction and image defogging) and composite skills that integrate multiple individual skills. When evaluated on the 100-Driver dataset[2] the proposed framework achieved state-of-the-art performance, achieving accuracy over 93%, precision over 95%, and recall over 93%.

Considering forecasting tasks, the ability to predict future traffic conditions or human mobility patterns is very important for transport engineering. Therefore, a body of research focuses on leveraging LLMs for traffic and human mobility forecasting.

xTP-LLM (Guo et al., 2024) is a LLM powered traffic prediction framework. The approach leverages a carefully crafted prompt which contains domain knowledge, CoT prompting, spatial and temporal attributes, historic data, date and holiday information, and weather information, all represented in textual format. To adapt the model, the authors fine-tune Meta's open-source LLaMA-2 7b on the CATraffic (source not given by the authors) dataset from California, USA with the help of the LoRA technique. xTP-LLM achieved competitive performance compared to the state-of-the-art while also generating explanations for its predictions. Notably, xTP-LLM demonstrated strong generalization and robustness, maintaining good accuracy despite class imbalance in the training data (where favourable weather conditions such as sunny days heavily outweighed adverse ones like snow). Furthermore, its effectiveness on a separate dataset, TaxiBJ[3], confirmed its high generalizability.

Another study ST-LLM+ (C. Liu et al., 2025) addresses the challenges of traffic forecasting by incorporating graph-based attention to better capture the complex spatial and temporal dependencies that cannot be adequately represented through simple tokenisation and node embeddings. To achieve this, the authors introduce a special spatial-temporal embedding function along with a Partially Frozen Graph Attention (PFGA) mechanism fine-tuned using LoRA. PFGA allows some layers or weights of the graph attention network to be kept fixed (not updated) during training, while others are allowed to learn and adapt. This helps stabilize training and leverage prior knowledge, allowing only selected parts of the model to specialize on the new task or dataset. It was proposed to overcome the limitations of using fully frozen pre-trained transformers for tasks like traffic forecasting as they may miss critical short and long-term dependencies needed for accurate prediction, so PFGA introduces graph-based attention in the unfrozen layers to better capture spatio-temporal relationships while retaining the benefits of pre-training in the frozen layers. ST-LLM+ consistently outperformed state-of-the-art traffic prediction models on the CHBike[4] and NYCTaxi[5] datasets.

A study by (Jiang et al., 2025) proposed the Knowledge-informed Dynamic Correlation Modelling (KIDCM) framework to address challenges like dynamic and non-linear spatial

[2] https://100-driver.github.io/
[3] https://www.kaggle.com/datasets/qils1964/datap
[4] https://citibikenyc.com/system-data
[5] https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

correlation, the limited generalisability of deep learning frameworks and high computational costs associated with LLMs. They propose a General Spatial Dynamics Modelling (GSDM) method which leverages LLMs to generate unbiased traffic data, which is subsequently used to model spatial correlations. The framework employs the Deepseek-V2.5 LLM and demonstrates superior performance over baselines on the I-24 Motion dataset[6]. To mitigate the high computational costs of the LLM, the authors developed a surrogate model, a hybrid of T-GCN, GCN and GRU architectures, which is trained via knowledge distillation from the LLM. While KIDCM offers superior generalisation ability when compared to the state-of-the-art, the prediction accuracy remains only comparable to that of the pure deep learning models. The authors also highlight the strong generalisability capabilities and robustness (to noise) of the framework.

Another key prediction task is human mobility prediction. Predicting the locations an individual might visit in the future is very important for efficient planning and operation of the transport infrastructure.

MobTCast (Beneduce et al., 2025) is a next-location prediction framework which uses LLMs as the predictors. The framework constructs a prompt incorporating a user's historic stays (long term mobility patterns), context stays (short term mobility patterns) and temporal information. This prompt is given as input to an LLM which outputs a ranked list of the possible next locations along with explanatory reasoning. The authors first evaluated MobTCast against other deep learning baselines (RNNs, ST-RNNs, etc) for geographical transferability. When trained on data from New York and tested in Tokyo[7], or vice versa, MobTCast achieved the highest accuracy. Next, the authors examined the performance 15 LLMs under zero-shot, one-shot and few-shot settings, finding that additional shots did not significantly improve accuracy. Based on these findings, the authors conclude that LLMs are effective next-location predictors in zero-shot scenarios and exhibit strong geographical generalizability.

Another study, LLM-MDC (Guo et al., 2024), proposes next location prediction as a multi-round, interactive task. In the first round, the framework performs activity prediction as an auxiliary task to narrow down potential locations. In the second round, an activity aware prompt is provided to the LLM, enabling it to make predictions for next locations alongside explanatory reasoning. In the third and final round, the LLM makes necessary corrections by integrating the prediction results of deep learning models (Time-LSTM (Zhu et al., 2017), Flashback (Yang et al., 2020) and MSSRM). While the LLM helped in the reasoning process, the overall framework did not surpass the baseline models, despite the LLM having access to the predictions made by the deep learning models, highlighting that the method could be improved further. The dataset used is CATraffic, a portion of which can be found on the GitHub repo of the project[8]

Another study, Mobility-LLM (Gong et al., 2024), leverages LLMs to analyse check-in sequences to learn visiting intentions and travel preferences. As LLMs cannot directly interpret raw check-in sequences, the authors introduce a visiting intention memory network (VIMN) to capture the visiting intentions at each record and a shared pool of human travel preference prompts (HTPP) to guide the LLM in understanding user travel preferences. The framework fine-tunes TinyLlama-1b (P. Zhang et al., 2024) using the LoRA technique while jointly training the VIMN and developing HTPP. Experiments conducted on 4 datasets, Gowalla[9], WeePlace

[6] https://i24motion.org/data
[7] https://sites.google.com/site/yangdingqi/home/
[8] https://github.com/Guoxs/xTP-LLM
[9] https://snap.stanford.edu/data/loc-Gowalla.html

(source not given by the authors), Brightkite[10] and FourSquare[11], show that the framework outperforms the baselines in next locations prediction and trajectory user linking tasks.

The paper by (Yu et al., 2024) addresses the challenge of cross-city origin-destination (OD) flow prediction in urban transport, where data scarcity often hinders accurate modelling. The authors propose LLM-COD, a framework which takes origin POIs, timestamps and candidate POIs to predict destinations and travel distances. It leverages the LLaMA-2 7B and Gemma models, trained with LoRA on real-world taxi and ride-sharing data from Beijing[12], Xi'an, and Chengdu[13]. The method achieves substantial performance gains, reducing RMSE by up to 46% against the state-of-the-art deep learning GODDAG baseline. A major strength lies in its generalisability and transferability enabled by a novel loss function, making it useful for cities with limited OD data. However, a critical limitation is the lack of reasoning or justification in the predictions, which reduces interpretability for practitioners. This could be addressed by integrating chain-of-thought prompting or explanation-aware training, ensuring the model not only predicts flows but also provides interpretable insights into mobility drivers. Overall, the study demonstrates that fine-tuned multimodal LLMs can meaningfully advance transport planning and resource optimization.

The paper TSI-LLM (Luo et al., 2024) introduces a framework for inferring rich semantic information from human mobility trajectories using GPT-4, addressing the limitations of traditional models that depend on auxiliary datasets or produce only shallow activity insights. By inputting spatio-temporal trajectory data like region IDs, POI vectors, and temporal information into structured prompts, the model outputs occupational categories, activity sequences, and natural language trajectory descriptions. The approach outperforms baselines like logistic regression, Bayesian rule-based models, hidden Markov models, and white-box approaches, offering logical and interpretable inferences. Its key strength lies in providing comprehensive semantic reasoning that enables deeper mobility behaviour analysis, which is valuable for urban planning, transportation optimization, and travel behaviour studies. However, the evaluation relies only on qualitative case studies, limiting rigor and generalizability. Future work should include benchmarking on large-scale datasets and clarifying downstream applications, which would establish stronger evidence for the framework's robustness and practical utility.

The paper (P. Tang et al., 2024) demonstrates that a fine-tuned LLaMA-3-8B can effectively forecast long-term citywide human mobility across multiple metropolitan areas. Using the Human Mobility Challenge 2024 dataset[14], the model (LLaMA-3-8B-Mob) predicts future trajectories in JSON format from historical trajectory data and time slot inputs. Evaluated with Dynamic Time Warping (DTW) and GEO-BLEU, it outperforms the 2023 Human Mobility Challenge winner LP-Bert (Boqing Zhu et al., 2023) and ranked in the top-10 of the challenge, showing that open-source models, when fine-tuned with LoRA, can rival specialized mobility predictors. Its strengths lie in leveraging instruction-tuning with structured Q&A prompts, achieving scalability and adaptability for urban planning, disaster response, and epidemic control. However, the study only compares against LP-Bert, without providing numeric results or a broader baseline set, limiting the rigor of its validation. Future work should expand benchmarking and include detailed quantitative comparisons to better establish performance advantages.

[10] https://snap.stanford.edu/data/loc-Gowalla.html
[11] https://foursquare.com/
[12] https://doi.org/10.1145/3183713.3183743
[13] https://gaia.didichuxing.com/
[14] https://wp.nyu.edu/humobchallenge2024/

The paper (Liang et al., 2024) investigates how GPT-4 can enhance mobility prediction by incorporating event-related textual data. The proposed framework, LLM-MPE, combines formatted event descriptions with historical mobility features, enabling the model to output inflow and outflow predictions with step-by-step reasoning on the NYC Taxi Dataset and Barclays Centre event data[15]. Evaluated using RMSE and MAE, it outperforms traditional models such as Linear Regression, ARIMAX, and XGBoost, particularly during event days when mobility patterns are highly irregular. Its strength lies in leveraging LLM reasoning to handle rare, complex event scenarios, offering practical benefits for transit planning and traffic management. However, the model struggles with previously unseen events, often producing inaccurate "guesses." Future improvements should introduce mechanisms to better generalize across novel events, such as hybrid approaches combining LLMs with statistical or simulation-based methods.

The paper HMP-LLM (Zhong et al., 2024) proposes a framework that leverages GPT-4 Turbo for human mobility prediction, with a particular focus on disruptions such as the COVID-19 pandemic. The model converts time-series mobility data - capturing trend, seasonal, and residual components – along with COVID-19 statistics into structured text prompts, from which the LLM predicts next-day mobility values. Evaluated using RMSE, MAE, and MAPE, HMP-LLM consistently outperforms traditional baselines including XGBoost, ARIMA, MLP, Random Forest, and DeepSTN, with notable improvements in short-term forecasting. Its strength lies in demonstrating effectiveness on real COVID-19 datasets, making it valuable for emergency response, traffic management, and public transport planning. However, the framework's reliance on COVID-specific data does not guarantee transferability to other rare events or different geographic contexts. Future work should generalize the input design to accommodate diverse disruptions and contexts, enhancing robustness and applicability.

Q. Liu et al., 2025) propose CPTR-LM, a collision risk prediction and driver takeover assessment framework leveraging radar–video fused sensor data. Using LLaMA 3, the model interprets textual descriptions of traffic scenes to reason stepwise about vehicle states, relative motion, and risk levels, ultimately suggesting when takeovers are required. The fine-tuned CPTR-LLM, trained with radar-video scene descriptions and conflict annotations, outperforms traditional models such as ARDL, AMG, XGBoost, and LSTM in both accuracy and reliability. Its key strength lies in integrating multimodal data for interpretable and context-aware decision-making in autonomous driving. However, the dataset's limited geographic and environmental diversity restricts generalizability. This can be addressed by expanding data collection across varied weather, temporal, and spatial conditions. Overall, the study demonstrates the promise of LLM-driven, multimodal reasoning for improving safety and situational awareness in intelligent transport systems.

Arteaga and Park, 2025 present an LLM-based framework for detecting underreported factors such as alcohol involvement in traffic crash narratives using LLMs like Flan-UL2 (20B), LLaMA-2-13B-Chat, and GPT-3.5-Turbo. The study found that Flan-UL2 achieved an F1 score of 96% (recall 1.0, precision 0.93), outperforming traditional machine learning baselines such as logistic regression and SVM. The model processes crash narratives through carefully engineered prompts to return a YES/NO output indicating alcohol involvement, thereby enabling automated correction of crash underreporting in safety databases. Its strength lies in the high interpretability and validation of results through human review. However, limited baseline comparisons and a lack of clarity on whether alcohol cues were explicitly present or inferred reduce methodological transparency. Future work should expand baseline models

[15] https://www.barclayscenter.com/events-tickets/event-calendar

and include qualitative examples to better illustrate the model's inference process and robustness.

The paper BRD-LLAMA (Gao, 2025) introduces a framework for bike rental demand prediction using LLaMA, adapted for time series forecasting. The model converts demand data into embeddings and structured prompts that incorporate domain knowledge, with outputs being predicted rental volumes. Fine-tuned with LoRA for efficiency, BRD-LLAMA is evaluated on the London Bike Rental Demand dataset[16] from 2015 to 2017 and achieves significantly lower errors than baselines such as XGBoost, LSTM, Transformer, and AutoFormer, reducing MAE by nearly 50% compared to AutoFormer. Its strength lies in the use of series embeddings and expert-driven prompts, which enable efficient and accurate time series analysis. However, the framework focuses narrowly on demand history, without considering external factors such as weather, holidays, or day-of-week effects. Addressing this limitation by incorporating contextual features could further improve prediction robustness and real-world applicability, making the framework more valuable for transport engineers in optimizing fleet allocation and service planning.

The paper LLMmove (Feng et al., 2024) investigates whether large language models can perform zero-shot next point-of-interest (POI) recommendation without task-specific training. Using GPT-3.5-Turbo, the framework takes user check-in histories along with candidate POIs (including ID, category, and distance) as input and generates top-K ranked recommendations with reasoning. Evaluated using Acc@K and MRR, LLMmove outperforms specialized baselines such as Popu, Dist, CZSR, LLMRank, and LLMMob variants, demonstrating strong generalization in zero-shot settings. Its strength lies in eliminating the need for training, making it adaptable and efficient across datasets. However, performance is highly dependent on the quality of the underlying LLM, with no optimization for task-specific nuances. This limitation could be addressed by incorporating fine-tuning or domain-adapted LLMs to achieve more robust and reliable predictions. The multi-step prompting strategy used further enhances contextual reasoning, making LLMmove a valuable tool for mobility prediction and transport planning.

The paper LLM-Next (Han et al., 2025) addresses the problem of next point-of-interest (POI) prediction by adapting LLMs for spatio-temporal understanding. Using LLaMA-3.1-8B, the authors fine-tune the model with LoRA on datasets such as Foursquare NYC, Tokyo, and Gowalla CA, where user check-in sequences are converted into structured text prompts containing spatial, temporal, and categorical information. The model outputs both the most probable next POI and its reasoning, achieving state-of-the-art performance by outperforming baselines including LSTM, STAN, GETNext, STHGCN, GPT-3.5-Turbo, LLM-Mob, and LLM-Move, particularly when considering the Accuracy@1 metric. Its strengths lie in structured prompt engineering and explainable predictions, making it valuable for transport applications like demand forecasting, route planning, and personalized mobility services. However, weaknesses include text modality conversion loss and prediction redundancy. These can be mitigated by embedding-based representations or improved data-to-text conversion methods to preserve spatio-temporal detail and reduce noise in predictions.

The paper by Qin et al., 2025 proposes a spatiotemporal prompt-driven framework using GPT-3.5-Turbo to predict individuals' next public transport boarding stations. The model structures trip history into prompts across long-, mid-, and short-term time scales, and outputs both the predicted next boarding location and an explanation. Evaluated using accuracy and weighted F1-score, LingoTrip outperforms baselines including 1-MMC, DeepMove, MobTcast, MHSA, and Mob-LLM, with especially strong performance in small and medium sample size

[16] Dataset not provided by authors

scenarios. Its strength lies in leveraging semantic reasoning without the need for task-specific training, enabling robust generalization and practical applications in personalized travel recommendations, crowd management, and disruption handling. However, the framework relies solely on historical trip data and does not incorporate additional contextual factors such as day of week, weather, or special events. Future work should integrate these features to enhance semantic richness and prediction reliability.

Overall, researchers have actively investigated the application of LLMs for classification and prediction tasks in transport engineering. While DDC-Chat, ST-LLM+ and (Yu et al., 2024) reported superior performance, establishing new state-of-the-art benchmarks, other studies like xTP-LLM and KIDCM emphasized the strong generalizability and robustness of LLM-based frameworks, particularly under noisy conditions. A recurring advantage is the ability of LLMs to provide explanatory reasoning alongside predictions.

4.2 Information Mining and Insight Generation

Today, transport systems generate huge amounts of data from vehicles, sensors, infrastructure, and passengers. While this data can be very useful for improving how transport systems work, the large volume makes it hard to manage and understand. Finding clear patterns and useful insights from all this information is not easy with traditional (statistical, clustering and rule-based) methods, which is why new approaches are needed. Information mining plays a crucial role here as it involves searching through massive datasets to find useful patterns, relationships and key facts. However, simply identifying such things is not enough, and hence, we turn to the process of insight generation.

Insight generation is the process of interpreting the mined information in an attempt to understand why we observe the patterns we see, why they matter, and their connection to the real world. It provides us with actionable knowledge and supports better planning and decision-making. This is where LLMs stand out. They use embeddings to capture the meaning and relationships hidden in data, so a timetable entry or a GPS point aren’t just numbers or words, they carry context. The attention mechanism in LLMs lets them focus on the most relevant parts of huge datasets, spotting connections that humans or traditional tools might miss. Through many layers of reasoning, they can combine these pieces into clear patterns and explanations. In this section, we consider how researchers are using LLMs for information mining and insight generation in transport engineering.

Public transport authorities collect a lot of data on travel behaviour; however, the effective utilisation of this data is challenging. To solve this issue, Ulan and Söderman, 2025 introduces an LLM powered framework and develops a prototype application through which public transport authorities can upload their data and then ask questions about the same data via a chatbot interface. The system employs retrieval-augmented generation (RAG), enabling the LLM to generate responses grounded in the uploaded data. For implementation, the authors fine-tuned LLaMA-2 using LoRA on data sourced from the Public Transport Barometer and other transport authorities. Despite the practical relevance of the idea, the study has several shortcomings. It lacks appropriate evaluation metrics, does not compare against baseline methods, and provides no quantitative evidence of its usefulness to transport authorities. Furthermore, as RAG is already a well-established technique, the study’s contribution lies more in domain application than in methodological innovation, limiting its overall scientific significance.

Venkatesh Raja et al., 2024 propose an AI-based case-based reasoning (CBR) system for road accident severity investigation and root-cause analysis, then delivers tailored troubleshooting recommendations, outperforming generic responses like those from

ChatGPT. The system uses similarity indices, retrieval accuracy, and severity scoring to identify and analyse accident patterns, achieving higher retrieval accuracy than ChatGPT, though detailed quantitative comparisons are not provided. By mapping accident attributes comprehensively, the framework supports intelligent decision-making and national road safety planning. Its strength lies in the structured and explainable reasoning process for accident diagnosis. However, the use of large-scale synthetic data (1,000 to 1,000,000 samples) risks introducing distributional bias and reducing real-world applicability. Future studies should validate performance on real crash datasets and quantitatively benchmark against LLMs using standardized prompts and metrics.

Bin Zaman Chowdhury et al., 2024 present Durghotona GPT, a framework that combines web scraping and LLMs to automatically generate structured road accident datasets in Bangladesh. Using GPT-4 (and testing against GPT-3.5, and LLaMA-3), the system categorizes news reports as general or specific accidents and extracts structured details (such as date, location, casualties, and vehicle types) from the latter. GPT-4 achieved 99% accuracy, outperforming LLaMA-3 (96%), GPT-3.5 (83%), and a Google BERT baseline. The approach provides timely, systematic crash data to support traffic safety analysis, urban planning, and public health research. Its strength lies in automating large-scale textual (news) data processing with good precision. However, restricting only to scrapable websites, while ethical, may lead to under coverage of less reported accidents. Future work should incorporate additional news sources and official records to improve completeness and dataset reliability.

The study by Z. Wang et al., 2025 investigates the joint influence of built environment factors on the urban rail peak hour patronage using LLMs. The authors consider a comprehensive set of variables including demographic factors (population density and average house pricing), land use (residencies, business, recreation, finance, healthcare, education, etc), station type (terminal, transfer or CBD), intermodal connections (bus stops, bus lines and parking) and external connectivity (distance to city, road density and number of road intersections). The data was collected from Baidu Map, Lianjia.com, OpenStreetMap, Tianditu and WorldPop; and DeepSeek was used as the LLM. The authors provided the model an engineered prompt which explains the relevant terms and provides guidelines about thinking and performing the analysis task. The LLM was asked to first output predictions for the next five working days, then it was given the ground truth for the next three working days and asked to update its prediction and reasoning for the remaining two days. The LLM identifies the key and latent factors using and introspective CoT analysis and makes refined predictions. It outperformed baselines like ARIMA and XGBoost for prediction task of peak hour ridership.

The study of Jaradat et al., 2024 proposed a multitask learning framework which leverages LLMs to perform real-time analysis of road traffic crashes from social media (X, formerly known as Twitter) data. The framework analyses tweets to perform both classification (e.g. crash severity, sentiment / emotion) and information extraction tasks (e.g. driver details, injury / death, location, contributing factors) to help with the analysis. The authors fine-tuned GPT-2 on approximately 19,000 tweets. The proposed framework outperformed baselines like XGBoost, and an even more capable (but not fine-tuned) LLM, GPT 4o-mini.

Another study by the same author, Jaradat et al., 2025b investigates multimodal data fusion of tabular and textual crash data using large language models to enhance traffic crash prediction and analysis. The study compares GPT-2 (fine-tuned), GPT-3.5, and GPT-4.5, finding that GPT-4.5 with few-shot prompting achieves the best performance—98.9% accuracy for crash severity and 98.1% for driver fault—surpassing fine-tuned GPT-2 and GPT-3.5 across accuracy, F1-score, and Jaccard metrics. The model processes serialized tabular data and crash narratives to classify outcomes such as fault, severity, and contributing factors,

offering interpretable insights for transport safety analysis. Its strength lies in the systematic evaluation of LLM learning paradigms and effective text–tabular integration. However, the study's lack of comparison with traditional ML or statistical models limits the depth of its performance claims. Future work should include baseline benchmarks to better position LLMs within established crash prediction frameworks.

Today, increasing numbers of people are using multiple transport modes in a single trip, hence, it's important to understand multimodal travel patterns to design better, efficient, and sustainable transport systems. The study by W. Li et al., 2025 aims to apply LLMs to better understand the complex multimodal travel patterns. First, the framework segments a multimodal trip and identifies the travel mode for each segment. Next, these travel features are converted into textual representations, which are transformed into semantic embeddings using BERT. BERT is a PLM and not an LLM, but it is excellent at generating embeddings. Finally, DBSCAN clustering is applied to measure the semantic similarity between these embeddings to identify distinct multimodal travel patterns. When the framework is applied to the dataset from the Geolife project (conducted by Microsoft Research Asia), it yields 35 classes or clusters. These include bus-dominated patterns (e.g. walk + bus + walk, bus + walk, car + bus), subway-dominated patterns (e.g. walk + subway, bus + subway, car + subway), car-dominated patterns (e.g. walk + car, car + bike, car + walk) and bike-dominated patterns (e.g. bike + walk, walk + bike). This helps in finding areas with insufficient accessibility to public transport (bus or subway), insufficient car parking, or a lack of bike lanes. The use of BERT semantics proves to be a significant enhancement in such classification tasks over conventional methods in capturing nuanced multimodal travel behaviours.

The paper (Zheng Zhang et al., 2024) evaluates the use of GPT-3.5-Turbo, GPT-4, and Claude-2 for anomaly detection in human mobility trajectories. The models take sequences of check-ins (time, location type, distance) as input and output anomaly judgments, sometimes with explanations. On the Geolife and Patterns-of-Life datasets, Claude-2 outperforms non-deep learning baselines and matches deep learning models, while GPT-3.5 surpasses all methods on Patterns-of-Life. Baselines include OMPAD, MoNav-TT, TRAOD, DSVDD, and DAE. The strength of this approach lies in showing that LLMs, even without task-specific training, can match or outperform specialized anomaly detection models while providing interpretable reasoning. However, the study lacks transparency in how anomaly hints were used and does not extend the analysis to important cases like infectious disease monitoring mentioned in the abstract. Future work should make prompt strategies explicit and broaden the scope of anomaly scenarios to validate generalizability.

The paper T2TrajLLM (L. Liu et al., 2025) presents a method for extracting individual mobility trajectories from textual narratives by combining GLM3, GPT-4, and Qwen2-72B LLMs with structured domain knowledge. Using a schema-guided prompting approach, the model converts free-form travel text into structured JSON outputs that capture actors, locations, and temporal information. Evaluated with accuracy, precision, recall, F1-score, and structural consistency, T2TrajLLM outperforms baselines such as BERT, mT5, and event-extraction models, improving accuracy by around 8% and showing robustness and transferability across contexts. A key strength is its ability to capture contextual details missing in GPS-based methods, enabling applications in urban transport planning, travel behaviour analysis, and even pandemic management. However, ambiguity in natural language remains a limitation, which could be addressed through human-in-the-loop validation. The framework's sophisticated prompt design which combines reasoning steps, constraint rules, and schema guidance enhances consistency, making it a promising direction for text-driven mobility analysis.

LLMs have proven to be effective tools for turning massive amounts of heterogeneous data streams generated by modern transport systems into insights. By leveraging features like embeddings and attention mechanism, LLMs can mine patterns from raw sensor feeds, timetables, GPS traces, and unstructured text, and translate them into interpretable explanations that support planning, safety analysis, demand forecasting, multimodal travel-behaviour understanding, and anomaly detection. This provides a significant advantage over conventional statistical or rule-based techniques in terms of contextual awareness, and in being able to work with disparate modalities and cohesive report generation. Nevertheless, the field still faces critical challenges: the need for rigorous quantitative evaluation against established baselines, mitigation of biases introduced by synthetic or limited datasets, and transparent prompt engineering to ensure reproducibility and reliability. Addressing these gaps will be essential for fully realizing LLMs' potential to enhance decision-making and efficiency across transport engineering domains.

4.3 Question Answering and Knowledge Retrieval

LLMs are very effective at answering questions because they have learned patterns from vast amounts of text, allowing them to connect concepts and infer answers, even when the exact wording hasn't been seen before. The deep neural networks create rich "embeddings" for every word or phrase, capturing subtle meanings and relationships, while the attention mechanism lets them focus on the most relevant parts of a question or context. This combination allows LLMs to understand questions, retrieve and synthesize information, and generate coherent answers, effectively acting as a flexible, general-purpose knowledge source.

In transport engineering, this ability means LLMs can act like smart assistants that quickly make sense of complex, scattered information. For example, they can answer questions from large technical reports, traffic sensor data descriptions, or policy documents without engineers manually searching through them. They can also summarize regulations, compare case studies from different cities, or explain model outputs in plain language. This helps transport agencies save time, make more informed decisions, and bridge the gap between raw data, technical knowledge, and practical planning or operations.

Considering public transport information management, the General Transit Feed Specifications (GTFS) is an open standard used to distribute relevant information about transit systems to passengers. The study by Devunuri et al., 2024 evaluates the capabilities and limitations of LLMs in context of semantic and data-augmented understanding of the GTFS static data. In the GTFS semantic testing, the authors noted that LLMs have some degree of understanding of GTFS concepts, definitions and rules based on official GTFS documents, presumably because their training dataset contained this information. In GTFS retrieval (data augmented) testing, GTFS static data from the Chicago Transit Authority was provided, and the LLM was tasked with answering related queries either by receiving the data directly in the prompt, or by generating an appropriate query to retrieve the required information. The later method gave significantly better results. Through this study, the authors established a foundational understanding of the capabilities and limitations of LLMs.

Building on top of the previous study, the authors introduced TransitGPT (Devunuri and Lehe, 2025), which leverages LLMs to answer natural language queries about static GTFS data via a chatbot interface. The framework generates python code based on the user's query to extract and manipulates the relevant data, which is then given to the LLM to summarise and answer the user query. This approach illustrates how LLMs can handle complex, data-grounded queries rather than relying solely on knowledge from pretraining, demonstrating their potential as practical tools for public transport information management.

The paper by Padoan et al., 2024 presents a conversational framework that leverages GPT-4-Turbo to support decision-making with mobility datasets. The system translates natural language queries into SQL statements, executes them on mobility data, and returns structured answers with reasoning or visualizations. Demonstrations show that the chatbot can handle complex analytical queries and SQL-based reasoning, offering a more intuitive alternative to traditional tools like Tableau or Power BI. Its strength lies in simplifying data analysis for non-technical users, enabling transport engineers to explore scenarios and insights through natural conversation. However, the work lacks formal evaluation metrics or systematic benchmarking, making it difficult to assess robustness and accuracy. Future research should incorporate quantitative evaluations and standardized benchmarks to validate performance and reliability.

Costa et al., 2024 introduce an LLM–powered framework which integrates urban risk levels, cycling infrastructure, and open geospatial data using a three-step pipeline: first, data ingestion and cleaning from tools like CityZones and OpenStreetMap; second, orchestration of LLM-powered agents that process user text or audio queries with prompt engineering; and third decision execution, where the system returns safety insights, risk maps, and tailored recommendations for cyclists via formats like GeoJSON, text, and audio, supporting safer route planning and urban cycling decisions. Built on GPT-3.5-Turbo, the system categorizes cycling route risks into low, medium, or high, providing real-time feedback and safety advice to cyclists and urban planners. Its strength lies in combining LLM reasoning with geospatial analytics for practical transport applications. However, the study lacks real-world validation and baseline comparisons, limiting evidence of its effectiveness. Future work should benchmark the model against traditional safety assessment tools and test it with real-world cycling data to establish reliability and generalizability.

LLMs excel at question answering because their deep embeddings and trained to capture nuanced meanings, while the attention mechanisms they use can isolate the most relevant information, enabling them to synthesize knowledge from vast textual corpora and infer answers even when exact phrasing is different. In transport engineering this translates into versatile “smart assistants” that can parse technical reports, sensor descriptions, policy documents, and domain-specific standards such as GTFS, then retrieve, summarize, or explain the content in plain language for engineers and planners. By coupling natural-language queries with data-augmented workflows, for e.g., generating code or SQL statements to extract and manipulate transit schedules, mobility datasets, or geospatial risk layers, LLMs help bridge raw data and actionable insights, supporting tasks from schedule exploration and scenario analysis to cyclist safety mapping. While these capabilities promise faster decision-making and broader accessibility of complex transport information, the field still requires systematic benchmarking, robustness testing, and real-world validation to ensure reliability and to quantify performance against traditional analytical tools.

4.4 Decision Support and Prescriptive Optimisation

LLMs can help in decision making because they can take unstructured and messy real-world data, summarise and extract key points, and find an optimum solution based on the goal, while explaining their reasoning in plain language. Similarly, they can help in optimisation tasks by reasoning about candidate solutions and choosing and justifying the best course of action to achieve certain goal.

This decision-making ability serves a key role in Traffic Signal Control (TSC) tasks. Preliminary studies like Yiqing Tang et al., 2023 assess the possibility of applying ChatGPT in traffic control scenarios. The authors found that ChatGPT can help traffic managers quickly acquire domain knowledge. They then tasked ChatGPT with analysing road structures using SUMO simulation road network code, which it successfully accomplished. Following this,

ChatGPT was asked to analyse traffic flow and provide control suggestions, and later to translate natural language traffic policies into directly deployable code. While the study presents an interesting application of LLMs, it neither explicitly mentions the version of GPT or nor presents any quantitative measures for the domain knowledge understanding and road network analysis ability of ChatGPT. Moreover, the study does not offer concrete recommendations for improvement, limiting its overall scientific contribution.

The study Y. Tang et al., 2024a explores a novel approach for the designing and implementing green wave control for urban arterial roads using LLMs, specifically GPT-4. The process begins by providing the road network file to the LLM for a quick road network analysis. Next, the LLM is provided with relevant information like network connectivity relationships, distances between intersections, lane speed limits, and traffic signal public cycle information, and is tasked with generating a green wave control policy. Finally, the LLM evaluates the policy by producing a spatio-temporal graph based on the green wave control policy generated during the preceding step. The authors tested the newly generated policy using SUMO and observed an increase in average road speed after deployment of the LLM policy. They concluded that LLMs can be used for both generating and evaluating green wave control policies. However, since the study does not compare the LLM-derived policy against baseline methods, or test in real-world scenarios, the practical value of the approach remains uncertain.

Another study, RAGTraffic (Zhendong Zhang et al., 2024), leverages Retrieval Augmented Generation (RAG) to enhance LLM performance in TSC. In this framework, the LLM agent is provided with scene and task descriptions, general traffic management knowledge, a set of possible actions (i.e. set of lanes granted green signal) and real-time external factors like weather, accidents and construction activities. RAG is employed to enrich the LLM's understanding of the traffic scenario by retrieving pertinent traffic-related common sense and information based on the current input. To further improve this process, the authors introduce Traffic Real-time Information Scanner (TRIS), a mechanism for dynamic RAG activation which considers the token uncertainty and the importance of the token in the context. Experimental results show that the proposed framework outperforms transport engineering-based (e.g. Maxpressure), RL-based (e.g. CoLight and PressLight) and simpler LLM-based baselines on key metrics like Average Queue Length, Average Waiting Time and Average Intersection Throughput, evaluated on datasets from Hangzhou, New York, Jinan, Los Angeles and Tokyo. The authors also highlighted the strong generalisation capability and robustness of the framework under high-traffic scenarios.

The study Movahedi and Choi, 2025 proposes two LLM agents for adaptive traffic control systems. The first agent, referred to as the actor agent in the paper, follows a Zero-Shot Chain of Thought (ZS-CoT) approach. It is provided with a general problem description specifying the task (e.g., generating traffic signal plans) along with real-time state parameters extracted from the simulation environment and is responsible for decision making. The second agent, termed the Generally Capable Agent (GCA) pairs a modified actor agent with a critique agent. The modified actor agent incorporates consolidated text-based knowledge, enabling it to make more informed decisions by leveraging past experience. The critic agent evaluates the actor agent's interactions and updates the consolidated text-based knowledge. Together, the GCA agent follows an actor-critic paradigm, improving performance over time through iterative feedback and knowledge accumulation. Simulation experiment results show that LLM-based controllers outperform traditional baselines (fixed-time, gap-based, and delay-based) at both intersection and approach levels, achieving fewer phase changes, optimized cycle times, reduced vehicle delays, and improved traffic speeds.

The study by Y. Tang et al., 2024b introduces a TSC framework that leverages both LLMs and human expertise. The framework maintains a library of traffic scenarios. Detectors gather data from real traffic systems and assess and select the best traffic strategy from the library. This process is termed “Autonomous Mode” in the paper. If the current traffic scenario is not present in the library, the framework shifts to “Human Feedback Mode”, where the LLM provides the human traffic engineer with a description of the scenario and proposes a strategy, and based on the human feedback, the LLM optimises and deploys the strategy. The framework also includes a “Human Takeover Mode”, in which the LLM translates human instructions into machine commands to guide the strategy optimization module. In this mode, humans can also override the LLM and directly control parameters. Overall, the study primarily leverages LLMs as an interface for human–machine interaction within a memory-based TSC strategy. However, the absence of experiments or benchmarking means its practical applicability is unclear.

Several studies have sought to augment LLMs with the help of agents and tools to enhance their capabilities.

The study by Yao et al., 2025 incorporates vision-based perception into an LLM agent augmented with specialised tools for TSC. The framework utilises YOLO based visual perception, along with tools for identifying possible signal phases and assessing occupancy or clearance urgency (e.g., for emergency vehicles). It further incorporates tools to get initial decisions (from RL model) and then refines and explains a traffic control strategy based on this. Simulation results show that the proposed framework outperforms traditional approaches (fixed-time), RL-based methods (e.g., simple RL and DQN), and simpler LLM-based baselines.

Open-ti (Da et al., 2024b) presents a tool augmented LLM to achieve Turing indistinguishable (human like) traffic intelligence for traffic signal control. The framework leverages a suite of tools to perform complex tasks such as data acquisition, demand generation, simulation, visualisation, optimisation and explanation generation. Experimental results demonstrate that the proposed framework outperforms traditional baselines and is comparable to the state-of-the-art RL-based baselines.

Reinforcement Learning based TSC has shown promising results, however, tuning the weights of a multi-objective reward function remains a significant challenge. To mitigate this, the study (Choi and Lim, 2025) proposes an algorithm which leverages LLMs for dynamic weight adjustment of the RL reward function. In the paper, the authors demonstrate the effectiveness of LLM based RL reward function with the help of an RL based TSC algorithm with the weighted sum of waiting time and queue length as the reward function, where the weight is determined by the parameter α. A higher α gives more weight to waiting time whereas a lower α gives more weight to queue length. The LLM dynamically adjusts the value of α based on its current value, the reward from the last episode, and the value of epsilon (exploration-exploitation balance). Experimental results demonstrate that the proposed optimisation technique-based RL algorithm outperforms fixed-time, simple DQN and simple LLM based baselines. While the study highlights the potential of LLMs in optimizing multi-objective RL reward functions, the contribution is limited, as the authors primarily demonstrate feasibility by instructing the LLM to perform the adjustment without deeper methodological innovation.

Current research pertaining to LLMs and traffic-signal control is converging on a set of LLM-centric paradigms that turn natural-language reasoning into concrete signal-timing actions. One line of work treats the LLM as an “intelligent analyst” that ingests unstructured network descriptions, simulation files or real-time incident reports, extracts key geometric and

operational parameters, and then generates or refines control policies (e.g., green-wave schedules) expressed directly in executable code or configuration scripts. Another strand augments the language model with retrieval-augmented generation or external knowledge bases so that it can pull relevant traffic-management heuristics, weather forecasts, or construction updates before deciding on phase allocations. A third family couples LLMs with tool-oriented agents—vision modules for occupancy detection, simulation engines for rapid policy evaluation, or reinforcement-learning loops whose reward weights are dynamically tuned by the model's textual feedback. Across these approaches the LLM's ability to perform chain-of-thought reasoning, translate high-level objectives into low-level commands, and explain its choices in plain language yields several practical benefits: faster creation of context-aware signal plans, reduced reliance on hand-crafted heuristics, improved adaptability to unforeseen events, and a more transparent human-machine interface that lets engineers audit and steer decisions. Collectively, these methods demonstrate how LLMs can move from passive question answering to active, data-driven optimization in the transport engineering domain.

4.5 Synthetic Data and Scenario Generation

LLMs can mimic the style and structure of real data, so they're useful for creating believable, diverse scenarios that fill gaps in training or testing sets. In transport engineering this ability can be leveraged for generating additional traffic records, rare crash or congestion scenarios, and synthetic sensor readings to cover missing or statistically rare data.

A performance drop is observed when simulator trained RL models are deployed in the real world. This drop can primarily be attributed to differences in the traffic dynamics of the simulator and the real world. To address this, the paper by Da et al., 2024a), introduces PromptGAT, an LLM powered Grounded Action Transformer which leverages the abilities of LLMs to generate synthetic but realistic traffic dynamics to improve the simulator-based training. The authors report that PromptGAT improves the model's performance on key metrics like average travel time, throughput, queue length and delays. However, the evaluation is limited to simulator-to-simulator transfer for reproducibility, rather than simulator-to-real deployment. This limitation undermines the core premise of PromptGAT, mitigating the simulation-to-reality performance gap, and thus it is unclear if this approach can effectively address the domain gap between simulated and real scenarios.

The paper TrajLLM (Ju et al., 2025a) proposes a modular LLM-enhanced agent-based framework for simulating realistic human mobility trajectories, addressing the challenge of aligning simulated movement with real-world patterns. The authors employ LLaMA-3.1 8B and GPT-4o-mini, prompting them with agent persona information and past activities to predict daily activity-location sequences. The framework demonstrates strong alignment with observed mobility data, outperforming baselines such as LLMob (pure LLM) and CoPB (LLM with, but with higher preference to populated and closer PoIs. The approach is hence termed as the 'gravity based model'), while maintaining scalability and interpretability. A key strength lies in its ability to capture behavioural aspects of human mobility for applications in urban planning, traffic management, and public health. However, the work lacks standardized benchmarks and baselines for evaluation, which limits comparability across studies. Addressing this limitation would require the establishment of shared benchmark datasets and evaluation protocols, enabling more rigorous validation and broader adoption of such LLM-driven mobility simulation frameworks.

The paper by Y. Zhang et al., 2024 explores the use of LLaMA-3 8B and GPT-2 to generate realistic daily human activity schedules for urban populations. By fine-tuning LLaMA-3 8B with

LoRA on the Tokyo Metropolitan Person Trip Survey dataset[17], the model outputs time-stamped sequences of activities based on personal attributes such as age, gender, and occupation. The results show that fine-tuned LLaMA-3 produces more diverse and realistic patterns than GPT-2 and vanilla LLaMA-3, with improvements in spatiotemporal consistency and activity diversity. Strengths include the systematic evaluation of prompting strategies (zero-shot, few-shot, and chain-of-thought) alongside fine-tuning, and the semantic grounding of generated activity patterns, which provides richer insights than trajectory-based approaches. However, the study only benchmarks against LLM baselines, with limited focus on spatial activity prediction, and some generated examples are overly simplistic. Future work should incorporate non-LLM baselines and extend the framework to spatial activity generation for broader applicability in transport and urban planning.

Clearly, a growing body of work treats LLMs as generative engines for filling the chronic data shortages that hinder transport research and reinforcement-learning deployment. By prompting LLMs with domain-specific schemas, agent personas or demographic attributes, researchers synthesize traffic records, rare crash or congestion episodes, sensor streams, and full daily activity schedules that mimic the statistical properties of real observations while preserving diversity. These synthetic streams are then injected into simulators to enrich training environments, thereby narrowing the simulator-to-real gap for RL controllers; they also serve as realistic mobility traces for urban-planning analyses, traffic-management studies, and public-health modelling. The key advantage lies in the LLM's ability to reproduce nuanced temporal patterns and human-behavioural logic without exhaustive manual data collection, enabling rapid scenario creation, stress-testing of algorithms under low-probability events, and more robust policy learning. As a result, transport engineering workflows become less constrained by scarce or biased datasets and can evaluate solutions on richer, higher-fidelity synthetic worlds that better reflect real-world dynamics.

4.6 Visual Scene Understanding

Multimodal large language models (MLLMs) and vision-language models (VLMs) are trained on huge numbers of images paired with descriptions. As such they learn to link objects appearance to text descriptions, learning to focus on the parts of an image that matter. This makes them effective at visual understanding tasks, such as recognising objects, spotting unusual situations, and describing relationships and actions in a scene — even when visual information is noisy, partially occluded, or unfamiliar. In transport engineering, and especially for traffic safety, this ability has many practical applications: these models can flag hazards (e.g., pedestrians stepping into traffic, stopped vehicles, debris), help predict risky behaviour, assist real-time monitoring of intersections and crash scenes, and speed up post-event analysis by turning video into clear, searchable reports. Because the models generalise across different camera views and can work with sparse labels, MLLMs and VLMs make safety systems more scalable and better at catching the rare but dangerous events that matter most.

The increasing complexity of traffic dynamics highlights the need for more advanced approaches for traffic safety description and analysis. To help with this, the study by Trinh Xuan et al., 2024 makes two key contributions. First, the authors propose a novel LLM powered segment extraction methodology which leverages LLMs to extract specific segments (appearance, location, environment, attention and action) from the video captions from the WTS vehicle and pedestrian safety dataset[18]. Second, the extracted segments are used to train Qwen-VL, a VLM developed by Alibaba Cloud, to perform the same extraction directly

[17] Source not provided by the authors

[18] https://woven-visionai.github.io/wts-dataset-homepage/

from video frames. The authors report that this approach leads to significant improvements in video-frame-to-caption generation accuracy.

Ahmed et al., 2024 proposes a framework for automated highway safety management which integrates multiple AI components. Specifically, YOLOv11 is used for real-time detection of critical traffic elements and anomalies like accident and fires; the Moondrea2 VLM is used to generate detailed scene description; and the GPT-4 LLM is used to produce incident reports and suggestions. The data is sourced from two public Roboflow datasets[19], Google creative commons licensed images, and ERA dataset. The framework also helps traffic management and emergency responses by notifying the relevant authorities automatically with accident information and actionable insights. The framework proves to be quite effective in simulated environment.

Tami et al., 2024 demonstrate how multimodal large language models (MLLMs) can be applied for automated detection and analysis of traffic safety-critical events using driving videos. Employing Gemini-Pro-Vision 1.5 with few-shot learning, the model interprets dashcam frames and structured, object-level QA prompts to identify scene types, vehicle directions, agents, risks, and recommended actions. The framework achieves around 79% accuracy, outperforming LLaVA-1.5 and other visual–language QA baselines, thereby showcasing the potential of MLLMs for scalable and interpretable traffic safety assessment. Its main strength lies in integrating visual perception with reasoning-based question answering to perform end-to-end hazard evaluation. However, being only tested on a single model (Gemini) and demonstrating only limited reliability in estimating hazard distances constrain its robustness. These weaknesses can be addressed by incorporating multiple MLLMs for comparative validation and enhancing spatial reasoning modules for distance estimation.

Calenzani et al., 2024 evaluate the effectiveness of GPT-4V in classifying risky driver behaviours from short video samples, using only a few frames per clip. The model processes 10 evenly spaced frames from 20-second videos to detect behaviours such as yawning, smoking, mobile phone use, and distraction, achieving accuracy up to 98.9%, often matching or exceeding human experts and outperforming traditional computer vision baselines. This demonstrates GPT-4V's strong multiclass behaviour recognition capability and potential for automated driver monitoring in transport safety and fleet management. The study's main limitation is its exclusive reliance on GPT-4V without comparisons to other LLMs or hybrid models, limiting generalizability. Future work should incorporate comparative benchmarking and explore model ensembles to validate robustness across broader driving conditions.

Yulianda et al., 2024 presents an Indonesian traffic sign recognition system combining deep learning and interactive voice feedback to improve driver safety. Using YOLOv8 for traffic sign recognition, the model achieves 97.5% mAP50 and 82.9% mAP50–95, outperforming baseline configurations and prior studies. GPT-3.5-Turbo and Google Text-to-Speech (TTS) are then integrated to generate concise voice instructions based on detected traffic sign names, enhancing driver awareness. The system demonstrates strong recognition accuracy and potential for real-time driver assistance applications. However, the LLM component plays a minimal role, offering limited added value. Future work should expand the LLM integration to provide context-aware, adaptive guidance that accounts for environmental and situational nuances, better leveraging the reasoning strengths of large language models.

Shafiq et al., 2024 propose a vision language model (VLM)–based approach for road anomaly detection in autonomous driving, using the Florence-2 VLM to identify unexpected obstacles

[19] https://universe.roboflow.com/yolo-and-car-accident-detection-xaltb/accident-detection-77mha and https://universe.roboflow.com/traffic-ai-8xnmy/car-fire-yssjr

from images. The model, fine-tuned via LoRA on the Lost and Found dataset (Pinggera et al., 2016) and tested on Road Anomaly dataset (Lis et al., 2019), achieves mAP@50–95 scores of 50.1 and 45.8, outperforming DETR, YOLOS, and YOLO V10, and competing closely with state-of-the-art baselines like Language Anchors and Perspective Aware models. The VLLM uses natural language prompts and visual inputs to generate bounding boxes and descriptive outputs, enhancing interpretability and situational awareness. Its strength lies in combining visual grounding with linguistic reasoning for robust anomaly detection. However, generalizability across diverse environments is not evaluated, limiting external validity. Future work should quantify cross-dataset performance and expand fine-tuning to varied geographical and lighting conditions to strengthen adaptability.

Xu et al., 2024 propose a multimodal LLM-based warning system for personalized driver assistance, leveraging GPT-4V and Mixtral-8x7B. The system inputs scenario images and driver profiles (age, driving experience, etc) and outputs multimodal warnings with detailed reasoning, tailoring alerts (visual, auditory and haptic) to individual drivers. While no quantitative baselines are reported, the framework demonstrates explainable, adaptive, and driver-centric warnings across diverse situations, enhancing interpretability and situational awareness. Strengths include user-focused personalization and integration of multiple feedback modalities. However, the lack of quantitative benchmarks, dataset details and RAG outputs limits reproducibility and assessment of generalizability. Future work should document data sources, evaluate performance against benchmarks, and quantify personalization effectiveness to validate system reliability.

Arefeen et al., 2024 present TrafficLens, a multi-camera traffic video analysis framework that accelerates video-to-text conversion using InternLM-Xcomposer2 (1.8B) and LLAVA-1.5-v2-7B, with ChatGPT employed in a RAG pipeline. The system ingests video chunks from multiple cameras that are converted into text, and responds to user queries about traffic scenes, achieving 2–4× faster processing than per-camera VLM baselines while maintaining or improving information diversity (BERT and ROUGE-L scores). This enables real-time monitoring, incident analysis, and intersection activity extraction for transport engineers. Its strength lies in efficiently fusing multi-camera feeds into a coherent textual understanding. However, the framework does not address limitations of VLMs such as their poor performance in counting objects, which may affect quantitative traffic metrics. Future work should evaluate and integrate counting capabilities to improve accuracy for vehicle and pedestrian quantification.

Lee et al., 2024 develop a multi-modal walking safety system for the visually impaired, integrating YOLOv5 nano for real-time object detection with the KoAlpaca LLM for natural language guidance in Korean. Detected object labels (e.g., “bollard,” “crosswalk,” “red light”) are fed into the LLM to generate context-specific warning sentences. The system achieves 88.84% object detection accuracy and 98.68% recognition rate, with positive pilot usability feedback (4.05/5), while maintaining efficiency on edge devices like a Jetson Nano. Strengths include a lightweight, robust framework that delivers personalized, interpretable guidance on devices like the Nvidia Jetson Nano. Limitations involve an inability to explicitly locate objects and challenges detecting fast-moving entities. These could be addressed by integrating LiDAR or enhanced sensor hardware and refining detection algorithms.

R. Zhang et al., 2025 introduce SeeUnsafe, a MLLM framework for video-based traffic accident analysis using GPT-4o. By combining structured textual and visual inputs such as segmented clips, object descriptions, and scene context, the model classifies accident events and provides interpretable justifications. SeeUnsafe achieves 76.3% accuracy and performs successful visual grounding, outperforming baselines including GPT-4o (vanilla), GPT-4o

mini, LLaVA-NeXT, and VideoCLIP across F1, BLEU, ROUGE-L, and the newly proposed Information Matching Score (IMS). The Information Matching Score (IMS) is a metric that uses a multimodal language model to rate semantic alignment between generated structured responses and ground truth for key safety attributes—scene context, object description, and justification—on a scale from 0 to 100, averaging scores for robust evaluation. Its strength lies in integrating reasoning with visual perception for transparent, automated crash analysis. However, the lack of clarity regarding supporting vision models (e.g., for object detection or segmentation) limits reproducibility. Future research should explicitly document model components and configurations to enhance transparency and replicability in multimodal transport safety applications.

Abdelrahman et al., 2025 introduce Video-to-Text Pedestrian Monitoring (VTPM), a privacy-preserving pedestrian monitoring framework that converts intersection video data into structured textual reports using phi-3-mini for real-time and phi-3-medium for historical analysis. Fine-tuned via QLoRA on over 600 Q&A pairs from research and real-world pedestrian data, the models outperform baselines (LLaMA-3-8B, Mistral-7B, Gemma-7B) on BLEU, METEOR, and ROUGE metrics and are rated highest in domain-specific summarization by human experts. The system processes extracted video features (counts, violations, conflicts, weather, time, location) to generate interpretable narratives, enabling real-time safety monitoring and historical trend analysis while preserving privacy. Strengths include automated, privacy-conscious reporting and clear narrative outputs. Limitations involve potential omission of critical safety details and lack of spatial context, which could be addressed by fine-tuning smaller models further and integrating spatial sensing, e.g., LiDAR.

Peruski et al., 2025 present an edge AI framework for traffic monitoring and anomaly detection using multimodal LLMs. The models process video frames and generate descriptive text summaries along with incident classifications, enabling real-time traffic analysis and rapid accident detection, particularly in remote areas with limited connectivity. The study evaluates LLaVA (LLaMA-based), VILA, and GPT-4 Vision, finding that VILA achieves the highest accuracy and most consistent response times on edge devices. Strengths include effective edge deployment through knowledge distillation and quantization-aware training and comprehensive prompt engineering for in-depth incident analysis. Limitations involve dataset imbalance, as training data are primarily urban while deployment targets rural environments. Future work should curate representative rural datasets to improve generalizability. The LLaVA model was fine-tuned iteratively using the CADP traffic video dataset, enhancing incident classification performance.

Risk prediction and accident analysis are essential components of traffic safety. However, performing such tasks require complex AI models (including LLMs) which are computationally complex, making it difficult to run them on vehicles. To mitigate this, the study (Hu et al., 2025b) proposes the use of cloud and edge computing to offload such complex workflows. The framework consists of three key modules. The first module leverages LLMs for road segment risk prediction from textual data. The second module performs the risk assessment based on the visual scene. The third module employs LLMs to classify the severity of accidents. In all cases, fine-tuned LLMs (GPT-o1, LLaMA-3.2 and deepseek-R1) outperformed ML model baselines (SVM, MLP an LSTM). The LLMs were fine-tuned using LoRA on an Austrian Highway dataset (Schlögl et al., 2019), Germany Highway dataset (de Winter et al., 2023) and UK traffic accident dataset (Grigorev et al., 2025) for the three modules respectively. While the proposed method demonstrates impressive results, concerns like network connectivity and data privacy persist with respect to the actual implementation.

The study (Ding et al., 2025) proposed an Urban Road Anomaly Visual Large Language Model (URA-VLM) to detect various anomalies (uneven surface, floods, fallen trees, fires, traffic accidents, garbage and traffic congestion) on urban roads, with the aim of improving traffic safety. The framework also incorporates multi-step prompting and RAG for anomaly detection, flood depth estimation and safety level assessment. URA-VLM uses Intern-VL optimised with prompts and RAG and outperforms deep learning models like ResNet34 by achieving an accuracy of over 93% as compared to the 81% by ResNet34.

While a lot of studies focus on accidents, near-miss traffic incidents offer crucial predictive insights for preventing crashes, the study by Jaradat et al., 2025a proposes a methodology for near-miss detection from crowdsourced videos. First, the authors evaluated three deep learning model (CNN, Vision Transformer and CNN+LSTM) to segment videos into clips that capture potential near-miss or crash incidents. To achieve this, they first leverage the vision models like CNN to perform frame-level classification. Then perform a rule based even level classification. These segments are then evaluated using an MLLM (GPT-4o) to generate narrative description. For video segmentation, the authors found CNNs to perform the best, and noted that for narrative generation, GPT-4o excelled at generalizing from diverse context without additional examples.

As we progress towards connected autonomous vehicles (CAVs), new challenges emerge. For example, malicious agents may use fake signs or broadcast forged motion information in order to deceive CAVs. The study by Hu et al., 2025a leverages LLMs to mitigate such malpractices. For fake sign detection, the framework identifies the sign and its distance, then queries the LLM to determine whether the sign is genuine. For forged motion information, the framework performs range plausibility analysis (whether the broadcaster is in the communicable range of the detector), location plausibility analysis (e.g. a vehicle cannot be inside a building) and speed plausibility analysis (exceeding speed limit or too slow). The authors report that LLMs (GPT-40, LLaMA-3.2 and Gemini 1.5 Pro) outperformed DNNs (ResNeXt, YOLOv8 and VGG 19) in fake sign detection, with GPT-4o achieving the best accuracy of 86%, while traditional ML models outperformed LLMs for forged motion information detection, with LSTM performing best with accuracy of 92%. It is to be noted that while GPT-4o performs the best amongst the baselines, its accuracy is still 86%, hence it does not mitigate all risks, and some things can still be missed.

It is worth noting that as VLMs and MLLMs are a new development, and not as refined as text only LLMs. As such, a lot of current approaches pair LLMs with existing neural network based solutions (like YOLO) for certain vision related tasks rather than using a single integrated model.

Recent studies show that VLMs and MLLMs are rapidly reshaping traffic-safety engineering by coupling visual perception with textual reasoning. Recent work demonstrates end-to-end pipelines that extract salient video segments, generate detailed incident reports, and personalize warnings using driver profiles; few-shot vision–language models can identify hazardous scenes from dash-cam footage, while large-vision models achieve robust anomaly detection across diverse road conditions. Integrations of object detectors, VLMs, and LLMs enable faster multi-camera video-to-text conversion and privacy-preserving pedestrian monitoring, and cloud-edge architectures offload heavy reasoning tasks to maintain real-time performance. Across studies, strengths lie in improved interpretability and scalability, yet common shortcomings include limited benchmarking, weak spatial reasoning, dataset bias, and opaque component configurations that hinder reproducibility. Future research should pursue standardized multimodal benchmarks, cross-domain validation, tighter sensor fusion

(e.g., LiDAR), on-vehicle lightweight inference, and transparent pipelines to fully realize safe, reliable autonomous transport systems.

5 Applications of LLMs in Autonomous Driving

Autonomous Driving (AD) is one of the most active fields at the intersection of LLMs and Transport Engineering. Researchers have applied LLMs (and VLMs and MLLMs) for a wide range of AD tasks such as perception, planning and decision making, predicting trajectories, human machine interface and synthetic data generation. Because the primary AD literature is both vast and already well-organized by previous reviewers, this paper adopts a high-level review approach and examines those existing review papers rather than re-reviewing hundreds of individual studies. We briefly summarise the findings, gaps and future directions identified by these papers. All the papers are enlisted in the **Table 2** below.

Table 2 LLM in Autonomous Driving Review Papers

Paper	Focus
(Li et al., 2024)	The paper focuses on how large language models (LLMs) are transforming autonomous driving by enabling more human-like decision-making, reviewing their applications in both modular and end-to-end AD systems, highlighting recent advancements, key challenges, and future research directions to bridge LLMs and human-centric autonomy.
(Zhou et al., 2024)	The paper provides a comprehensive survey and outlook on the applications, challenges, and future directions of Vision-Language Models (VLMs) and Large Language Models (LLMs) in autonomous driving systems.
(Cui et al., 2024c)	The paper systematically surveys recent advances, applications, challenges, and future directions of multimodal large language models (MLLMs) in autonomous driving systems.
(J. Li et al., 2025)	The paper comprehensively reviews and synthesizes the current state, applications, challenges, and future directions of large language models and multimodal large models in autonomous driving systems.
(Ashqar et al., 2025)	The paper focuses on reviewing and empirically evaluating the advancements, challenges, and applications of Multimodal Large Language Models (MLLMs) for object detection and reasoning in autonomous driving and transportation systems.
(Sathyam and Li, 2025)	The main focus of the paper is to survey and critically analyse the core capabilities, applications, and challenges of foundation models (which encompasses LLMs and VLMs and MLLMs) in advancing autonomous driving perception systems.
(Zhu et al., 2025)	The focus of the paper is to comprehensively review how large language models are being applied to various aspects of autonomous driving, including perception, decision-making, and interaction within intelligent vehicles.

5.1 Current Applications

LLMs can perform a multitude of AD related tasks. These are each discussed briefly below.

Perception

Models like LiDAR-LLM (Yang et al., 2025) leverage LLMs to understand and reason about 3D scenes from sensor data (LiDAR + camera), providing rational plans and explanations. Other frameworks such as Talk2BEV (Choudhary et al., 2024), BEV-TSR (Tang et al., 2025),

DriveVLM (Tian et al., 2024), OmniDrive (S. Wang et al., 2025), GPT4V-AD (Wen et al., 2023b), CarLLaVA (Renz et al., 2024) use MLMs and LLMs to generate semantic scene descriptions, enhance vehicle perception by synthesising a bird's eye view and multi-view sensor formats, and support reasoning for control and safety decisions.

Prediction

The methods of LC-LLM (Peng et al., 2025), LLM-PCMP (Zheng et al., 2024), LG-Traj (Chib and Singh, 2024), Traj-LLM (Ju et al., 2025b) all convert driving scenes and historical agent trajectories into natural language or visual prompts for intention prediction and trajectory forecasting. The use of chain-of-thought reasoning and LLMs increases interpretability and accuracy compared to traditional black-box deep learning models.

Planning and Decision making

GPT-Driver (Mao et al., 2023), LLM-ASSIST (Sharan et al., 2023), VELMA (Schumann et al., 2024), DaYS (Cui et al., 2024a), DiLu (Wen et al., 2023a), MTD-GPT (Liu et al., 2023), RRaR (Cui et al., 2024b), EoLLM (Tanahashi et al., 2023) all use LLMs to generate real-time planning decisions, interpret navigation instructions, associate landmarks, reflect and improve decisions via memory/reflection modules, and address multitask planning in complex scenarios. LLMs provide explainable language-based outputs, improving transparency and safety.

Scenario Generation

DKD (Yun Tang et al., 2023), TARGET (Deng et al., 2023), ADEPT (Wang et al., 2022), OmniTester (Lu et al., 2024), LCTGen (Tan et al., 2023), TransGPT (Wang et al., 2024) all use LLMs to automate scenario creation for simulation and testing, including rule-based scenario scripting, parsing accident reports to structured data, creating scenario ontologies, and generating dynamic traffic scenes from natural language descriptions.

Human Machine Interaction

LLM based frameworks like Drive as you speak (Cui et al., 2024a) and Receive, Reason and React (Cui et al., 2024b) facilitate more natural communication between occupants and autonomous vehicles by allowing human like interaction with the vehicle.

Multitask Processing

DOLPHINS (Ma et al., 2025), EMMA (Hwang et al., 2024), TrafficGPT (S. Zhang et al., 2024), ESR(Nouri et al., 2024), LMDrive (Shao et al., n.d.), DriveGPT4 (Xu et al., 2024) use LLMs and MLMs to unify multiple driving tasks (e.g., object detection, road map estimation, motion planning, control signal generation) into a single framework, often combining vision and language inputs with real-time task allocation and execution.

5.2 Current Challenges

Inference Speed and Computational Cost

LLMs require substantial computational resources, and their inference latency can be problematic for real-time, safety-critical tasks like planning and control. Deploying these large models on-board vehicles is challenging due to power, compute, and cost limitations. Offloading inference to remote servers raises issues with bandwidth, reliability, and response time. This issue is highlighted by all the review papers.

Hallucination

LLMs sometimes generate information that is plausible-sounding, but inaccurate or fabricated. This presents substantial risk in autonomous driving, as incorrect outputs can threaten safety-critical decisions. Models sometimes invent ("hallucinate") objects or attributes which are not present in the scene, leading to dangerous perception or decision errors in AD settings. This issue is also highlighted by all the review papers.

Limited Physical and Common-Sense Understanding

While LLMs excel at handling language data, they lack an intrinsic grasp of the physical world, dynamics, or common sense—key for navigating and making decisions in complex real-world environments. Their knowledge is based on patterns in text or data, not direct experience or sensorimotor interaction. (Li et al., 2024)

Fine-Detail and Small Object Detection

MLLMs are less reliable at identifying small, occluded, or fine-detail objects (like thin road barriers or faded markings), which are critical for AD safety but easily overlooked. (Ashqar et al., 2025)

Generalization in Diverse and Unseen Scenarios

LLMs have made progress in traditional NLP and vision-language settings, but they often fail to generalize to out-of-distribution or real-world driving scenarios without extensive retraining and data augmentation. (Ashqar et al., 2025)

Lack of Unified Multi-Modal Reasoning

Effectively grounding LLMs in physical driving contexts requires improved fusion of diverse inputs (visual, geographic, textual) for deeper situational understanding. Current multi-modal training frameworks need further advancement. Most current approaches rely only on vision and language, while robust AD systems need to combine LiDAR, radar, and other sensors. Absent or naive fusion limit the model's reliability. (Zhou et al., 2024; Sathyam and Li, 2025)

Ethical, Social and Legal challenges

Issues of privacy, accident liability, and the lack of clear regulatory standards complicate both development and practical deployment at scale. Issues like bias, fairness, privacy in language-driven models, and risk mitigation for real-world deployment are underexplored. (Li et al., 2024; J. Li et al., 2025)

5.3 Future directions

Real-Time Multimodal Fusion and Reasoning

There is a need to design models that can reliably fuse heterogeneous sensory data (images, LiDAR, maps, and text) and reason over these modalities in real-time. This includes improved architectures for modality alignment, feature representation, and language-vision fusion pipelines. (Ashqar et al., 2025; J. Li et al., 2025)

Optimizing Latency & Model Efficiency

Innovative solutions in model compression, quantization, pruning, and distillation are required to make very large foundation models (LLMs/VLMs/VFMs) feasible for real-time, safety-critical autonomous vehicle applications. Similarly, the development of efficient deployment strategies and hardware-aware AI pipelines is essential. All the papers highlight this.

Safety, Reliability, and Real-World Validation

There is a lack of rigorous real-world testing and benchmarking protocols to that expose models to edge cases, rare events, and operational safety challenges. These are needed to test and refine models using diverse datasets (beyond simulation), including live road conditions, traffic incidents, and adverse weather. This is also highlighted by all papers.

Spatial and Temporal Reasoning

Existing methods are limited with respect to spatial reasoning, with foundation models limited in their ability to interpret complex 3D environments, long-tail unstructured scenes, and intricate physical relationships. Improve spatial and temporal reasoning is needed to drive future work in temporal modelling so systems can anticipate, predict, and reason over future states and object permanence, rather than just instantaneous perception. (Cui et al., 2024c; Zhu et al., 2025)

Integration with City Planning and ITS

(Ashqar et al., 2025) highlights that multimodal large language models (MLLMs) can help transport engineering by integrating diverse data streams—such as images, text, and sensor data—to enhance object detection for autonomous driving, infrastructure planning, and intelligent transportation systems (ITS). Specifically, it emphasizes that MLLMs can support city planners and traffic management by providing detailed insights into traffic patterns and pedestrian behaviour, enabling better infrastructure design and optimized traffic control. The future directions include expanding MLLM applications through richer data fusion (e.g., adding LiDAR and radar), improving model robustness to complex real-world scenarios, and reducing computational barriers for large-scale, real-time deployment. Ultimately, the paper envisions MLLMs as foundational elements for next-generation, smart, and adaptive urban transportation networks that are safer, more efficient, and capable of meeting the demands of evolving urban environments.

6 Discussion

6.1 Performance and Novelty of LLM Methodologies

Some tasks that were previously considered difficult or impossible are now achievable thanks to LLMs. We start this section by highlighting these novel capabilities based on the literature reviewed so far.

Ulan and Söderman, 2025 introduces a novel framework for textual data management. The framework allows analysis of huge volume of data through a chatbot interface. This was not possible until the advent of LLMs as no prior AI model could understand the question and provide an appropriate answer by identifying relevant information from a large corpus. However, LLMs combined with RAG allows us to do this.

TransitGPT (Devunuri and Lehe, 2025) is another framework which allows users to retrieve data about public transport services via a chatbot. Under the hood, the LLM receives the query, generates python code to fetch data, and based on the output of the code, provides the user with an appropriate response.

In (Trinh Xuan et al., 2024) , Qwen-VL (a VLM) demonstrated an impressive capability to generate traffic safety specific captions from given images. Similarly, in (Jaradat et al., 2025a), GPT-4o (an MLLM) demonstrates impressive ability to generate textual narration of the given video segment.

End-to-end autonomous driving frameworks with the ability of natural language prediction are only possible due to LLMs and VLMs/MLLMs. These models can comprehend a visual scene, assess and predict vehicle trajectories and make driving decisions, all while providing the users with a natural language interface.

In several other areas, the use of LLMs has seen methods match or surpass the performance of existing state-of-the-art methods.

ST-LLM+ (C. Liu et al., 2025), a spatio-temporal traffic prediction framework which establishes a new state-of-the-art by leveraging the advanced autoregressive abilities of LLMs and overcoming its inability to interpret spatial and temporal data by introducing spatiotemporal embeddings and PFGA.

DDC-Chat (Liao and Lin, 2025) establishes a new state-of-the-art for distracted driver classification by augmenting a VLM with tools (pose detection, instance segmentation, etc) to overcome their limitations for these specific downstream tasks.

BEV-TSR (Tang et al., 2025) proposes a framework which takes in descriptive text as input to retrieve corresponding scene in BEV, and it achieves a new state-of-the-art on the novel nuScenes-retrieval benchmark.

LLM-COD (Yu et al., 2024) establishes a new state-of-the-art for cross city OD prediction by instruction-tuning LLMs on OD dataset to capture transferable urban mobility semantics, integrating rich spatial and POI features and aligning predictions using a novel loss function.

LLM-Next (Han et al., 2025) establishes a new state-of-the for POI prediction. It achieves this by transforming spatio-temporal trajectory and POI data in textual prompts for LLM, combined with adaptive statistical profiling of user behaviour, and fine tuning an LLM using LoRA.

An observation to be made here is that all the frameworks which have achieved a new state-of-the-art fine-tune the LLM. This implies that fine-tuned LLMs perform quite well for downstream tasks.

6.2 LLM Strengths in Transport Engineering

LLMs bring unique versatility and intelligence to transport engineering. Their ability to understand and reason in human language allows them to perform a number of complex tasks which the other models might struggle with. In this section, we look at the benefits that LLMs bring throughout the literature.

Explainability

One of the distinctive advantages of LLMs in transport applications lies in their explainability. Unlike conventional deep learning models that often operate as black boxes, LLMs can articulate the reasoning behind their predictions using natural language, offering transparency into the decision-making process. This capacity to generate human-interpretable explanations not only enhances trust and accountability but also provides domain experts with insights that can support validation, error analysis, and informed decision-making in safety-critical contexts such as transport engineering. xTP-LLM (Guo et al., 2024) is a great example, which offer competitive prediction capability to the state-of-the-art while providing human understandable text explanations.

Generalisability and Robustness

LLMs have shown notable strengths in terms of generalisability and robustness. Unlike many deep learning models that struggle to adapt across regions or datasets, LLM-based frameworks have demonstrated the ability to transfer effectively across diverse geographic contexts and data distributions (Guo et al., 2024; Beneduce et al., 2025), even in zero-shot settings. Moreover, several studies highlight their resilience to noisy or imbalanced data (Guo et al., 2024; Jiang et al., 2025), maintaining competitive accuracy under challenging conditions. These properties make LLMs particularly promising for transport applications, where data heterogeneity, sparsity, and variability are common challenges.

Natural Language Interface

LLMs allow for effortless human-machine interaction by allowing the users to communicate in natural (human) language. Transport engineers can ask questions in plain English and receive textual and numeric answers, plus a short reasoning trace, (Ulan and Söderman, 2025). This reduces the barrier to entry for advanced transport systems and reduces reliance on specialist data scientist.

Multimodal Data Processing

LLMs can handle structured as well as unstructured data. This makes them very useful for transport data processing, where the LLMs can ingest a wide range of data such as accident reports, sensor logs, GTFS tables, JSON data, etc (Devunuri and Lehe, 2025). VLMs and MLLMs go beyond text and numbers and are able to ingest visual data and have shown impressive visual scene understanding capabilities (Ding et al., 2025). This reduces the pre-processing time and complexity and makes the solution more versatile and robust.

No need for training

Before LLMs, most ML models needed to undergo an extensive training process on a specific dataset, and their learnings were largely non-transferable. For example, a classifier model trained to distinguish cat images from dog images could not be used to predict the future prices of stocks. However, LLMs have demonstrated good capabilities to perform a multitude of tasks out-of-the-box, without being explicitly trained on the task specific datasets. In context of transport engineering, take the example of an RL based TSC algorithm. It needs to be trained extensively with the help of data, rules and simulation. An LLM based TSC however does not need that. The model can simply be explained the intent, and it will reason to reach to the optimum controller decisions. The point is that models like RL will give out gibberish values without training, whereas LLMs will output sensible values without being trained. With this being said, approaches like fine-tuning do help improve the performance of the LLM for a specific task.

Adaptability

When the performance of LLMs falls short for a given task, they can be specialized through fine-tuning (DDC-Chat (Liao and Lin, 2025)), or their outputs can be grounded in external data using retrieval-augmented generation (RAG) (RAGTraffic (Zhendong Zhang et al., 2024). This adaptability makes LLMs highly versatile, enabling them to address the diverse and evolving demands of transport engineering applications.

Imputation and Augmentation

LLMs can tap into their internal pool of knowledge and world-understanding to generate data. With only a few example prompts, they can generate demand forecasts, travel-time distributions and even signal timing suggestions. They can answer “what-if” questions by adjusting their chain of reasoning and allow for testing of hypothetical but realistic scenarios.

This ability can also be leveraged for data imputation. PromptGAT (Da et al., 2024a) is an example of this.

6.3 LLM Weaknesses in Transport Engineering

6.3.1 LLM Weaknesses Specific to the Transport Domain

While the critical limitations of individual studies can be found in Appendix A, this section intends to discuss the limitations across papers, providing deeper insights.

Lack of Strong Baselines

Many studies do not evaluate their methods against strong baselines. Papers like (Y. Tang et al., 2024a, 2024b) lack any comparative results despite tackling well defined problems like Traffic Signal Control. Studies like (Zhen et al., 2024; Movahedi and Choi, 2025) compare exclusively with transport specific, statistical or ML models, without considering other LLMs. Furthermore, papers such as (Ulan and Söderman, 2025) introduce entirely new frameworks that lack a direct baseline, making relative performance difficult to assess. Therefore, it is important for publications to test the proposed framework extensively against good baselines to establish the usefulness of the framework.

Contradictory results

Often studies point to opposite facts related to the applications of LLMs for certain transport specific downstream tasks. For example, the studies (Feng et al., 2024) and (Han et al., 2025) both tackle the next Point-of-Interest (POI) prediction task for human mobility, however the first argues that the zero-shot prediction capabilities of LLMs are sufficient, whereas the later argues that they are insufficient and fine-tuning is needed. This is also partly due to the benchmarks and baselines (emphasising the previous point) being different in the two cases.

Minimal scientific contributions

Papers like (Aldieri and Voß, 2023) serve more as a showcase paper rather than scientific contributions. Others, while demonstrating how LLMs can be used, don't validate the usefulness through evaluations. For example, Y. Tang et al., 2024b proposes an LLM powered framework for urban traffic signal control, however the paper doesn't discuss any metric, or qualitative benefit of the method, undermining its impact.

6.3.2 General LLM Weaknesses

High computational complexity

State-of-the-art LLMs require significant computational resources for training and inference. For example, OpenAI GPT-3 175B required 314 zettaflops (1 zettaflop = 10^{21} floating point operations) for training (Brown et al., 2020). The Meta LLaMA 3 405B model with 16-bit floating point precision requires 810GB of VRAM for inference[20] . This makes real-time edge deployment, say at an intersection for traffic signal control, impractical. Instead, current solutions rely on cloud computing which drives up costs and complexity and increases latency. For autonomous vehicles, where safety critical decisions need to be made in split seconds, this presents an acute challenge.

Training bias

[20] https://huggingface.co/blog/llama31#inference-memory-requirements

The pre-training corpora might introduce social, cultural, political or gender biases into the model, which if left unchecked can skew service recommendations and perpetuate inequitable access. Especially in cases where LLMs are used for behavioural models, by giving the LLM a persona and then asking it to make decisions, a study (Campbell et al., 2025) found that not only do LLMs replicate the biases, but they often exaggerate them.

Hallucination and Stochasticity

Another major problem with LLMs is hallucination. LLMs often "hallucinate" facts, i.e., produce a plausible sounding but false or nonsensical fact, instead of accepting that the model does not know. This presents a significant challenge for safety critical applications such as traffic safety. Furthermore, even with the temperature parameters set to zero and providing the same prompt, LLMs often yield slightly different results (say traffic signal phase plans). This inconsistency or stochasticity makes real-world adoption challenging. Hallucination mitigation techniques involve grounding the responses in data through RAG, CoT prompting, and supervised fine-tuning.

No built-in physics or spatio-temporal understanding

This follows closely with the last point (hallucination) that the values predicted by LLMs are purely statistical and based on observed patterns. LLMs might output unrealistic values such negative travel times, queue length > 10 km, traffic volumes more than capacity, etc. Therefore, at this stage, it is crucial to check the values output by the LLM. Another important aspect for transport engineering is understanding and predicting spatial and temporal data. A recent study has shown that LLMs often struggle with spatial reasoning (H. Zhang et al., 2025) due to hallucinations.

Visual Shortcomings

MLLMs and VLMs, while demonstrating impressive visual scene understanding, often struggle with basic visual analyses such as counting the number of objects or identifying the orientation of entities (especially when there are a greater number of entities in an image), something which conventional Vision models like YOLO excelled in. They also suffer from hallucination issues and might omit an item which is present in the image or may add something into their interpretation which is absent from the image. This presents to be an acute problem for applications like autonomous driving.

Data Privacy

As LLMs often rely on sensitive user and operational data such as individual trip histories, real-time location traces, etc. they pose significant privacy concerns. For example, when LLMs are fine-tuned with proprietary datasets, there is a potential for unintended memorization of identifiable information, which might get exploited by malicious actors. Therefore, there is a pressing need to ensure that data privacy best practices are followed.

Ethical Considerations

The versatility of LLMs allows them to be applied to tasks such as decision-making and visual perception in transport systems. However, they are prone to errors, and this raises important ethical concerns. For instance, if a transport engineer follows a strategy suggested by an LLM and an accident occurs, accountability becomes unclear, as LLMs themselves cannot be held responsible. Moreover, LLMs lack inherent moral reasoning or virtues, a critical consideration when they are involved in enforcing rules, regulations, or safety-critical decisions.

6.4 Future Direction

So far, we have discussed current applications of LLMs in the domain of transport engineering. LLMs have already begun to reshape data-driven workflow. However, many promising avenues remain untapped. In this section, we outline prospective research directions that could provide better and more impactful LLM powered systems.

Meta-analysis

Future research could focus on comparing LLMs with traditional (statistical, machine learning and deep learning) methods on parameters like accuracy, compute cost and scalability. Such analysis would help make informed decisions about whether to use LLMs or not. Future studies can also perform quantitative and qualitative meta-analysis of the existing LLM based approaches as current literature doesn't provide enough information for a meta-analysis.

Another important points are that these methods often use different LLMs and techniques (Few-shot vs Fine-tuning), even when doing the same task. And while they are evaluated extensively against existing baselines, there is a lack of studies on how different LLMs compare for any given task or how different techniques perform. How important is the underlying LLM, vs the use of fine-tuning. And how big of an impact prompting plays. Future studies can investigate these aspects.

Improving LLMs

A lot of papers rely only on pre-trained LLMs for their abilities like reasoning and general-knowledge. Hence, improvements in LLMs will yield better results across the board for applications in the transport engineering domain. Furthermore, if adequate resources are available, smaller LLMs can be fine-tuned (either on domains specific data or undergo knowledge distillation) to reduce the computational complexity with little to no impact on the accuracy.

Hallucination Mitigation

Hallucination is one of the prominent issues with LLMs and can be caused due to a variety of factors such as poor training-data, improper context poor understanding of real-world dynamics. While some studies like (Ziwei Xu et al., 2024) suggest it cannot be completely eliminated it can be reduced by adopting techniques like Chain-of-though Prompting, domain specific fine-tuning, grounding results in trusted data sources using RAG and the introduction of guardrail systems.

Ethical Considerations

The versatility of LLMs allows them to be applied to tasks such as decision-making and visual perception in transport systems. However, they are prone to errors, and this raises important ethical concerns. For instance, if a transport engineer follows a strategy suggested by an LLM and an accident occurs, or the AV under LLM's control hits a pedestrian, accountability becomes unclear, as LLMs themselves cannot be held responsible. Furthermore, LLMs lack inherent moral reasoning or virtues, a critical consideration when they are involved in enforcing rules, regulations, or safety-critical decisions. Therefore, to properly integrate LLMs in Transport Engineering systems, ethical considerations need to be addressed with care.

Standardised Benchmarks

Given that LLMs are a new technology, there is a lack of benchmarks that quantifies the abilities of LLMs. A common trend seen in these papers is "decision-making" / "reasoning". However, as yet, there exists no metric to assess how good the decision-making or reasoning

capabilities of an LLM truly are. Next, as LLMs (specially VLMs and MLLMs) provide new capabilities (say visual scene understanding), there are no good benchmarks to assess the performance and give confidence to transport engineers to adopt these solutions for real-world applications.

Multilingual Machine Translation

Transit agencies often serve multilingual communities, making accurate translation essential. LLMs are exposed to vast amounts of multilingual text during training, and their architecture allows them to capture patterns of grammar, meaning, and context across languages. This enables them to perform well on translation tasks, often producing outputs on par with a junior human translator (Yan et al., 2024) in widely used languages, as shown in recent studies. Because they can be deployed with little or no fine-tuning, LLMs offer a practical, low-cost solution for transit agencies. However, their performance drops sharply for languages that are poorly represented (or entirely absent) in the training data, which limits their usefulness in such cases.

Agentic AI (LLM)

Agentic AI systems are those which leverages LLMs primarily for reasoning and communication, and operate in order to pursue autonomy, planning and adaptive decision making with minimal human intervention. Such systems work autonomously and proactively towards a goal, leveraging LLMs to understand requirements, break down complex commands into simple, executable steps, and then act upon them with the help of tools to achieve the goal. While LLMs are good at natural language processing, they are not so good at number crunching (the area where traditional ML models shine), hence Agentic AI can use these ML models as “tools” to empower LLM powered Agentic AI to build a better, synergised system. Many tasks in transport engineering can be formulated and solved with Agentic AI as demonstrated by some of the studies presented in (Da et al., 2024b; Devunuri and Lehe, 2025). MDAgents (Nori et al., 2023) is a novel agentic framework for the domain of medicine which copies how real medical teams make decisions sometimes solo, sometimes in groups, and outperforms older AI methods by flexibly adjusting to each problem’s difficulty and encouraging collaboration, leading to more accurate and robust medical answers

Drawing inspiration from other domains

In the Medical field, LLMs have helped in tasks like clinical documentation summarisation and generation, clinical information extraction, diagnostic and clinical decision support and patient communications and engagement (Jung, 2025). In Business and Management, LLMs help by supporting data-driven decision making, automating tasks such as customer support, marketing content generation and workflow optimisation, and enhancing knowledge management (Batz et al., 2025). In the domain of Education, LLMs have facilitated more interactive, collaborative and self-directed learning, promoting access to knowledge (Peláez-Sánchez et al., 2024).

Drawing inspiration from these fields, Transport engineers can leverage LLMs’ abilities like language processing, reasoning, etc. to perform or automate more and more tasks.

7 Conclusion

In this paper we systematically reviewed the emerging applications of Large Language Models across key transport engineering domains including public transport, traffic signal control, traffic forecasting, and traffic safety. LLMs exhibit versatile capabilities, such as advanced

reasoning, general-knowledge integration and high adaptability, which enables a wide range of applications from classification and prediction optimization, information mining and insight generation, question answering and knowledge source, decision support and optimisation, synthetic data generation and visual scene. In several cases, LLM based methods either establish a new state-of-the-art (e.g. human mobility prediction), or provide better explainability and interpretability than existing methods while maintaining comparable performance to the state-of-the-art approaches. The growing use of VLMs and MLLMs further demonstrates the potential for enhanced visual scene understanding, strengthening vision-driving transport and mobility solutions.

Despite this promise, several challenges limit real world deployment. These include high computational complexity, hallucinations risks, poor physical understanding of the real world. Addressing these limitations will require closer alignment between model development and transport domain needs.

Looking ahead, several research avenues must be pursued to fully unlock the potential of LLMs in transport engineering. Comparative meta-analyses are needed to benchmark performance against traditional methods and to clarify the roles of prompting, fine-tuning, and model selection. Further advances in domain-specific model design, cost-efficient adaptation, and integration with validated transport data will be essential to improve robustness and reduce hallucinations. Ethical and governance frameworks must be strengthened to ensure safe deployment in operational environments. Additionally, standardised benchmarks for reasoning, decision-making, and multimodal perception will support more reliable evaluation, while improved multilingual capabilities can enhance accessibility within diverse communities. Finally, agentic AI architectures and cross-domain knowledge transfer present promising directions for developing more autonomous, adaptive, and impactful transport solutions.

The continued integration of LLM advancements, holds significant potential to redefine how transport systems are analysed, managed and optimised.

8 References

Abdelrahman AS, Abdel-Aty M, Wang D. Video-to-Text Pedestrian Monitoring (VTPM): Leveraging Large Language Models for Privacy-Preserve Pedestrian Activity Monitoring at Intersections. Proceedings - 2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACVW 2025 2025:329–38. https://doi.org/10.1109/WACVW65960.2025.00043.

Abu Tami M, Ashqar HI, Elhenawy M, Glaser S, Rakotonirainy A. Using Multimodal Large Language Models (MLLMs) for Automated Detection of Traffic Safety-Critical Events. Vehicles 2024, Vol 6, Pages 1571-1590 2024;6:1571–90. https://doi.org/10.3390/VEHICLES6030074.

Ahmed A, Farhan M, Eesaar H, Chong KT, Tayara H. From Detection to Action: A Multimodal AI Framework for Traffic Incident Response. Drones 2024, Vol 8, Page 741 2024;8:741. https://doi.org/10.3390/DRONES8120741.

Aldieri L, Voß S. Bus Bunching and Bus Bridging: What Can We Learn from Generative AI Tools like ChatGPT? Sustainability 2023, Vol 15, Page 9625 2023;15:9625. https://doi.org/10.3390/SU15129625.

Arefeen MA, Debnath B, Chakradhar S. TrafficLens: Multi-Camera Traffic Video Analysis Using LLMs. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC 2024:3974–81. https://doi.org/10.1109/ITSC58415.2024.10920144.

Arteaga C, Park JW. A large language model framework to uncover underreporting in traffic crashes. J Safety Res 2025;92:1–13. https://doi.org/10.1016/J.JSR.2024.11.009.

Ashqar HI, Jaber A, Alhadidi TI, Elhenawy M, Hu ABJ. Advancing Object Detection in Transportation with Multimodal Large Language Models (MLLMs): A Comprehensive Review and Empirical Testing. Computation 2025, Vol 13, Page 133 2025;13:133. https://doi.org/10.3390/COMPUTATION13060133.

Batz A, D'Croz-Barón DF, Vega Pérez CJ, Ojeda-Sanchez CA. Integrating machine learning into business and management in the age of artificial intelligence. Humanit Soc Sci Commun 2025;12:1–20. https://doi.org/10.1057/S41599-025-04361-6;SUBJMETA=4000,4001,4008;KWRD=BUSINESS+AND+MANAGEMENT,INFORMATION+SYSTEMS+AND+INFORMATION+TECHNOLOGY.

Beneduce C, Lepri B, Luca M. Large Language Models are Zero-Shot Next Location Predictors. IEEE Access 2025;13:77456–67. https://doi.org/10.1109/ACCESS.2025.3565297.

Boqing Zhu C, Li D, Zhu contributed equally Authors B, Yi M, He Y, Zhu B, et al. Multi-task Pre-training Language Model for Semantic Network Completion. ACM Transactions on Asian and Low-Resource Language Information Processing 2023;22. https://doi.org/10.1145/3627704.

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners. Adv Neural Inf Process Syst 2020;33:1877–901.

Calenzani JFG, Neves VN, Ramos LT, Junior LJL, Magnago LCS, Badue C, et al. A Study on the Effectiveness of GPT-4V in Classifying Driver Behavior Captured on Video Using Just a Few Frames per Video. Proceedings of the International Joint Conference on Neural Networks 2024. https://doi.org/10.1109/IJCNN60899.2024.10650751.

Campbell H, Goldman S, Markey PM. Artificial intelligence and human decision making: Exploring similarities in cognitive bias. Computers in Human Behavior: Artificial Humans 2025;4:100138. https://doi.org/10.1016/J.CHBAH.2025.100138.

Chen Y, Chi B, Li C, Zhang Y, Liao C, Chen X, et al. Toward Interactive Next Location Prediction Driven by Large Language Models. IEEE Trans Comput Soc Syst 2025. https://doi.org/10.1109/TCSS.2024.3522965.

Chib PS, Singh P. LG-Traj: LLM Guided Pedestrian Trajectory Prediction 2024.

Choi S, Lim Y. Optimizing Traffic Signal Control Using LLM-Driven Reward Weight Adjustment in Reinforcement Learning. Journal of Information Processing Systems 2025;21:43–51. https://doi.org/10.3745/JIPS.04.0334.

Choudhary T, Dewangan V, Chandhok S, Priyadarshan S, Jain A, Singh AK, et al. Talk2BEV: Language-enhanced Bird's-eye View Maps for Autonomous Driving. Proc IEEE Int Conf Robot Autom 2024:16345–52. https://doi.org/10.1109/ICRA57147.2024.10611485.

Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: Scaling Language Modeling with Pathways. Journal of Machine Learning Research 2022;24.

Corrias R, Gjoreski M, Langheinrich M. Exploring Transformer and Graph Convolutional Networks for Human Mobility Modeling. Sensors 2023, Vol 23, Page 4803 2023;23:4803. https://doi.org/10.3390/S23104803.

Costa DG, Silva I, Medeiros M, Bittencourt JCN, Andrade M. A method to promote safe cycling powered by large language models and AI agents. MethodsX 2024;13:102880. https://doi.org/10.1016/J.MEX.2024.102880.

Cui C, Ma Y, Cao X, Ye W, Wang Z. Drive as You Speak: Enabling Human-Like Interaction with Large Language Models in Autonomous Vehicles. 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW) 2024a:902–9. https://doi.org/10.1109/WACVW60836.2024.00101.

Cui C, Ma Y, Cao X, Ye W, Wang Z. Receive, Reason, and React: Drive as You Say, With Large Language Models in Autonomous Vehicles. IEEE Intelligent Transportation Systems Magazine 2024b;16:81–94. https://doi.org/10.1109/MITS.2024.3381793.

Cui C, Ma Y, Cao X, Ye W, Zhou Y, Liang K, et al. A Survey on Multimodal Large Language Models for Autonomous Driving 2024c:958–79.

Da L, Gao M, Mei H, Wei H. Prompt to Transfer: Sim-to-Real Transfer for Traffic Signal Control with Prompt Learning. Proceedings of the AAAI Conference on Artificial Intelligence 2024a;38:82–90. https://doi.org/10.1609/AAAI.V38I1.27758.

Da L, Liou K, Chen T, Zhou X, Luo X, Yang Y, et al. Open-ti: open traffic intelligence with augmented language model. International Journal of Machine Learning and Cybernetics 2024b;15:4761–86. https://doi.org/10.1007/S13042-024-02190-8/FIGURES/21.

Deng Y, Yao J, Tu Z, Zheng X, Zhang M, Zhang T. TARGET: Automated Scenario Generation from Traffic Rules for Testing Autonomous Vehicles via Validated LLM-Guided Knowledge Extraction 2023.

Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient Finetuning of Quantized LLMs. Adv Neural Inf Process Syst 2023;36.

Devlin J, Chang M-W, Lee K, Google KT, Language AI. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North 2019:4171–86. https://doi.org/10.18653/V1/N19-1423.

Devunuri S, Lehe L. TransitGPT: a generative AI-based framework for interacting with GTFS data using large language models. Public Transport 2025:1–27. https://doi.org/10.1007/S12469-025-00395-W/TABLES/3.

Devunuri S, Qiam S, Lehe LJ. ChatGPT for GTFS: benchmarking LLMs on GTFS semantics.. and retrieval. Public Transport 2024;16:333–57. https://doi.org/10.1007/S12469-024-00354-X/FIGURES/8.

Ding H, Du Y, Xia Z. Urban Road Anomaly Monitoring Using Vision–Language Models for Enhanced Safety Management. Applied Sciences 2025, Vol 15, Page 2517 2025;15:2517. https://doi.org/10.3390/APP15052517.

Feng S, Lyu H, Li F, Sun Z, Chen C. Where to Move Next: Zero-shot Generalization of LLMs for Next POI Recommendation. Proceedings - 2024 IEEE Conference on Artificial Intelligence, CAI 2024 2024:1530–5. https://doi.org/10.1109/CAI59869.2024.00277.

Gao C. BRD-LLAMA based Bike Rental Demand Prediction. Proceedings of 2024 4th International Conference on Signal Processing and Communication Technology, SPCT 2024 2025:176–80. https://doi.org/10.1145/3712464.3712497.

Gong L, Lin Yan, Zhang X, Lu Y, Han X, Liu Y, et al. Mobility-LLM: Learning Visiting Intentions and Travel Preference from Human Mobility Data with Large Language Models. Adv Neural Inf Process Syst 2024;37:36185–217.

Grigorev A, Saleh K, Ou Y, Mihăiţă AS. Enhancing Traffic Incident Management with Large Language Models: A Hybrid Machine Learning Approach for Severity Classification. International Journal of Intelligent Transportation Systems Research 2025;23:259–80. https://doi.org/10.1007/S13177-024-00448-7/FIGURES/14.

Guo X, Zhang Q, Jiang J, Peng M, Hao, Yang, et al. Towards explainable traffic flow prediction with large language models. Communications in Transportation Research 2024;4:100150. https://doi.org/10.1016/J.COMMTR.2024.100150.

Han Q, Yoshikawa A, Yamamura M. Adapting Large Language Model for Spatio-Temporal Understanding in Next Point-of-Interest Prediction. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 2025. https://doi.org/10.1109/ICASSP49660.2025.10889866.

Hochreiter S. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. Https://DoiOrg/101142/S0218488598000094 2011;6:107–16. https://doi.org/10.1142/S0218488598000094.

Hu E, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022 - 10th International Conference on Learning Representations 2021.

Hu Y, Wang F, Ye D, Wu M, Kang J, Yu R. LLM-Based Misbehavior Detection Architecture for Enhanced Traffic Safety in Connected Autonomous Vehicles. IEEE Trans Veh Technol 2025a. https://doi.org/10.1109/TVT.2025.3551327.

Hu Y, Zheng S, Zhang Z, Wang S, Ye D, Wu M, et al. Leveraging LLMs in Cloud-Edge Networks for Traffic Risk Prediction and Accident Severity Analysis. IEEE Trans Netw Sci Eng 2025b. https://doi.org/10.1109/TNSE.2025.3584432.

Hwang J-J, Xu R, Lin H, Hung W-C, Ji J, Choi K, et al. EMMA: End-to-End Multimodal Model for Autonomous Driving 2024.

Jaradat S, Elhenawy M, Ashqar HI, Paz A, Nayak R. Leveraging Deep Learning and Multimodal Large Language Models for Near-Miss Detection Using Crowdsourced Videos. IEEE Open Journal of the Computer Society 2025a;6:223–35. https://doi.org/10.1109/OJCS.2025.3525560.

Jaradat S, Elhenawy M, Nayak R, Paz A, Ashqar HI, Glaser S. Multimodal Data Fusion for Tabular and Textual Data: Zero-Shot, Few-Shot, and Fine-Tuning of Generative Pre-Trained Transformer Models. AI 2025, Vol 6, Page 72 2025b;6:72. https://doi.org/10.3390/AI6040072.

Jaradat S, Nayak R, Paz A, Ashqar HI, Elhenawy M. Multitask Learning for Crash Analysis: A Fine-Tuned LLM Framework Using Twitter Data. Smart Cities 2024, Vol 7, Pages 2422-2465 2024;7:2422–65. https://doi.org/10.3390/SMARTCITIES7050095.

Jiang R, Wang S, Ma W, Zhang Y, Fan P, Jia D. A knowledge-informed dynamic correlation modeling framework for lane-level traffic flow prediction. Information Fusion 2025;124:103327. https://doi.org/10.1016/J.INFFUS.2025.103327.

Ju C, Liu J, Sinha S, Xue H, Salim F. TrajLLM: A Modular LLM-Enhanced Agent-Based Framework for Realistic Human Trajectory Simulation. WWW Companion 2025 - Companion Proceedings of the ACM Web Conference 2025 2025a:2847–50. https://doi.org/10.1145/3701716.3715201.

Ju C, Liu J, Sinha S, Xue H, Salim F. TrajLLM: A Modular LLM-Enhanced Agent-Based Framework for Realistic Human Trajectory Simulation. WWW Companion 2025 - Companion Proceedings of the ACM Web Conference 2025 2025b:2847–50. https://doi.org/10.1145/3701716.3715201.

Jung KH. Large Language Models in Medicine: Clinical Applications, Technical Challenges, and Ethical Considerations. Healthc Inform Res 2025;31:114. https://doi.org/10.4258/HIR.2025.31.2.114.

Kingma DP, Ba JL. Adam: A Method for Stochastic Optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings 2014.

Lee J ;, Cha K-A;, Lee M, Lee Jekyung, Cha K-A, Lee Miran. Multi-Modal System for Walking Safety for the Visually Impaired: Multi-Object Detection and Natural Language Generation. Applied Sciences 2024, Vol 14, Page 7643 2024;14:7643. https://doi.org/10.3390/APP14177643.

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Adv Neural Inf Process Syst 2020;33:9459–74.

Li J, Li Jingyuan, Yang G, Yang Lie, Chi H, Yang Lichao. Applications of Large Language Models and Multimodal Large Models in Autonomous Driving: A Comprehensive Review. Drones 2025, Vol 9, Page 238 2025;9:238. https://doi.org/10.3390/DRONES9040238.

Li W, Ding L, Zhang Y, Pu Z. Understanding multimodal travel patterns based on semantic embeddings of human mobility trajectories. J Transp Geogr 2025;124:104169. https://doi.org/10.1016/J.JTRANGEO.2025.104169.

Li Y, Katsumata K, Javanmardi E, Tsukada M. Large Language Models for Human-Like Autonomous Driving: A Survey. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC 2024:439–46. https://doi.org/10.1109/ITSC58415.2024.10919629.

Liang Y, Liu Y, Wang X, Zhao Z. Exploring large language models for human mobility prediction under public events. Comput Environ Urban Syst 2024;112:102153. https://doi.org/10.1016/J.COMPENVURBSYS.2024.102153.

Liao C, Lin K. DDC-Chat: Achieving accurate distracted driver classification through instruction tuning of visual language model. Journal of Safety Science and Resilience 2025;6:250–64. https://doi.org/10.1016/J.JNLSSR.2024.10.001.

Lis K, Nakka K, Fua P, Salzmann M. Detecting the Unexpected via Image Resynthesis 2019:2152–61.

Liu C, Hettige KH, Xu Q, Long C, Xiang S, Cong G, et al. ST-LLM+: Graph Enhanced Spatio-Temporal Large Language Models for Traffic Prediction. IEEE Trans Knowl Data Eng 2025. https://doi.org/10.1109/TKDE.2025.3570705.

Liu J, Hang P, Qi X, Wang J, Sun J. MTD-GPT: A Multi-Task Decision-Making GPT Model for Autonomous Driving at Unsignalized Intersections. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC 2023:5154–61. https://doi.org/10.1109/ITSC57777.2023.10421993.

Liu L, Pei T, Fang Z, Yan X, Zheng C, Wang X, et al. Extracting individual trajectories from text by fusing large language models with diverse knowledge. International Journal of Applied Earth Observation and Geoinformation 2025;141:104654. https://doi.org/10.1016/J.JAG.2025.104654.

Liu Q, Yu R, Cai Y, Yuan Q, Wei H, Lv C. Collision risk prediction and takeover requirements assessment based on radar-video integrated sensors data: A system framework based on LLM. Accid Anal Prev 2025;218:108041. https://doi.org/10.1016/J.AAP.2025.108041.

Lu Q, Wang X, Jiang Y, Zhao G, Ma M, Feng S. Multimodal Large Language Model Driven Scenario Testing for Autonomous Vehicles. Automotive Innovation 2025 2024:1–15. https://doi.org/10.1007/S42154-025-00364-W/FIGURES/12.

Luo Y, Cao Z, Jin X, Liu K, Yin L. Deciphering Human Mobility: Inferring Semantics of Trajectories with Large Language Models. Proceedings - IEEE International Conference on Mobile Data Management 2024:289–94. https://doi.org/10.1109/MDM61037.2024.00060.

Ma Y, Cao Y, Sun J, Pavone M, Xiao C. Dolphins: Multimodal Language Model for Driving. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 2025;15103 LNCS:403–20. https://doi.org/10.1007/978-3-031-72995-9_23.

Mao J, Qian Y, Ye J, Zhao H, Wang Y. GPT-Driver: Learning to Drive with GPT 2023.

Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings 2013.

Movahedi M, Choi J. The Crossroads of LLM and Traffic Control: A Study on Large Language Models in Adaptive Traffic Signal Control. IEEE Transactions on Intelligent Transportation Systems 2025;26:1701–16. https://doi.org/10.1109/TITS.2024.3498735.

Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making 2023.

Nouri A, Cabrero-Daniel B, Torner F, Sivencrona H, Berger C. Engineering Safety Requirements for Autonomous Driving with Large Language Models. Proceedings of the IEEE International Conference on Requirements Engineering 2024:218–28. https://doi.org/10.1109/RE59067.2024.00029.

Openai AR, Openai KN, Openai TS, Openai IS. Improving Language Understanding by Generative Pre-Training 2018.

Padoan L, Cesetti M, Brunello L, Antonelli M, Zamengo B, Silvestri F. Mobility ChatBot: supporting decision making in mobility data with chatbots. Proceedings - IEEE International Conference on Mobile Data Management 2024:295–300. https://doi.org/10.1109/MDM61037.2024.00061.

Peláez-Sánchez IC, Velarde-Camaqui D, Glasserman-Morales LD. The impact of large language models on higher education: exploring the connection between AI and Education 4.0. Front Educ (Lausanne) 2024;9:1392091. https://doi.org/10.3389/FEDUC.2024.1392091/XML.

Peng M, Guo X, Chen X, Chen K, Zhu M, Chen L, et al. LC-LLM: Explainable lane-change intention and trajectory predictions with Large Language Models. Communications in Transportation Research 2025;5:100170. https://doi.org/10.1016/J.COMMTR.2025.100170.

Pérez B, Resino M, Seco T, García F, Al-Kaff A. Innovative Approaches to Traffic Anomaly Detection and Classification Using AI. Applied Sciences 2025, Vol 15, Page 5520 2025;15:5520. https://doi.org/10.3390/APP15105520.

Peruski R, Saroj A, Zhou W, Djouadi S, Cao C. Edge AI-Enhanced Traffic Monitoring and Anomaly Detection Using Multimodal Large Language Models. International Conference on Transportation and Development 2025: Transportation Safety and Emerging Technologies - Selected Papers from the International Conference on Transportation and Development 2025 2025:429–38. https://doi.org/10.1061/9780784486191.038.

Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 2018;1:2227–37. https://doi.org/10.18653/v1/n18-1202.

Pinggera P, Ramos S, Gehrig S, Franke U, Rother C, Mester R. Lost and found: Detecting small road hazards for self-driving vehicles. IEEE International Conference on Intelligent Robots and Systems 2016;2016-November:1099–106. https://doi.org/10.1109/IROS.2016.7759186.

Qin Z, Zhang P, Wang L, Ma Z. LingoTrip: Spatiotemporal context prompt driven large language model for individual trip prediction. J Public Trans 2025;27:100117. https://doi.org/10.1016/J.JPUBTR.2025.100117.

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning Transferable Visual Models From Natural Language Supervision. Proc Mach Learn Res 2021;139:8748–63.

Renz K, Chen L, Marcu A-M, Hünermann J, Hanotte B, Karnsund A, et al. CarLLaVA: Vision language models for camera-only closed-loop driving 2024.

Sathyam R, Li Y. Foundation Models for Autonomous Driving Perception: A Survey Through Core Capabilities. IEEE Open Journal of Vehicular Technology 2025. https://doi.org/10.1109/OJVT.2025.3604823.

Schlögl M, Stütz R, Laaha G, Melcher M. A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset. Accid Anal Prev 2019;127:134–49. https://doi.org/10.1016/J.AAP.2019.02.008.

Schumann R, Zhu W, Feng W, Fu TJ, Riezler S, Wang WY. VELMA: Verbalization Embodiment of LLM Agents for Vision and Language Navigation in Street View. Proceedings of the AAAI Conference on Artificial Intelligence 2024;38:18924–33. https://doi.org/10.1609/AAAI.V38I17.29858.

Shafiq S, Awan HM, Khan AA, Amin W. Driving Like Humans: Leveraging Vision Large Language Models for Road Anomaly Detection. 2024 3rd International Conference on

Emerging Trends in Electrical, Control, and Telecommunication Engineering, ETECTE 2024 - Proceedings 2024. https://doi.org/10.1109/ETECTE63967.2024.10823889.

Shao H, Hu Y, Wang L, Song G, Waslander SL, Liu Y, et al. LMDrive: Closed-Loop End-to-End Driving with Large Language Models n.d.

Sharan SP, Pittaluga F, G VKB, Chandraker M. LLM-Assist: Enhancing Closed-Loop Planning with Language-Based Reasoning 2023.

Sun Y, Shi Y, Jia K, Zhang Z, Qin L. A Dual-Stream Cross AGFormer-GPT Network for Traffic Flow Prediction Based on Large-Scale Road Sensor Data. Sensors 2024, Vol 24, Page 3905 2024;24:3905. https://doi.org/10.3390/S24123905.

Tami MA, Ashqar HI, Elhenawy M, Glaser S, Rakotonirainy A. Using Multimodal Large Language Models (MLLMs) for Automated Detection of Traffic Safety-Critical Events. Vehicles 2024, Vol 6, Pages 1571-1590 2024;6:1571–90. https://doi.org/10.3390/VEHICLES6030074.

Tan S, Ivanovic B, Weng X, Pavone M, Krähenbühl P. Language Conditioned Traffic Generation. Proc Mach Learn Res 2023;229.

Tanahashi K, Inoue Yuichi, Yamaguchi Y, Yaginuma H, Shiotsuka D, Shimatani H, et al. Evaluation of Large Language Models for Decision Making in Autonomous Driving 2023.

Tang P, Yang C, Xing T, Xu X, Jiang R, Sezaki K. Instruction-Tuning Llama-3-8B Excels in City-Scale Mobility Prediction. 2nd ACM SIGSPATIAL International Workshop on the Human Mobility Prediction Challenge, HuMob-Challenge 2024 2024:1–4. https://doi.org/10.1145/3681771.3699908.

Tang T, Wei D, Jia Z, Gao T, Cai C, Hou C, et al. BEV-TSR: Text-Scene Retrieval in BEV Space for Autonomous Driving. Proceedings of the AAAI Conference on Artificial Intelligence 2025;39:7275–83. https://doi.org/10.1609/AAAI.V39I7.32782.

Tang Yun, Da Costa AAB, Zhang X, Patrick I, Khastgir S, Jennings P. Domain Knowledge Distillation from Large Language Model: An Empirical Study in the Autonomous Driving Domain. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC 2023:3893–900. https://doi.org/10.1109/ITSC57777.2023.10422308.

Tang Y, Dai X, Lv Y. Large Language Model-Assisted Arterial Traffic Signal Control. IEEE Journal of Radio Frequency Identification 2024a;8:322–6. https://doi.org/10.1109/JRFID.2024.3384289.

Tang Yiqing, Dai X, Lv Y. ChatGPT Participates in Traffic Control as a Traffic Manager Assistant. 2023 IEEE 3rd International Conference on Digital Twins and Parallel Intelligence, DTPI 2023 2023. https://doi.org/10.1109/DTPI59677.2023.10365318.

Tang Y, Dai X, Zhao C, Cheng Q, Lv Y. Large Language Model-Driven Urban Traffic Signal Control. 2024 Australian and New Zealand Control Conference, ANZCC 2024 2024b:67–71. https://doi.org/10.1109/ANZCC59813.2024.10432823.

Tian X, Gu J, Li B, Liu Y, Wang Y, Zhao Z, et al. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. Proc Mach Learn Res 2024;270:4698–726.

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, et al. LLaMA: Open and Efficient Foundation Language Models 2023.

Trinh Xuan K, Nguyen Nguyen K, Hoang Ngo B, Dinh Xuan V, An M-H, Dinh Q-V, et al. Divide and Conquer Boosting for Enhanced Traffic Safety Description and Analysis with Large Vision Language Model 2024:7046–55.

Ulan M, Söderman M. Making sense of data: leveraging AI to empower data-driven decision making in public transport. Transportation Research Procedia 2025;86:700–7. https://doi.org/10.1016/J.TRPRO.2025.04.087.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. Adv Neural Inf Process Syst 2017;2017-December:5999–6009.

Venkatesh Raja K, Siddharth R, Yuvaraj S, Ramesh Kumar KA. An Artificial Intelligence based automated case-based reasoning (CBR) system for severity investigation and root-cause analysis of road accidents – Comparative analysis with the predictions of ChatGPT. Journal of Engineering Research 2024;12:895–903. https://doi.org/10.1016/J.JER.2023.09.019.

Wang P, Wei X, Hu F, Han W. TransGPT: Multi-modal Generative Pre-trained Transformer for Transportation. Proceedings - 2024 International Conference on Computational Linguistics and Natural Language Processing, CLNLP 2024 2024:96–100. https://doi.org/10.1109/CLNLP64123.2024.00026.

Wang S, Sheng Z, Xu J, Chen T, Zhu J, Zhang S, et al. ADEPT: A Testing Platform for Simulated Autonomous Driving. ACM International Conference Proceeding Series 2022. https://doi.org/10.1145/3551349.3559528.

Wang S, Yu Z, Jiang X, Lan S, Shi M, Chang N, et al. OmniDrive: A Holistic Vision-Language Dataset for Autonomous Driving with Counterfactual Reasoning 2025:22442–52.

Wang Z, Zheng X, Meng F, Wang K, Wu X, Yu D. Exploring the Joint Influence of Built Environment Factors on Urban Rail Transit Peak-Hour Ridership Using DeepSeek. Buildings 2025, Vol 15, Page 1744 2025;15:1744. https://doi.org/10.3390/BUILDINGS15101744.

Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Adv Neural Inf Process Syst 2022;35:24824–37.

Wen L, Fu D, Li X, Cai X, Ma T, Cai P, et al. DiLu: A Knowledge-Driven Approach to Autonomous Driving with Large Language Models. 12th International Conference on Learning Representations, ICLR 2024 2023a.

Wen L, Yang X, Fu D, Wang X, Cai P, Li X, et al. On the Road with GPT-4V(ision): Early Explorations of Visual-Language Model on Autonomous Driving 2023b.

de Winter J, Hoogmoed J, Stapel J, Dodou D, Bazilinskyy P. Predicting perceived risk of traffic scenes using computer vision. Transp Res Part F Traffic Psychol Behav 2023;93:235–47. https://doi.org/10.1016/J.TRF.2023.01.014.

Xu Z, Chen T, Chen S. A LLM-based Multimodal Warning System for Driver Assistance. IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC 2024:1527–32. https://doi.org/10.1109/ITSC58415.2024.10919879.

Xu Ziwei, Jain S, Kankanhalli M. Hallucination is Inevitable: An Innate Limitation of Large Language Models 2024.

Xu Z, Zhang Y, Xie E, Zhao Z, Guo Y, Wong KYK, et al. DriveGPT4: Interpretable End-to-End Autonomous Driving Via Large Language Model. IEEE Robot Autom Lett 2024;9:8186–93. https://doi.org/10.1109/LRA.2024.3440097.

Yan J, Yan P, Chen Y, Li J, Zhu X, Zhang Y. GPT-4 vs. Human Translators: A Comprehensive Evaluation of Translation Quality Across Languages, Domains, and Expertise Levels 2024.

Yan Y, Liao Y, Xu G, Yao R, Fan H, Sun J, et al. Large Language Models for Traffic and Transportation Research: Methodologies, State of the Art, and Future Opportunities 2025.

Yang D, Fankhauser B, Rosso P, Cudre-Mauroux P. Location Prediction over Sparse User Mobility Traces Using RNNs. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence 2020;2021-January:2184–90. https://doi.org/10.24963/IJCAI.2020/302.

Yang S, Liu J, Zhang R, Pan M, Guo Z, Li X, et al. LiDAR-LLM: Exploring the Potential of Large Language Models for 3D LiDAR Understanding. Proceedings of the AAAI Conference on Artificial Intelligence 2025;39:9247–55. https://doi.org/10.1609/AAAI.V39I9.33001.

Yao J, Li J, Xu X, Tan C, Yap KH, Su R. Incorporating vision-based artificial intelligence and large language model for smart traffic light control. Appl Soft Comput 2025;179:113333. https://doi.org/10.1016/J.ASOC.2025.113333.

Yu C, Xie X, Huang Y, Qiu C. Harnessing LLMs for Cross-City OD Flow Prediction. 32nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM SIGSPATIAL 2024 2024;12:384–95. https://doi.org/10.1145/3678717.3691308.

Yulianda Y, Misbullah A, Farsiah L, Husaini, Sukiakhy KM, Junidar. Traffic Sign Recognition Using Deep Learning with Interactive Voice Output for Drivers. 2024 9th International Conference on Informatics and Computing, ICIC 2024 2024. https://doi.org/10.1109/ICIC64337.2024.10957387.

Bin Zaman Chowdhury MT, Islam MR, Hossain M. Durghotona GPT: A Web Scraping and Large Language Model Based Framework to Generate Road Accident Dataset Automatically in Bangladesh. 2024 27th International Conference on Computer and Information Technology, ICCIT 2024 - Proceedings 2024:50–5. https://doi.org/10.1109/ICCIT64611.2024.11021969.

Zhang H, Deng H, Ou J, Feng C. Mitigating spatial hallucination in large language models for path planning via prompt engineering. Sci Rep 2025;15:1–13. https://doi.org/10.1038/S41598-025-93601-5;SUBJMETA=117,639,705,794;KWRD=COMPUTER+SCIENCE,SOFTWARE.

Zhang P, Zeng G, Wang T, Lu W. TinyLlama: An Open-Source Small Language Model 2024.

Zhang R, Wang B, Zhang J, Bian Z, Feng C, Ozbay K. When language and vision meet road safety: Leveraging multimodal large language models for video-based traffic accident analysis. Accid Anal Prev 2025;219:108077. https://doi.org/10.1016/J.AAP.2025.108077.

Zhang S, Fu D, Liang W, Zhang Z, Yu B, Cai P, et al. TrafficGPT: Viewing, processing and interacting with traffic foundation models. Transp Policy (Oxf) 2024;150:95–105. https://doi.org/10.1016/J.TRANPOL.2024.03.006.

Zhang Y, Zhang K, Pang Y, Sekimoto Y. Agentic Large Language Models for Generating Large-Scale Urban Daily Activity Patterns. Proceedings - 2024 IEEE International Conference on Big Data, BigData 2024 2024:6815–22. https://doi.org/10.1109/BIGDATA62323.2024.10825138.

Zhang Zheng, Amiri H, Liu Z, Zhao L, Züfle A. Large Language Models for Spatial Trajectory Patterns Mining. GEOANOMALIES 2024 - Proceedings of the 1st ACM SIGSPATIAL International Workshop on Geospatial Anomaly Detection 2024:52–5. https://doi.org/10.1145/3681765.3698467.

Zhang Zhendong, Shen Z, Yuan M, Zhu F, Ali H, Xiong G. RAGTraffic: Utilizing Retrieval-Augmented Generation for Intelligent Traffic Signal Control. 2024 International Annual Conference on Complex Systems and Intelligent Science, CSIS-IAC 2024 2024:728–35. https://doi.org/10.1109/CSIS-IAC63491.2024.10919289.

Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A Survey of Large Language Models 2025.

Zhen H, Shi Y, Huang Y, Yang JJ, Liu N. Leveraging Large Language Models with Chain-of-Thought and Prompt Engineering for Traffic Crash Severity Analysis and Inference. Computers 2024, Vol 13, Page 232 2024;13:232. https://doi.org/10.3390/COMPUTERS13090232.

Zhen H, Yang JJ. Tab-Text: Bridging tabular data and natural language for enhanced traffic safety analysis and modeling. Expert Syst Appl 2025;290:128450. https://doi.org/10.1016/J.ESWA.2025.128450.

Zheng X, Wu L, Yan Z, Tang Y, Zhao H, Zhong C, et al. Large Language Models Powered Context-aware Motion Prediction in Autonomous Driving. IEEE International Conference on Intelligent Robots and Systems 2024:980–5. https://doi.org/10.1109/IROS58592.2024.10802397.

Zhong X, Xiang Y, Yi F, Li C, Yang Q. HMP-LLM: Human Mobility Prediction Based on Pre-trained Large Language Models. Proceedings - 2024 IEEE 4th International Conference on Digital Twins and Parallel Intelligence, DTPI 2024 2024:687–92. https://doi.org/10.1109/DTPI61353.2024.10778764.

Zhou X, Liu M, Yurtsever E, Zagar BL, Zimmer W, Cao H, et al. Vision Language Models in Autonomous Driving: A Survey and Outlook. IEEE Transactions on Intelligent Vehicles 2024. https://doi.org/10.1109/TIV.2024.3402136.

Zhu Y, Li H, Liao Y, Wang B, Guan Z, Liu H, et al. What to do next: Modeling user behaviors by Time-LSTM. IJCAI International Joint Conference on Artificial Intelligence 2017;0:3602–8. https://doi.org/10.24963/IJCAI.2017/504.

Zhu Y, Wang Shiyi, Zhong W, Shen N, Li Y, Wang Siqi, et al. A Survey on Large Language Model-Powered Autonomous Driving. Engineering 2025. https://doi.org/10.1016/J.ENG.2025.07.038.

Appendix A

Public Transport

Paper	LLM	Task	Focus	Result	Limitations
(Aldieri and Voß, 2023)	GPT-3.5, GPT-4	Question Answering & Knowledge Retrieval	Exploring how generative AI chatbots (ChatGPT and Bing) can be applied to analyse and discuss public-transport disturbance scenarios—specifically bus bunching and bus bridging—within a sustainability context.	LLMs are able to answer questions related to bus bunching and bus bridging, demonstrating that they have a basic understanding of the phenomena. However, they may produce inaccurate responses sometimes	The paper is not solving any problem, nor have the authors mentioned how their findings will lead to a solution. Lack of quantitative metrics or discussion on the qualitative judging criteria. Specific versions of LLMs are not clearly mentioned.
(Devunuri et al., 2024)	GPT-3.5-Turbo, GPT-4	Question Answering & Knowledge Retrieval	A study to examine and benchmark the understanding of LLMs on the GTFS file format and feeds on two benchmarks: GTFS semantics and GTFS retrieval	Creation of two public benchmarks and demonstration that GPT-4 (via program synthesis) can achieve up to 93 % accuracy on simple retrieval tasks.	The paper is not solving any problems.
(Devunuri and Lehe, 2025)	GPT-4o, Claude 3.5 Sonnet	Question Answering & Knowledge Retrieval	**TransitGPT**: Using LLMs to generate code to answer user queries by fetching GTFS Static database	Off-the-shelf LLMs are capable of generating code to answer user queries from GTFS static dataset.	Doesn't work with GTFS real-time
(Ulan and Söderman, 2025)	LLaMA-2 (Fine-tuned)	Information Mining and Insight Generation	Presenting a prototype AI-based data management tool that leverages LLMs and RAG to help public-transport authorities query,	A web-based prototype tool that demonstrated significant usability improvements between two evaluation iterations for data-driven decision support.	Not good with Swedish data. Evaluation might be biased

			summarize, and chat with their data.		
(Z. Wang et al., 2025)	DeepSeek-R1	Information Mining and Insight Generation	Introducing an LLM-based analytical framework using DeepSeek-R1 to explore the joint influence of built environment and station characteristics on urban rail transit peak-hour ridership (Morning & Evening boarding & alighting)	The LLM with built environment factors achieved 12.8% MAPE, 0.89 R^2, and 87.5% A20—outperforming XGBoost (16.0% MAPE, 0.86 R^2) and ARIMA (20.1% MAPE, 0.82 R^2). Proximity to the CBD and population density proved to be the two most important factors overall.	Reliance on Euclidean PCA delineation, simplified POI data, outdated station coverage, and potential LLM hallucinations.

Traffic Signal Control

Paper	**LLM**	**Task**	**Focus**	**Result**	**Limitations**
(Yiqing Tang et al., 2023)	GPT (version not mentioned)	Decision Support and Prescriptive Optimisation	The goal of this paper is to investigate how the capabilities of ChatGPT can be extended to the field of traffic control management, helping traffic manager to improve the efficiency of traffic management and enhance the intelligence of traffic information control.	ChatGPT can support traffic managers by quickly providing domain knowledge, analysing road network topologies, and extracting insights from traffic flow data. It can also suggest optimization strategies, helping improve efficiency in traffic control policy deployment.	Lack of concrete metrics and comparative study. The research gap is not mentioned, neither is how the outcome can be used to actually improve TSC
(Da et al., 2024b)	Llama2-7b, Llama2-13b, ChatGPT−3.5 and GPT−4.0.	Decision Support and Prescriptive Optimisation	**Open-TI**, an augmented language model for comprehensive traffic intelligence.	Open-TI, an LLM agent capable of question answer, processing raw map data, executing simulations, training traffic light policies and conduct demand optimisation.	Aligning intermediate outputs across complex tool sequences is difficult, may result in failure

(Da et al., 2024a)	GPT-4.0	Decision Support and Prescriptive Optimisation	**PromptGAT**: When policies trained on simulators are implemented in real world, a performance gap arises due to the differences in the dynamics of real-world and simulator. PromptGAT leverages the knowledge of LLMs and prompts them to generate real world like system dynamics to train better policies to reduce the performance gap	The use of PromptGAT reduces the simulation to real-world gap the most when compared to baselines like direct-transfer and Vanilla-GAT.	The dynamics given by the LLM could be wrong due to bias or hallucinations. The authors didn't discuss the accuracy of the metrics given by LLMs. Also, the authors didn't actually check in real-world.
(Zhendong Zhang et al., 2024)	*Not mentioned!!*	Decision Support and Prescriptive Optimisation	**RAGTraffic**, a framework for TSC that integrates historical data with real-time information for better TSC	RAGTraffic outperforms various transportation, RL-based and ML based baselines	High dependence on the quality of knowledge base and retrieval mechanism. Not compared against sensor-to-action baselines.
(Y. Tang et al., 2024a)	GPT-4	Decision Support and Prescriptive Optimisation	Leveraging LLMs to generate Green Wave policies for arterial roads with the help of traffic signal control policies	The Average Road Speed increase in most cases after the deployment of LLM policy	Not compared against any baseline models or human, hence, the actual effectiveness is not verified.
(Y. Tang et al., 2024b)	*Not mentioned!!*	Decision Support and Prescriptive Optimisation	Proposes LLM driven urban TSC methods. Three modes are proposed autonomous, feedback and human takeover mode based on the artificial systems, computational	3 TSC modes are described Autonomous, Human Feedback (HF), and Human Takeover (HT). Autonomous model relies on traffic scenarios and strategy library; HF and HT modes are useful to ensure reliability and handle edge cases.	None of the proposed methods are tested or compared with baselines, so while the method is explained, the effectiveness and the fine details are not discussed

			experiments, parallel execution (ACP) method.		
(Movahedi and Choi, 2025)	GPT-3.5-turbo	Decision Support and Prescriptive Optimisation	Introduces 2 LLM TSC controllers (1) ZS-CoT and (2) Actor-critique Generally Capable Agent (GCA) for traffic signal control	LLM based controllers outperform traditional (fixed-time, gap-based and delay-based TSCs). GCA outperforms ZS-CoT	Method not compared with advanced RL based TSC controllers. At this stage, the study is limited on only one isolated intersection, and other modes of transport (PT, emergency vehicles) are not considered here.
(Choi and Lim, 2025)	GPT-4o	Decision Support and Prescriptive Optimisation	**D3QN-LLM,** an algorithm that leverages a large language model (LLM) to dynamically adjust the weights of the RL reward function in real time, enabling efficient traffic signal control at intersections.	D3QN-LLM improves ATT, AQL and AWT by 25%, 33% and 40% respectively over fixed weights method. It also outperforms other baselines (PPO + LLM, DQN and fixed-time)	Only limited to one intersection, simple scenario. Also, no discussion of the exact method of how LLM makes decision based on epsilon and reward, may be prone to hallucination
(Yao et al., 2025)	GPT, Qwen, LLaMA, DeepSeek and Mistral	Decision Support and Prescriptive Optimisation	Improving TSC over RL based models by using vision-based perception tool to extract traffic information and an LLM agent controller with domain specific logical reasoning.	Proposed method outperforms various baselines (conventional, RL, DQN and LLM-assisted Light) for both general passenger car and emergency vehicle	Limited to single intersection, more vehicle types need to be included

Traffic Forecasting

Paper	LLM	Task	Focus	Result	Limitations

(Corrias et al., 2023)	Temporal Fusion Transformer (TFT)	Classification & Prediction (Forecasting)	Explores use of General-Purpose Transformer (GPT) (Not the same as GPT in ChatGPT) and Graph Convolutional Networks (GCN) based models for next-place prediction.	3 methods were used: TFT, Multivariate Time Series Graph Convolutional Neural Network (MTGCN) and Flashback-LSTM. All performed well on Dense dataset. Flashback-LSTM performed better on sparse dataset. TFT slightly better than MTGCN.	The authors only used one Transformer based model (and only one GCN based model). They also didn't explain the reason of choosing this over other extant models.
(Guo et al., 2024)	Fine-tuned LLaMA2 7B	Classification & Prediction (Forecasting)	xTP-LLM: LLM powered traffic flow prediction models which generates explainable traffic flow predictions.	xTP-LLM outperforms baselines (LSTM, DCRNN, STGCN, ASTGCN, GWNET, AGCRN, STTN, STGODE and DSTAGNN) on both short-term and long-term prediction tasks.	No LLM based model present in the baseline. Furthermore, instead of giving the last 12h of data, wouldn't it be better if the relevant time data is given, for example, to predict traffic flow during a sport match, it'll be more helpful to provide data from previous sport matches than last 12h.
(Sun et al., 2024)	Uses GPT-2 with AGFormer (arXiv:2305.07521)	Classification & Prediction (Forecasting)	AGFormer-GPT traffic flow forecasting model that treats historical road occupancy and speed as dual “prompts” and fuses them using cross-attention layers. By combining an adaptive graph neural network—capturing the spatial relationships between road segments—with transformer-style attention mechanisms	AGFormer-GPT outperforms baselines (LSTM, STGCN, ASTGCN, STSGCN, STFGNN, STTNs and STGNCDE) on both accuracy metrics (RMSE, MAE & MAPE) and correlation metrics (Pearson and Spearman)	The "prompt engineering" used by the authors is different from conventional prompt engineering. Instead, the authors discuss the cross-attention mechanism. Hence, the use of the term "prompt engineering" is misleading here. The authors also mention GPT-2 as transformer encoder,

			from large language models, it jointly learns spatial and temporal dependencies in traffic data.		however, GPT-2 is transformer decoder.
(Gong et al., 2024)	TinyLlama-1b (arXiv:2401.02385)	Classification & Prediction (Forecasting)	Mobility-LLM, an LLM powered framework for check-in analysis task which includes location prediction, trajectory user link and time prediction by extracting semantics of check-in sequence. The authors also introduce Visiting Intention Memory Network (VIMN) to capture user's visiting intentions and a shared pool of Huan Travel Preference Prompts (HTTP) to enable comprehensive understanding of human travel preferences	Mobility-LLM outperform baselines (12 for LP and 9 for TUL). TinyLlama outperforms other baseline LLMs (7, including GPT-2)	None of the baselines are LLM based. The authors also didn't compare to vanilla LLMs, i.e., without additions such as VIMN and HTPP. The model is not generalisable as it is fine-tuned only on one dataset and different dataset might have different number of POIs
(Beneduce et al., 2025)	LLaMA 2 (7,13,70 B, normal and chat), GPT-3.5,4,4o, LLaMA 3 (8, 70 B normal and instruct), LLaMA 3.1 8b and Mistral 7B	Classification & Prediction (Forecasting)	Explores the use of (15) LLMs as next-location prediction tasks on three real-world dataset. The paper also assesses data contaminations risks and potential of using LLMs for explainability.	The LLM based approached outperformed DL baselines significantly. Furthermore, bigger models achieved higher score. LLaMA 3 70B being the best in zero and few shot, GPT-4o in one shot prompting	First, the authors limited strictly to the zero-shot prediction task with sparse data availability and geographical generalisation. In other cases, however, DL models are able to compete with or outperform LLMs. Second, the training dataset for the LLMs could be

					biased or incomplete, which might lead to undesirable outputs.
(Jiang et al., 2025)	Deepseek-v2.5	Classification & Prediction (Forecasting)	KIDCM, Knowledge-informed Dynamic Correlation Modelling which uses LLM with traditional predictive methodologies to achieve balance of accuracy and generalisability in lane-level traffic flow prediction. The authors also introduce General Spatial Dynamics Modelling (GSDM) uses LLM-generated synthetic traffic data to model dynamic spatial correlations over time.	KIDCM outperforms baselines (GRU, T-GCN, FDL, MDL, PIDL, PIGAT and vanilla LLM) on accuracy and generalisability (for generalisability, vanilla LLM is excluded from comparison). Dataset: I-24 motion	The authors claim that the data generated by LLM is unbiased pertaining to the fact that it isn't trained on a specific highway. However, it cannot be overlooked that the data on which LLMs are pre-trained might contain biases, which might lead to undesirable outcomes.
(W. Li et al., 2025)	BERT	Information Mining & Insight Generation	To apply LLMs to better understand the complex multimodal travel patterns of the urban residents using LightGMB to infer travel modes for each segment, then using BERT to convert them to semantic embeddings, and finally clustering them using DBSCAN.	The framework is able to cluster multi-modal trips more effectively using BERT embeddings. This allows for analysis of multi-modal transit and helps finding gaps.	Data set is biased (dominated by younger population). While some parts of the frameworks are critically analysed, the framework as a whole is not analysed against baselines.

(C. Liu et al., 2025)	GPT-2 and LLaMA 2 7b	Classification & Prediction (Forecasting)	**ST-LLM+**, LLM powered spatio-temporal traffic forecasting. It incorporates spatio-temporal embedding and uses Partially Frozen Graph Attention (PFGA) fine-tuned using LoRA.	ST-LLM+ outperforms baselines (7 GNN based, 3 attention based and 4 LLM based) on NYCTaxi and CHbike dataset on zero and few shot cases	Can only choose place from where the user has already been.
(Chen et al., 2025)	ChatGLM3-6b	Information Mining & Insight Generation	**LLM-MDC,** Interactive next location prediction using LLM via a multi-round continuous dialogue mechanism and a candidate enhancement method TOPSIS.	Comparable performance to baselines (6 DL & 6 LLM based models) on NYC and Tokyo check-in dataset and prioritizes interpretability.	LLM-MDC can only predict a location that has already been visited. Fails to incorporate other aspects like public holidays, weather, etc
(Yu et al., 2024)	LLaMA2 7B and Gemma (version not specified)	Classification & Prediction (Forecasting)	**LLM-COD:** A novel framework using large language models (LLMs) for cross-city origin-destination (OD) flow prediction in urban transport.	LLM-COD significantly outperforms state-of-the-art methods like Random Forest, Gravity Model, GBRT, and GODDAG, reducing RMSE by up to 46% against GODDAG for high precision OD flow prediction; LLM-COD achieves better scores in all main metrics	No reasoning / justifications provided, something which is very common in such LLM based prediction framework
(Ju et al., 2025a)	LLaMA-3.1 8b Instruct and GPT-4o-mini	Synthetic Data Generation	**TrajLLM** The main aim of the paper is to create a modular framework that uses large language models (LLMs) to generate realistic human mobility trajectories for simulation.	LLM-driven simulation closely matches real-world patterns and is both scalable and interpretable; the model is compared to LLMob (pure LLM) and CoPB (LLM with gravity model) baselines and achieves more realistic and adaptable simulations.	Lack of proper benchmarks and baselines
(Han et al., 2025)	LLaMA-3.1-8B	Classification & Prediction (Forecasting)	**LLM-Next:** The main aim of the paper is to adapt large language models to better	LLM-Next outperforms state-of-the-art baselines—including LSTM, STAN, GETNext, STHGCN, GPT-3.5-Turbo,	Text modality conversion loss, prediction redundancy,

			understand spatio-temporal data for predicting the next point-of-interest (POI) a user will visit.	LLM-Mob, and LLM-Move—on all datasets in Acc@1	
(Luo et al., 2024)	GPT-4	Classification & Prediction (Forecasting)	**TSI-LLM** The main aim of the paper is to infer detailed semantic information from human mobility trajectories using large language models, overcoming the limitations of previous methods that rely on auxiliary datasets or provide shallow activity analysis.	TSI-LLM framework produces logical and interpretable trajectory semantic inferences across three levels, outperforming traditional approaches that rely on travel survey data; baselines mentioned include logit regression, Bayesian rule-based, hidden Markov, and white box models	more quantitative and comprehensive evaluation on large-scale datasets is necessary to rigorously assess semantic reasoning performance and generalizability. Downstream application not clear
(Zhong et al., 2024)	GPT-4 Turbo	Classification & Prediction (Forecasting)	The main aim of the paper is to use large language models (LLMs) to predict human mobility, especially under interventions such as the COVID-19 pandemic, using a novel framework called HMP-LLM.	HMP-LLM outperforms baselines (XGBoost, ARIMA, MLP, Random Forest, DeepSTN) with lower errors, especially in short-term predictions	The process of updating prediction based on covid data is not completely clear and is not transferable to other rare events or geographies.
(L. Liu et al., 2025)	GLM3, GPT-4 and Qwen2-72b-instruct	Information Mining & Insight Generation	**T2TrajLLM**: The main aim of the paper is to develop a new method for extracting individual mobility trajectories from text using large language models fused with structured domain knowledge	T2TrajLLM achieves about 8% higher accuracy than baselines, outperforming methods like BERT, mT5, and event-extraction models in robustness and transferability.	Language ambiguity

(Feng et al., 2024)	GPT-3.5-turbo	Classification & Prediction (Forecasting)	**LLMmove** The main aim of the paper is to explore if large language models (LLMs) can predict the next point of interest (POI) a user will visit, without training task-specific models.	LLMmove framework using gpt-3.5-turbo outperforms all baselines, including Popu, Dist, CZSR, LLMRank, ListRank, LLMMob, LLMMob(-Time), and LLMMob(+Geo), in zero-shot next POI recommendation accuracy	The performance depends higly on the underlying models, so a poor model will yield poor performance, and no optimisation is done.
(Gao, 2025)	LLaMA	Classification & Prediction (Forecasting)	**BRD-LLAMA**, a prediction framework for bike rental demand using a large language model adapted to handle time series data	BRD-LLAMA achieves much lower MAE (24.94) and RMSE (36.22) than all baseline models (XGBoost, LSTM, Transformer, AutoFormer), reducing MAE by nearly 50% compared to AutoFormer	It's primarily time series prediction, and doesn't take into account other factors like day of week, holiday, weather, etc.
(Y. Zhang et al., 2024)	LLaMA-3 and GPT-2	Synthetic Data Generation	to use large language models to generate realistic daily human activity patterns for urban populations, helping urban planning and transport research	LLaMA-3 8B fine-tuned with LoRA generates more diverse and realistic activity patterns than GPT-2 or baseline models, with baselines including vanilla GPT-2 and vanilla LLaMA-3 8B.	Only compared agains LLM baselines. Limited to temporal activity prediction only. Some of the examples are too basic.
(P. Tang et al., 2024)	LLaMA-3 8B	Classification & Prediction (Forecasting)	The main aim of the paper is to show that instruction-tuning Llama-3-8B improves long-term citywide human mobility prediction for multiple cities using large-scale trajectory data.	Llama-3-8B-Mob outperforms the LP-Bert baseline in both DTW and GEO-BLEU across cities, achieving top-10 ranking in a major challenge.	While LP-Bert is a good baselines, it's the sole baseline. There too, no numeric comparison is provided by the paper.
(Padoan et al., 2024)	GPT-4-Turbo	Question Answering & Knowledge Retrieval	The main aim of the paper is to develop a chatbot that supports decision-making using mobility datasets by	Key outcomes show the chatbot can answer complex queries, generate SQL statements, and visualize data; it handles SQL-based reasoning but	No evaluations of the performance or accuracy.

			interacting with data in natural language	struggles with certain error cases, such as division by zero, and is compared against traditional dashboard tools like Tableau and Power BI.	
(Zheng Zhang et al., 2024)	GPT-3.5-turbo, GPT-4 and Claude-2	Information Mining & Insight Generation	The main aim of the paper is to assess how well large language models (LLMs) can detect anomalous patterns in human mobility trajectories using both real and simulated datasets.	Claude-2 outperformed all non-deep learning baselines and matched deep learning models on Geolife, while GPT-3.5 outperformed all methods on the Patterns-of-Life dataset; baselines include OMPAD, MoNav-TT, TRAOD, DSVDD, and DAE.	The hints used to improve performance are not shown, the outliers are pattern showing cases (hunger, social and work), however cases like infectious disease monitoring (mentioned in abstract) are not discussed
(Liang et al., 2024)	GPT-4	Classification & Prediction (Forecasting)	The main aim of the paper is to explore how large language models can improve human mobility prediction during public events, especially by leveraging event textual data	GPT-4 (LLM-MPE) outperforms traditional models like Linear Regression, ARIMAX, and XGBoost, especially on event days, by better incorporating textual event data	In case of previously unseen events, the model "guesses" the mobility which can be often wrong.
(Qin et al., 2025)	GPT-3.5-Turbo	Classification & Prediction (Forecasting)	The main aim is to use a large language model (LLM) with specially designed spatiotemporal prompts to predict an individual's next public transport trip start location.	LingoTrip outperforms 1-MMC, DeepMove, MobTcast, MHSA, and Mob-LLM, especially for small and medium training sample sizes, and achieves the highest accuracy and F1 score among all compared methods.	While it predicts based on historic data, it doesn't include additional factors for semantics like day of week, weather, etc.

Traffic Safety

Paper	LLM	Task	Focus	Result	Limitations
(Jaradat et al., 2024)	Fine-tuned GPT-2 (and GPT-3.5 for data extraction)	Information Mining & Insight Generation, Information Mining & Insight Generation	Leveraging LLM for road traffic crashes analysis from tweets using classification and multitask learning	The fine-tuned GPT-2 performed better on classification (Accuracy, precision, recall and F1), BLEU, ROUGE and WER than baselines (XGBoost and GPT-4o mini)	This study relies on the fact that all information on social media is true and reliable, which is often not the case. The authors justified the choice of GPT-2 over bigger models by saying that they'll be to computationally expensive to fine-tune. However, they could have used LoRA or QLoRA.
(Trinh Xuan et al., 2024)	Mistral 7B for segment extraction, MiniLM for embedding extraction and Qwen-VL for training.	Visual Scene Understanding	Introduces a framework for extraction of precise traffic information for video for traffic safety. Crucial segments are first extracted, and then they are critically analysed to generate detailed descriptions.	A framework that has better capabilities of 1) segmenting crucial segments from video and 2) capturing temporal information using vision LLM (Qwen-VL). The authors secured second rank for this approach in Track 2 in the AI City Challenge 2024.	No comparison against baselines (other teams in the competition). The metrics are only used in ablation studies.
(Zhen et al., 2024)	GPT-3.5-Turbo, LLaMA 3-8B and LLaMA4-70B	Information Mining & Insight Generation	Using LLMs for crash severity analysis as a classification task.	LLaMA3-70B consistently exhibits superior performance. Techniques like CoT and prompt engineering improve the performance. GPT-3.5's macro-F1 rose from 0.1812 to 0.2073 in zero-shot CoT	LLMs are not compared any against other non-LLM baselines
(Ahmed et al., 2024)	GPT-4-Turbo and Moondream-2(for vision tasks)	Visual Scene Understanding	Introduces a framework for automated highway safety management which integrates computer vision and natural language	Creation of a custom dataset combining drone and CCTV videos from different public sources. A framework that detects accident using YOLOv11, Moondream2 for	Not compared with any baselines. Only 22 images were used for training YOLO, which is not a good

			processing for real-time monitoring, analysis and reporting of traffic incidents from CCTV and drone imagery	vision-language interpretation and GPT-4-Turbo for contextual summarisation and response generation.	amount to train or test a system
(Zhen and Yang, 2025)	*No LLM used.* Instead uses ELECTRA (transformer based)	Information Mining & Insight Generation	Leveraging LLMs to integrate tabular data and descriptive narrative to perform crash severity modelling.	Outperforms baselines (CatBoost, Transformer based, and multinomial logit model) in crash severity inference.	Lack of explainability and use of a rigid template may overlook scenario specific nuances.
(Hu et al., 2025a)	ChatGPT-2, LLaMA-1, CogVLM2, LLaMA-3.2, Gemini 1.5 Pro and ChatGPT-4o	Visual Scene Understanding	LLM based framework to detect things like fake traffic signs and forged motion information by fine-tuning LLMs using LoRA	LLMs outperform DNNs for fake traffic sign and forged motion detection in terms of accuracy metrics but have higher latency. In LLMs, GPT4o performs the best but has higher latency.	Rarely, LLMs might take an actual traffic sign for a fake when, what are the fail safes for such scenarios. For forged motion, the algorithm relies on plausible data like the speed limit, however in edge cases, a vehicle could actually be exceeding it.
(Hu et al., 2025b)	GPT-o1, LLaMA-3.2 and deepseek-R1	Information Mining & Insight Generation, Visual Scene Understanding	Traffic safety framework leveraging LLMs on cloud and edge networks for risk prediction and accident analysis. The framework consists of a) road segment risk prediction, b) scene level risk prediction and c) accident severity analysis	LLMs outperform ML baselines (SVM, MLP and LSTM) for segment risk prediction and accident severity analysis	This papers works on the assumption that cloud-edge network are accessible all the time, and no problem will be encountered in network communications (like packet drops).
(Ding et al., 2025)	InternVL	Visual Scene Understanding	**URA-VLM**, a gen-ai based framework for monitoring	URA-VLM outperforms the baseline (ResNet34) in terms of accuracy in identifying the type of safety	Why is the framework compared only against ResNet34. Comparison

			of diverse urban road anomalies.	hazards. Optimized prompt achieved better performance	against more robust models is required
(Jaradat et al., 2025a)	GPT-4o	Visual Scene Understanding	Leveraging deep learning models to segment video streams and identify potential near-miss or crash, then using MLLMs to further analyse and extract narrative information from the event.	CNN outperforms Vision Transformer and CNN+LSTM in segmenting videos, and GPT-4o produces narratives for the near miss events	Human verification bottleneck, near-miss are subjective. Even boundary ambiguity, not good at identifying when the event started.
(Liao and Lin, 2025)	LLAVA	Information Mining & Insight Generation, Visual Scene Understanding	**DDC-Chat,** Leveraging fine-tuned VLM for distracted driving classification.	DDC-Chat outperforms baselines (LLAVA variants and ML models) in classification task	Bias towards "safe driving" output in ambiguous cases.
(Ashqar et al., 2025)	*Various*	Visual Scene Understanding	Comprehensive review and empirical testing of MLLMs and VLMs for road safety object detection	A critical review of 8 papers, highlighting the benefits, potentials, challenges and limitations	While the performance of each model is presented, the improvement over their respective baselines is not discussed. Only Object detection improvement is discussed
(Q. Liu et al., 2025)	ChatGPT, CLIP and derivatives	Visual Scene Understanding	Review of traffic anomaly detection approaches using ML methods like CNN, RNN, GAN, MLLMs, etc.	A comparison of ML, CNN, GAN, Transformer and MLLM based anomaly detection approaches	No standardised comparison between the methods, or even within each category (like CNN based models)
(Pérez et al., 2025)	LLaMA 3	Classification & Prediction (Forecasting)	The main aim is to develop a system using an LLM to predict collision risk and determine when driver takeovers are needed,	The CPTR-LLM model outperforms statistical and AI baselines (like ARDL, AMG, XGBoost, LSTM), offering higher accuracy and	Dataset is limited to only one location, and is monotonous

			based on radar-video sensor data.	reliability in collision risk prediction and takeover assessment.	
(Jaradat et al., 2025b)	GPT-2 (fine-tuned), GPT-3.5, and GPT-4.5, with GPT-4.5	Information Mining & Insight Generation	The main aim of the paper is to show that combining tabular and textual crash data using large language models improves traffic crash analysis and predictive accuracy	GPT-4.5 few-shot outperforms the baselines (fine-tuned GPT-2 and GPT-3.5), achieving 98.9% and 98.1% accuracy for crash severity and driver fault; it also attains the highest Jaccard scores for crash factor and driver action extraction	Not compared with other statistical / ML baselines
(Abu Tami et al., 2024)	Gemini-Pro-Vision 1.5	Visual Scene Understanding	The paper aims to show that multimodal large language models (MLLMs) can automatically detect and analyse traffic safety-critical events from driving videos in a scalable, accurate way	Few-shot learning with Gemini-Pro-Vision 1.5 achieved the highest overall accuracy (about 79%), outperforming zero-shot learning, self-ensemble learning, image-augmented methods, and baselines like LLaVA-1.5 and previous visual-language QA models	Using LLM to find the distance of traffic hazard is not very reliable. Only Gemini is used
(Calenzani et al., 2024)	GPT-4V (vision)	Visual Scene Understanding	The main aim is to evaluate how well GPT-4V can classify risky driver behaviours in videos using just a few frames from each video.	GPT-4V achieved up to 98.9% accuracy for yawning, 98.4% for smoking, 95.7% for cell phone use, 94.1% for face not visible, and 91.7% for distraction, generally outperforming prior computer vision baselines and matching or surpassing human expert performance; explicit baselines include traditional computer vision systems and human labelling	Paper only restricted to GPT-4V, not compared with other baselines, LLM or otherwise.
(Arteaga and Park, 2025)	Flan-UL2 (20B), Llama-2-13b-chat, and ChatGPT (gpt-3.5-turbo)	Classification & Prediction (Forecasting)	The paper aims to create and evaluate a large language model (LLM) framework to detect underreported factors,	The Flan-UL2 LLM achieved up to 96% F1 (recall 1.0, precision 0.93), significantly outperforming traditional manual review and machine learning text classification	Limited baselines comparisons. Also, it is not discussed how complex or simple is it to extract labels. Is the label (say alcohol)

			specifically alcohol involvement, in traffic crash reports by analysing crash narratives.	baselines like logistic regression and SVM used in prior studies.	mentioned in the text, or did the LLM infer it?
(Venkatesh Raja et al., 2024)	ChatGPT is used for comparison purposes, but the exact model name (e.g., GPT-3.5, GPT-4) is not specified in the paper	Information Mining & Insight Generation	The paper aims to develop a case-based reasoning system to investigate and troubleshoot the causes of road accidents using AI techniques	The proposed CBR system demonstrated high retrieval accuracy compared to ChatGPT predictions, but specific numerical accuracy values and detailed baseline comparisons are not provided in the visible content.	Synthetic data expansion (1,000 to 1,000,000) may introduce unwanted biases and may not reflect real-world distributions
(Costa et al., 2024)	GPT-3.5-turbo	Question Answering & Knowledge Retrieval	The main aim of the paper is to develop a method that uses large language models (LLMs), AI agents, and open geospatial data to promote safer cycling in cities by providing user-friendly safety information.	AI system that gives cyclists and planners reliable, real-time safety information for cycling journeys—helping promote safer, more sustainable urban mobility.	No real-world validation, lack of baselines comparison
(Yulianda et al., 2024)	GPT-3.5-turbo	Visual Scene Understanding	The main aim of the paper is to develop an Indonesian traffic sign recognition system using deep learning with interactive voice output to enhance driver safety.	The best model achieves up to 97.5% mAP50 and 82.9% mAP50-95 using YOLOv8, outperforming the baseline (YOLOv8 without augmentation, precision 72.8%, mAP50 76.1%); comparison baselines include prior research and YOLOv8 with default settings.	Pretty rudimentary. To some extent, LLM is not needed in the pipeline
(R. Zhang et al., 2025)	GPT-4o	Visual Scene Understanding	The main aim of the paper is to develop and evaluate SeeUnsafe, a multimodal large language model	The key result is that SeeUnsafe (with GPT-4o) achieves the highest accuracy (76.3%) and is the only method capable of successful visual	Some of the models used (object detection and segment masking) are not mentioned explicitly. Paper

			(MLLM) framework for video-based traffic accident analysis to improve road safety.	grounding, outperforming baselines including GPT-4o (vanilla), GPT-4o mini (vanilla), LLaVA-NeXT (video), and VideoCLIP.	just says state-of-the-art models
(Shafiq et al., 2024)	Florence-2	Visual Scene Understanding	The main aim of the paper is to use vision large language models (VLLMs) to improve detection of unexpected road anomalies for autonomous vehicles.	The key result is that the proposed VLLM-based method achieves mAP@50–95 of 50.1 (Lost and Found) and 45.8 (Road Anomaly), competitive with state-of-the-art baselines like Language Anchors, Perspective Aware, and Erasing Objects; it outperforms DETR, YOLOS, and YOLO V10.	The generalisability post fine-tuning (caption bounding) is not discussed for diverse locations
(Bin Zaman Chowdhury et al., 2024)	GPT-4, GPT-3.5, and Llama-3	Information Mining & Insight Generation	To automate the creation of comprehensive road accident datasets in Bangladesh by combining web scraping with Large Language Models (LLMs).	GPT-4 performed best (99% accuracy), Llama-3 was nearly as good (96%), and both outperformed GPT-3.5 and the previously used Google BERT baseline; GPT-3.5 lagged behind at 83% accuracy.	As only scrapable websites are considered, framework might miss out on accidents not reported by scrapable websites.
(Xu et al., 2024)	GPT-4V and Mixtral-8x7B	Visual Scene Understanding	The main aim of the paper is to develop a large language model-based multimodal warning system that delivers personalized driver assistance through multimodal interactions.	The key results show the system generates personalized and explainable multimodal warnings tailored for different driver profiles across various scenarios; baseline models are not explicitly compared in experiments.	Databases not listed. What does RAG returns is also not discussed
(Arefeen et al., 2024)	InternLM-Xcomposer2 (InternLM-1.8B) and LLAVA-1.5-v2-7B; ChatGPT for RAG	Visual Scene Understanding	The main aim of the paper is to speed up multi-camera traffic video analysis by quickly converting videos to text with optimized Vision-	TrafficLens achieves 2–4× faster video-to-text conversion than baseline per-camera VLM processing, while maintaining or improving information diversity,	A common limitation with VLMs is their inability to count, the paper doesn't assess this.

			Language Model (VLM) and Large Language Model (LLM) processing.	using InternLM-1.8B and LLAVA-7B as baselines.	
(Lee et al., 2024)	KoAlpaca	Visual Scene Understanding	The main aim of the paper is to develop a system that helps visually impaired people walk safely by combining real-time multi-object detection with natural language sentence generation for guidance.	The key results show high object detection accuracy (average 88.84%) and recognition (98.68%) using YOLOv5, with usability scores of 4.05 from pilot users; the YOLOv5 nano model was used for efficiency and compared to the small model with only ~1% lower performance but much smaller size.	While the framework helps with what is ahead, it cannot locate things explicitly. The authors also didn't discuss about moving objects which might be too quick for proper detection.
(Abdelrahman et al., 2025)	phi-3-mini (for real-time reports) and phi-3-medium (for historical analysis).	Visual Scene Understanding	The main aim of the paper is to introduce a privacy-preserving pedestrian monitoring framework (VTPM) that generates concise textual reports from intersection video footage using large language models.	The key results show that the fine-tuned phi-3-mini LLM outperforms baselines like Llama-3-8B, Mistral-7B, and Gemma-7B in BLEU, METEOR, and most ROUGE scores; phi-3-medium is also rated highest in domain-specific summarization and knowledge by human experts.	Video-to-text, especially by smaller models might omit safety critical details. Might not be able to capture spatial attributes (how close pedestrian is standing to conflict zone).
(Peruski et al., 2025)	LLaVA (based on the Llama series); VILA and GPT-4 Vision are also evaluated.	Visual Scene Understanding	The main aim is to improve traffic monitoring and anomaly detection in remote areas using multimodal large language models on edge AI devices.	VILA outperforms LLaVA and GPT-4 Vision, showing higher average accuracy and more consistent response times; GPT-4 Vision has lower accuracy and safety limitations as a baseline.	data imbalance. Dataset used primarily represents urban area, whereas target deployment is rural area.