# Large Language Models for Traffic and Transportation Research: Methodologies, State of the Art, and Future Opportunities\*

Yimo Yan<sup>a,b</sup>, Yejia Liao<sup>c</sup>, Guanhao Xu<sup>d</sup>, Ruili Yao<sup>e</sup>, Huiying Fan<sup>f</sup>, Jingran Sun<sup>g</sup>, Xia Wang<sup>h</sup>, Jonathan Sprinkle<sup>h</sup>, Ziyan An<sup>h</sup>, Meiyi Ma<sup>h</sup>, Xi Cheng<sup>i</sup>, Tong Liu<sup>j</sup>, Zemian Ke<sup>k</sup>, Bo Zou<sup>b</sup>, Matthew Barth<sup>l</sup>, Yong-Hong Kuo<sup>a</sup>

Abstract The rapid rise of Large Language Models (LLMs) is transforming traffic and transportation research, with significant advancements emerging between the years 2023 and 2025 – a period marked by the inception and swift growth of adopting and adapting LLMs for various traffic and transportation applications. However, despite these significant advancements, a systematic review and synthesis of the existing studies remain lacking. To address this gap, this paper provides a comprehensive review of the methodologies and applications of LLMs in traffic and transportation, highlighting their ability to process unstructured textual data to advance transportation research. We explore key applications, including autonomous driving, travel behavior prediction, and general transportationrelated queries, alongside methodologies such as zero- or few-shot learning, prompt engineering, and fine-tuning. Our analysis identifies critical research gaps. From the methodological perspective, many research gaps can be addressed by integrating LLMs with existing tools and refining LLM architectures. From the application perspective, we identify numerous opportunities for LLMs to tackle a variety of traffic and transportation challenges, building upon existing research. By synthesizing these findings, this review not only clarifies the current state of LLM adoption and adaptation in traffic and transportation but also proposes future research directions, paving the way for smarter and more sustainable transportation systems.

Keywords: Large Language Models, Natural Language Processing, Transportation, Traffic, Logistics

<sup>&</sup>lt;sup>a</sup>Department of Data and Systems Engineering, the University of Hong Kong, Hong Kong SAR, China

<sup>&</sup>lt;sup>b</sup>Department of Civil, Materials, and Environmental Engineering, University of Illinois Chicago, IL, USA

<sup>&</sup>lt;sup>c</sup>Department of Electrical Engineering, University of California, Riverside, CA, USA

<sup>&</sup>lt;sup>d</sup>Buildings and Transportation Science Division, Oak Ridge National Laboratory, TN, USA

<sup>&</sup>lt;sup>e</sup>University of California, Riverside, CA, USA

<sup>&</sup>lt;sup>f</sup>School of Civil and Environmental Engineering, Georgia Institute of Technology, GA, USA

<sup>&</sup>lt;sup>g</sup>University of Texas at Austin, TX, USA

<sup>&</sup>lt;sup>h</sup>Department of Computer Science, Vanderbilt University, TN, USA

<sup>&</sup>lt;sup>i</sup>Cornell University, NY, USA

<sup>&</sup>lt;sup>j</sup>Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign, IL, USA

<sup>&</sup>lt;sup>k</sup>Google Inc., CA, USA

<sup>&</sup>lt;sup>1</sup>Department of Electrical Engineering, University of California, Riverside, CA, USA

<sup>\*</sup>This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The publisher acknowledges the US government license to provide public access under the DOE Public Access Plan (https://www.energy.gov/doe-public-access-plan).

#### 1. Introduction

Traffic and transportation have been pivotal in shaping human civilization throughout history. From the rise and fall of empires driven by maritime trade routes to the development of intricate road networks facilitating urban expansion, the movement of people and goods has always been a cornerstone of societal advancement since the 20th century Before Christ (Gianpaolo et al., 2013). Efficient transportation systems have enabled economic growth, cultural exchange, and technological progress, while also presenting challenges related to congestion, safety, and environmental impact.

In the 20th century, the advent of computer technologies revolutionized traffic and transportation research. The introduction of optimization algorithms and prediction models has allowed for more systematic and efficient planning of transportation networks. These advancements have enabled better traffic management, route optimization, and forecasting of transportation demands, significantly improving the functionality of transportation systems. However, despite these technological strides, several persistent issues remain unresolved. Modern transportation systems generate vast amounts of heterogeneous data, encompassing numerical metrics, videos, images, and unstructured textual information from diverse sources such as traffic reports, social media, and sensor logs. Traditional optimization and prediction algorithms, while powerful, often struggle to integrate and interpret this multifaceted data effectively.

Recent developments in artificial intelligence, particularly Large Language Models (LLMs), have the potential to address these challenges. LLMs, such as generative pre-trained transformer(GPT)-4, bidirectional encoder representations from transformers (BERT), and their derivatives are advanced artificial intelligence (AI) systems trained on extensive datasets to understand, generate, and manipulate human language with high proficiency. These models leverage Transformer architectures (Vaswani et al., 2017), enabling them to capture complex linguistic patterns and contextual relationships. Beyond natural language processing (NLP), LLMs exhibit capabilities in reasoning, data integration, and multimodal understanding, making them well-suited for applications in traffic and transportation research.

LLMs can undertake a variety of tasks critical to enhancing transportation systems. They can automate the extraction and summarization of information from unstructured data sources, improve the accuracy of traffic forecasts by integrating textual and numerical data, assist in scenario generation for planning and emergency response, and facilitate better decision-making through sophisticated data analysis and interpretation. These capabilities not only enhance the efficiency and safety of transportation systems but also contribute to sustainability by optimizing resource allocation and reducing emissions.

The purpose of this paper is to provide a comprehensive review of recent methodologies and applications of LLMs in traffic and transportation research. Our goal is to present and highlight the state of the art and the potential of LLMs within the traffic and transportation research community, thereby

outlining promising directions for future research. The specific research questions to be addressed include:

- In which areas of traffic and transportation research are LLMs more promising for adoption?
- Which LLM methods are more appropriate to tackle specific traffic and transportation problems?
- What are the challenges and future opportunities for LLMs in traffic and transportation research?

Our paper is organized as follows. In Section 2, we introduce the background and the core methodologies in LLMs. In Section 3, applications are classified into two broad categories, namely *traffic* and *transportation*. In Section 4, we present statistics of current research trends and future directions. Finally, we conclude this paper in Section 5. The abbreviations used in the paper are presented in Table 1.

## 2. Background of LLMs

Between six and eleven months, a child typically starts to learn its language from the surrounding environment (Health, 2025). A newborn is exposed to an overwhelming amount of linguistic input – parents talking, sibling chattering, TV sounds, and even books they see. Initially, these exposures are noise. But gradually, through consistent exposure and the pattern recognition repertoire of infants' brains, the child begins to make sense of the sea of information (Jurafsky and Martin, 2025).

This remarkable process mirrors how LLMs learn, beginning with their data foundation. Just as children absorb massive amounts of language input during their formative years, LLMs begin their development with enormous text datasets. The parallel extends to the processing mechanism: just as human sensory organs (eyes and ears) perform initial pre-processing of linguistic input before neural transmission, LLMs employ sophisticated pre-processing techniques to transform raw text into processable tokens. The neural networks underlying LLMs mirror our brain's architectural principles, though in a simplified and artificial form.

Research shows that children learn approximately seven to ten words daily through reading and exposure, accumulating thousands of words by adulthood (Jurafsky and Martin, 2025). Similarly, LLMs process vast amounts of text during pretraining, building their foundational knowledge. This learning process is guided by the "distributional hypothesis," which suggests that both children and LLMs can learn meaning through observing how words appear together in context, forming intricate patterns of understanding through repeated exposure and contextual learning.

Knowledge acquisition follows a similar trajectory in both systems. Children progress from basic vocabulary to complex language understanding, much like how LLMs develop from basic pattern recognition to sophisticated language processing. After acquiring basic language ability, children begin to learn domain-specific knowledge, very much similar to that of post-training procedures such as fine-tuning of LLMs. This specialization phase allows both human learners and artificial systems to develop expertise in specific areas while building upon their foundational language understanding.

Table 1 Table of Abbreviations							
AD	Autonomous Driving	AI	Artificial Intelligence				
ALPACA	Instruction-Fine-Tuned LLaMA	ASR	Attack Success Rate				
ASR	Automatic Speech Recognition	ATC	Air Traffic Control				
ATFM	Air Traffic Flow Management	AVs	Autonomous Vehicles				
BERT	Bidirectional Encoder Representations from Transformers	BLIP	Bootstrapped Language Image Pretraining				
C-CLUE	Chinese Corpus for Language Understanding Evaluation	CLIP	Contrastive Language–Image Pretraining				
CMA	Cascaded Multi-Scale Attention	COPT	Cardinal Optimizer				
CoT	Chain-of-Thought	DL	Deep Learning				
DPO	Direct Preference Optimization	DRL	Deep Reinforcement Learning				
FL	Federated Learning	FLAN	Finetuned Language-Action Network				
FLRT	Fluent Student-Teacher Redteaming	FSL	Few-Shot Learning				
FSM	Flight Schedule Manager	FT	Fine-Tuning				
GCG	Greedy Coordinate Gradient	GDP	Ground Delay Program				
GRU	Gated Recurrent Unit	GPT	Generative Pre-trained Transformer				
GUI	Graphical User Interface	ICL	In-Context Learning				
IR	Information Retrieval	KD	Knowledge Distillation				
KG	Knowledge Graph	LLaMA	Large Language Model Meta AI				
LLMs	Large Language Models	LVLM	Large Vision Language Model				
LSTM	Long Short-Term Memory	LoRA	Low-Rank Adaptation				
MC	Model Construction / Multimodal Content	MFD	Multi-Modal Fusion Discriminator				
ML	Machine Learning	MLLM	Multi-modal Large Language Model				
MMLMs	Multimodal Large Language Models	MMBench	Multimodal Model Benchmark				
MT	Machine Translation	MT-Bench	Multitask Benchmark				
NER	Named Entity Recognition	NLG	Natural Language Generation				
NLP	Natural Language Processing	NLU	Natural Language Understanding				
NAS	National Airspace System	NN	Neural Network				
PaLM	Pathways Language Model	PEFT	Parameter-Efficient Fine-Tuning				
PPO	Proximal Policy Optimization	POS	Part-of-Speech (Tagging)				
QA	Question Answering	QLoRA	Quantized Low-Rank Adaptation				
RAG	Retrieval-Augmented Generation	RLAIF	Reinforcement Learning from AI Feedback				
RLHF	Reinforcement Learning from Human Feedback	RL	Reinforcement Learning				
RNN	Recurrent Neural Network	SA	Sentiment Analysis				
Т0	T5-like Multitask Text-to-Text Transfer Transformer	T5	Text-to-Text Transfer Transformer				
TTS	Text-to-Speech	TC	Text Classification				
$\mathrm{TL}$	Transfer Learning	TSC	Traffic Signal Control				
UAV	Uncrewed Aerial Vehicle	ULM	Unsupervised Language Model				
V2I	Vehicle-to-Infrastructure	V2V	Vehicle-to-Vehicle				
V2X	Vehicle-to-Everything	ViT	Vision Transformer				
ZSL	Zero-Shot Learning						

This acquired knowledge then transforms into practical application. Children eventually use their language knowledge for various purposes – asking questions, telling stories, expressing emotions, and engaging in complex dialogues. Similarly, LLMs can apply their learned patterns to diverse tasks ranging from summarization and question answering to text completion and translation under

appropriate interaction techniques. The versatility of both systems demonstrates how fundamental language understanding can adapt to serve numerous practical purposes.

This section establishes the foundation for understanding LLMs in traffic and transportation applications. It encompasses the essential components: data, training, integration with other tools, interaction techniques, evaluation metrics, and mainstream LLMs. Figure 1 shows some key aspects of LLMs.

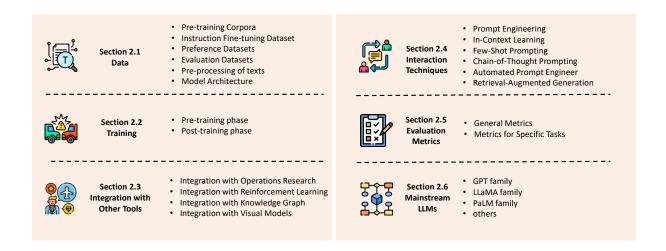


Figure 1 Key aspects of LLMs

## 2.1. Data

Just as a baby learns to understand and communicate by being exposed to varied and abundant stimuli – listening to conversations, observing facial expressions, and interacting with its surroundings – LLMs also rely on their "environment" to develop their extraordinary capabilities. For LLMs, this environment is the vast array of datasets that serve as the foundation for training, fine-tuning, and evaluation. Without high-quality datasets akin to the rich experiences of a child, an LLM cannot achieve its full potential in understanding and generating language.

In this section, we explore the "lifecycles" of datasets that shape LLMs, from the foundational pre-training corpora that teach basic language understanding to the instruction fine-tuning datasets that help LLMs follow human commands. We also examine preference datasets that align models with human expectations and evaluation datasets that measure their performance. These datasets, much like the milestones in a child's learning journey, are crucial for the development of an LLM's capabilities.

**2.1.1. Pre-training Corpora** The foundation of an LLM's "early learning" lies in its pre-training corpora, akin to the vast sensory input a baby receives during its formative years. These

Author: LLMs for Traffic and Transportation Research
Article submitted to; manuscript no.

extensive collections of text data enable LLMs to acquire fundamental language skills, such as grammar, semantics, and contextual understanding. Pre-training corpora are the largest and most diverse datasets in the LLM lifecycle, encompassing both general knowledge and specialized domain expertise.

## General Pre-Training Corpora

General pre-training corpora, like a child's exposure to everyday conversations and activities, provide LLMs with broad and diverse language knowledge from mixed domains. These corpora include a few major categories:

Webpages: Crawled texts from the internet, characterized by their massive scale, dynamic updates, multilingual content, and diverse themes (Xue, 2020; Penedo et al., 2023; Raffel et al., 2020).

Code: Datasets containing programming languages like Python, Java, and C++, enabling LLMs to perform tasks such as code comprehension and generation. Examples include The Stack (Kocetkov et al., 2022) and BIGQUERY (Nijkamp et al., 2022).

Parallel Corpus: Text pairs in different languages, vital for machine translation and cross-lingual tasks (Bañón et al., 2020; Ziemski et al., 2016).

**Encyclopedia**: Authoritative resources like Wikipedia and Baidu Baike, offering structured knowledge edited by experts.

#### Domain-Specific Pre-Training Corpora

As a child matures, it gains a deeper understanding in specific areas, like recognizing the nuances of a parent's profession or learning specialized vocabulary. Similarly, domain-specific pre-training corpora are tailored to particular fields, such as transportation, machine learning, or environmental protection. These datasets enhance LLMs' performance in specialized domains by building on the foundational knowledge acquired from general corpora.

For example, in the transportation field, the TransGPT-pt corpus \* provides data on transportation literature, engineering, and management, supporting model development in traffic-related applications.

2.1.2. Instruction Fine-tuning Datasets Instruction fine-tuning datasets act like a child learning to follow specific instructions or commands, such as "tie your shoes" or "say thank you." These datasets teach LLMs to follow human instructions across tasks like classification, summarization, and code generation, improving their ability to align with human expectations.

 $<sup>^*\,</sup>https://github.com/DUOMO/TransGPT$ 

# General Instruction Fine-tuning Datasets

General instruction datasets are designed to help LLMs follow commands across a wide range of tasks. These datasets are constructed using four main methods:

**Human Generation (HG)**: Manually created by annotators, offering high quality but limited scalability due to cost and time constraints (e.g., Databricks-dolly-15K (Conover et al., 2023), OASST1 (Wang et al., 2023a)).

Model Construction (MC): Generated by LLMs, providing abundant and cost-effective data (e.g., Alpaca (Taori et al., 2023a), UltraChat (Ding et al., 2023)).

Collection and Improvement of Existing Datasets (CI): Integrating and refining open-source datasets for diversity and scale (e.g., Flan 2022 (Longpre et al., 2023), InstructDial (Gupta et al., 2022)).

Combination Method: Combining approaches for optimal results (e.g., Firefly <sup>†</sup>, COIG (Zhang et al., 2023a)).

# Domain-specific Instruction Fine-tuning Datasets

Domain-specific datasets, like a child practicing a specific skill (e.g., playing an instrument), finetune LLMs for specialized tasks. For instance, TransGPT-sft <sup>‡</sup> is a transportation LLM fine-tuned based on domain-specific documents and can generate traffic-related dialogues.

2.1.3. Preference Datasets Preference datasets are akin to a parent guiding a child's behavior by providing feedback on what actions are good or bad. These datasets evaluate multiple responses to a single instruction, capturing human or model preferences through voting, sorting, or scoring (Zhao et al., 2023).

Vote: Selecting the better option between two or more answers, offering simplicity but limited granularity (e.g., Chatbot-arena-conversations (Zheng et al., 2023a), PKU-SafeRLHF (Ji et al., 2024), CValues (Xu et al., 2023), MT-Bench-human-judgments (Zheng et al., 2023a)).

**Sort**: Ranking responses based on predefined criteria, providing detailed insights but requiring significant effort (e.g., OASST1 (Wang et al., 2023a)).

Score: Assigning numerical values to responses for nuanced reflections of preferences (e.g., WebGPT (Nakano et al., 2021), Alpaca\_comparison\_data (Peng et al., 2023)).

Other: Alternative approaches, such as preference modeling (e.g., Medical-rlhf §).

A balanced approach, combining human and model feedback, ensures alignment with human expectations while mitigating bias.

<sup>†</sup> https://github.com/yangjianxin1/Firefly

<sup>&</sup>lt;sup>‡</sup> https://github.com/DUOMO/TransGPT

 $<sup>^{\</sup>S}$  https://github.com/shibing624/MedicalGPT

2.1.4. Evaluation Datasets Evaluation datasets represent the "student report cards" for LLMs, testing their performance across various domains and tasks. These datasets assess an LLM's knowledge, capabilities, and alignment with human values. The corresponding evaluation methods could be divided into three categories: code evaluation, human evaluation, and model evaluation.

General: Measures versatility across domains and ability to follow complex instructions (e.g., AlpacaEval (Dubois et al., 2024), MT-Bench (Zheng et al., 2023a)).

**Reasoning**: Tests logical reasoning and inference skills (e.g., Chain-of-Thought (CoT) Hub (Fu et al., 2023), TabMWP (Lu et al., 2022)).

Code: Evaluates programming proficiency, including code generation and debugging (e.g., HumanEval (Chen et al., 2021), CodeXGLUE (Lu et al., 2021a)).

Others: Covers specialized areas like safety, multilingualism, and academic tasks (e.g., Safety-Bench (Zhang et al., 2023b), C-CLUE ¶).

Much like the various stages of a child's development, the datasets used in LLMs play unique, critical roles in shaping their capabilities. From foundational learning in pre-training corpora to task-specific instruction fine-tuning, alignment with preferences, and rigorous evaluation, datasets are the driving force behind the success of LLMs.

2.1.5. Pre-processing of Texts Pre-processing transforms raw text into a standardized format for LLM analysis, critical for handling domain-specific data like transportation terminology, traffic reports, or infrastructure records. Key steps include data cleaning, normalization, and tokenization, which ensure consistency and computational interpretability.

#### **Data Cleaning and Normalization**

Cleaning removes irrelevant elements (e.g., HTML tags, sensor noise in traffic datasets) and standardizes domain-specific content, such as unifying road naming conventions (e.g., "St," "Street") or vehicle classifications. Normalization enforces uniformity by converting text to lowercase, expanding contractions, and harmonizing units (e.g., converting "mph" and "km/h" to a single metric). This step mirrors standardizing traffic signal terminology (e.g., "red phase" vs. "stop interval") to reduce ambiguity.

#### **Tokenization**

Tokenization splits text into subwords or characters, enabling LLMs to process complex domain vocabulary (e.g., decomposing "deceleration" into ["de," "celer", "ation"] or handling technical terms like "LiDAR"). Subword tokenization balances vocabulary size and out-of-vocabulary resilience, which is crucial for evolving domains like autonomous vehicles (AVs). Special tokens (e.g., classification

 $<sup>\</sup>P$  https://github.com/jizijing/C-CLUE

tokens and separator tokens) structure inputs for tasks such as classifying traffic incident reports or segmenting infrastructure descriptions. This approach aligns with scaling laws (Kaplan et al., 2020), optimizing model performance on sparse, domain-specific datasets.

2.1.6. Model Architecture Modern LLMs rely on advanced neural architectures to process tokenized text and generate human-like responses. The Transformer architecture, introduced by Vaswani et al. (Vaswani et al., 2017), revolutionized NLP by overcoming the limitations of earlier sequential models, such as RNNs and LSTMs. Its ability to handle long-range dependencies and enable parallel processing makes it the foundation of modern LLMs.

At the core of the Transformer are three key components: attention mechanism, which dynamically focuses on relevant input parts; positional encoding, which incorporates token order; and the encoder-decoder framework, which processes inputs and outputs. These components enable Transformers to excel in both understanding and generating text.

Transformers have evolved into three main architectural variants tailored to different tasks: decoderonly, encoder-only, and encoder-decoder models. Table 2 summarizes their primary characteristics and applications. Decoder-only models (e.g., GPT) specialize in text generation; encoder-only models (e.g., BERT) are optimized for understanding tasks; and encoder-decoder models (e.g., T5) excel in sequence-to-sequence tasks. This section provides a concise overview of these architectures.

For further details on the implementation of Transformers, readers can refer to more comprehensive resources (Hadi et al., 2023; Kalyan et al., 2021; Huang et al., 2023b).

## Attention Mechanism

The attention mechanism, particularly self-attention, is central to the Transformer's success (Vaswani et al., 2017). It evaluates relationships between tokens, enabling parallel processing and efficient modeling of long-range dependencies. Multi-head attention extends this by capturing diverse relationships within sequences. Variants like multi-query attention (Shazeer, 2019) and grouped-query attention (Ainslie et al., 2023) further improve scalability.

#### **Positional Encoding**

Transformers lack inherent sequential structure, so positional encoding incorporates token order into the model. Absolute positional encoding (Vaswani et al., 2017) assigns unique embeddings to each position, while relative positional encoding focuses on token distances, improving performance in tasks requiring contextual understanding.

#### Overview of Transformer Variants

Decoder-only models process text left-to-right for autoregressive generation tasks like text completion (e.g., GPT (Radford et al., 2018; Brown, 2020)). Encoder-only models, such as BERT (Devlin,

2018), use bidirectional self-attention for understanding tasks like classification and question answering. Encoder-decoder models (e.g., T5 (Raffel et al., 2020)) balance generation and understanding tasks, excelling in translation and summarization.

Training Architecture Components Primary Use Cases Attention Objective **Decoder-Only** Stacked Decoders Masked Autoregressive Text Generation (e.g., Self-Attention Next-Token GPT) (Unidirectional) Prediction Masked Language **Encoder-Only** Stacked Encoders Self-Attention Understanding Tasks (Bidirectional) Modeling (MLM) (e.g., BERT) Encoder-Stacked Encoders & Self-Attention Sequence-to-Translation, Decoder Decoders (Encoder) & Sequence Loss Summarization (e.g., Cross-Attention T5, BART) (Decoder)

Table 2 Comparison of Transformer Architectures

#### 2.2. Training

Building upon pre-processed data, LLMs undergo a two-stage training process akin to how children acquire language skills. This consists of pre-training for general language understanding and post-training for specialization. During pre-training, LLMs learn fundamental language patterns through tasks like next-token prediction (Radford et al., 2018) and Masked Language Modeling (MLM) (Devlin, 2018). In post-training, LLMs are fine-tuned to follow instructions, align with human preferences, or improve specific skills like coding and reasoning. For domain-specific tasks, fine-tuning on related datasets improves performance, enabling LLMs to better respond to domain-specific contexts.

**2.2.1. Pre-training Phase** Pre-training is unsupervised, involving large, diverse text corpora (e.g., books, websites). Two common pre-training methods are:

#### **Next-token Prediction**

This method (Radford et al., 2018) involves predicting the next word in a sequence, such as completing the sentence "The sun is shining in the..." with "sky." The objective is to minimize the autoregressive loss as shown in Equation (1):

$$\mathcal{L}_{AR}(\theta) = -\mathbb{E}_{(x_1, \dots, x_T) \sim \mathcal{D}} \left[ \sum_{t=1}^T \log P_{\theta}(x_t \mid x_{1:t-1}) \right]$$
(1)

## Masked Language Modeling (MLM)

MLM (Devlin, 2018) masks tokens in a sentence, requiring the model to predict them using bidirectional context. For example, in "The cat is [MASK] on the mat," the model predicts "sitting." The MLM loss is as shown in Equation (2):

$$\mathcal{L}_{\text{MLM}}(\theta) = -\mathbb{E}_{(x_1, \dots, x_T) \sim \mathcal{D}} \left[ \sum_{t \in M} \log P_{\theta}(x_t \mid x_{\setminus t}) \right]$$
 (2)

Where  $\t$  denotes the sequence without the token at position t.

**2.2.2.** Post-training Phase Post-training adapts pre-trained models to specific tasks and user needs.

#### **Instruction Tuning**

Fine-tuning refines models using datasets containing instructions, such as "Summarize this text," to improve their ability to handle diverse user requests. The associated loss function is based on next-token prediction and is applied to task-specific data, as shown in Equation (3):

$$\mathcal{L}_{\text{FT}}(\theta) = -\mathbb{E}_{(x_1, \dots, x_T) \sim \mathcal{D}_{\text{task}}} \left[ \sum_{t=1}^T \log P_{\theta}(x_t \mid x_{1:t-1}) \right]$$
(3)

## Alignment with Human Preferences

After fine-tuning for task-specific skills, aligning model behavior with human values involves methods like Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022). Human evaluators rank outputs (e.g., preferring  $x_1$  over  $x_2$ ), and a reward model  $R_{\phi}(x)$  is trained using pairwise ranking loss as shown in Equation (4):

$$\mathcal{L}_{\text{reward}}(\phi) = -\mathbb{E}_{(x_1, x_2) \sim \mathcal{D}_{\text{human}}} \left[ \log \frac{e^{R_{\phi}(x_1)}}{e^{R_{\phi}(x_1)} + e^{R_{\phi}(x_2)}} \right]$$
(4)

The LLM is fine-tuned using reinforcement learning to maximize expected rewards, with algorithms like Proximal Policy Optimization (PPO) as shown in Equation (5):

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_{(s,a) \sim \pi_{\text{old}}} \left[ \min \left( r(\theta) A(s,a), \text{clip} \left( r(\theta), 1 - \epsilon, 1 + \epsilon \right) \cdot A(s,a) \right) \right]$$
 (5)

where  $r(\theta)$  is the probability ratio between policies, A(s,a) is the advantage function.

#### Alignment with Human Preferences

Fine-tuning trains models on domain-specific datasets, improving their performance in fields like coding, law, or medicine. For example, fine-tuning for code generation uses the following loss as shown in Equation (6):

$$\mathcal{L}_{\text{code}}(\theta) = -\mathbb{E}_{(c_1, \dots, c_T) \sim \mathcal{D}_{\text{code}}} \left[ \sum_{t=1}^T \log P_{\theta}(c_t \mid c_{1:t-1}) \right]$$
(6)

Efficient techniques like adapter-based fine-tuning (Houlsby et al., 2019) or parameter-efficient fine-tuning (PEFT) (Hu et al., 2021) reduce computational costs by only updating small parts of the model.

#### 2.3. Integrating with Other Tools

Very much like a child learns to use different tools to facilitate his/her life, LLMs can learn to understand and utilize tools in the post-training phase, such as optimization solvers, reinforcement learning (RL), and knowledge graphs (KGs) to further enhance their productivity.

2.3.1. Integration with Operations Research Operations research (OR) is widely utilized in traffic and transportation domains to optimize decision-making under complex constraints and objectives. We find two streams of research that integrate OR with LLMs. The first stream of literature addresses the challenge of a steep learning curve for beginners and limited adoption in small businesses. The second stream of literature improves prompts with discrete optimization methods.

The first stream of literature combines LLMs with OR techniques to make these tools more accessible for industrial applications and education. Recently, a model named "Operations Research Language Model" (ORLM) based on fine-tuning is introduced to automate the process of optimization modeling (Huang et al., 2024a). Users can describe an optimization problem in natural language, and the model processes the input to identify key components such as decision variables, constraints, and objectives. It then translates these into a standardized mathematical representation. Once the mathematical model is generated, ORLM further converts it into code compatible with optimization solvers, such as COPT or Gurobi, enabling direct execution without manual intervention.

In addition to generating initial solutions, ORLM offers flexibility for dynamic business requirements. For instance, if a constraint like "airplanes must be used if ships are used" is introduced, ORLM can seamlessly update both the mathematical model and the solver-compatible code to reflect the new condition. This adaptability reduces the time and effort required to adjust models to evolving scenarios. An illustration is shown in Figure 2. In our literature search, we find that similar methods have already been used before the introduction of ORLM to address supply chain management problems (Li et al., 2023).

The second stream examines the application of operations research techniques to optimize prompts. However, relevant research has not been found in our literature search. Therefore, we direct readers to Subsection 4.2 under future research directions section.

2.3.2. Integration with Reinforcement Learning LLMs are increasingly being used to improve RL. A comprehensive review (Cao et al., 2024) highlights how LLMs leverage their pretrained capabilities – such as understanding, reasoning, and multimodal processing – to enhance RL efficiency and generalization in complex environments. There are four important aspects of research, as shown in Figure 3.

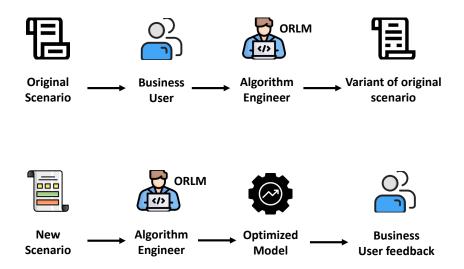


Figure 2 ORLM (Huang et al., 2024a)

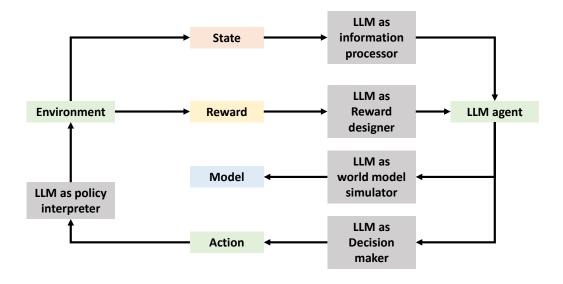


Figure 3 Integration of LLMs with Reinforcement Learning (Cao et al., 2024)

Information Processing: LLMs convert unstructured inputs like natural language instructions or visual data into structured representations usable by RL agents. This method has been used for traffic signal control (Villarreal et al., 2023; Pang et al., 2024a), which reduces ambiguity and noise in state features.

Reward Design: In environments with sparse feedback, LLMs automate reward function creation through two methods: generating implicit reward values via semantic analysis (e.g., prioritizing stability over speed for delicate objects) and producing explicit reward code that encodes task logic (Ma

et al., 2023). This method has been used in the public transportation domain for bus holding control to address the problem of sparse reward (Yu et al., 2024).

**Decision Guidance**: LLMs improve exploration efficiency by either directly generating actions (framing RL as sequence prediction) or constraining action spaces using semantic priors. This method has been used in traffic signal control (Pang et al., 2024a), drone control (Yang et al., 2024a) and bus holding control in public transportation (Prabhod, 2023), where unreasonable actions are removed to enhance efficiency.

Environment Generation: LLMs synthesize plausible environment dynamics and policy explanations, addressing model-based RL's need for accurate simulators. For example, Minecraft, a sandbox game, subgoal sequences are hypothesized to (e.g., "collect wood  $\rightarrow$  build tools") improve sample efficiency over pure RL methods (Nottingham et al., 2023).

2.3.3. Integration with Knowledge Graph LLMs are trained on vast amounts of unstructured textual data. Their knowledge is implicit, derived from statistical patterns in the data, rather than being explicitly organized or structured. While this allows them to generate coherent and contextually relevant responses, their outputs are not anchored in factual relationships or structured reasoning, resulting in hallucinations. To address this issue, researchers have explored integrating KGs with LLMs to enhance their reasoning capabilities and produce more reliable outputs.

There are two primary methods for combining KGs with LLMs: integration during pre-training and integration during inference (prompting). The latter is the most common and practical approach due to its flexibility.

A recent method introduces a plug-and-play, prompt-based approach to integrating KGs with LLMs. This approach involves using a KG as an input to the LLM during inference and designing prompts that guide the LLM to leverage the structured information from the KG for reasoning. For example, if a KG contains relationships about historical events, a prompt could explicitly reference this knowledge, enabling the LLM to generate a factually accurate and logically consistent response (Li et al., 2024c).

The knowledge solver framework, introduced by Feng et al., 2023, transforms information retrieval into an interactive, multistep decision-making process. Instead of treating KGs as static sources of information, this framework enables the LLM to interact dynamically with the KGs, retrieving and reasoning over relevant knowledge in iterative steps. This enhances the LLM's ability to answer complex, multi-faceted questions as well as solving problems requiring connections between multiple facts or entities. For example, if a KG contains relationships about historical events, a prompt could explicitly reference this knowledge, enabling the LLM to generate a factually accurate and logically consistent response.

Pan et al., 2024b provides a comprehensive review of recent efforts to integrate LLMs with KGs. We direct readers to this review for more detailed methodologies. In our literature review, we have identified papers applying combined LLM and KG methods in supply chain management (AlMahri et al., 2024) and autonomous driving (AD) (Hussien et al., 2025).

2.3.4. Integration with Visual Model Large vision-language models (LVLMs) are sophisticated AI systems designed to process and reason about both visual (e.g., images and videos) and textual (natural language) data simultaneously. By combining computer vision and NLP, LVLMs can understand visual content, generate descriptions, and answer questions about images or videos in a coherent and context-aware manner.

Popular vision-language models include contrastive language-image pretraining (CLIP) by OpenAI, Flamingo by DeepMind, and bootstrapped language-image pretraining (BLIP) by Salesforce. These models are generally pre-trained on large datasets of image-caption pairs, enabling them to associate visual features with text representations. Their core methodologies involve leveraging contrastive learning (e.g., CLIP aligns visual and textual embeddings in a shared latent space to maximize the similarity between matching image-text pairs) and multimodal transformers (e.g., Flamingo and BLIP use cross-attention mechanisms to integrate vision and language modalities). These architectural designs allow LVLMs to effectively adapt to new data and excel at diverse multimodal tasks with minimal additional fine-tuning (FT).

A common way to utilize these models is visual question answering (VQA), where LVLMs analyze an image and respond to specific queries about its content using zero-shot learning (ZSL) or few-shot learning (FSL). For example, when asked, "What does this sign represent?" while providing an image of a road sign, an LVLM might respond: "This sign indicates that all vehicles are only permitted to turn right or left" (Wang et al., 2024f). LVLMs also excel in object detection and scene understanding, where they identify objects in an image and reason about their spatial and contextual relationships. For instance, LVLMs can detect and describe the motion of vehicles to identify drivable paths (Wang et al., 2024i; Pan et al., 2024a; Li et al., 2024b; Li et al., 2024e), detect ships in maritime imagery (Zhang et al., 2024d), or interpret environmental contexts for drone navigation (De Curtò et al., 2023).

In autonomous systems, such as self-driving vehicles or drones, LVLMs combine visual perception with language-based reasoning to perform critical tasks. These include identifying obstacles, interpreting traffic signs, or following complex textual instructions. LVLMs can even be integrated into end-to-end systems that directly generate vehicle trajectory plans based on visual and textual inputs (Pan et al., 2024a). Their ability to unify vision and language into a single reasoning framework makes them a promising tool for intelligent and adaptive decision-making in real-world settings.

#### 2.4. Interaction Techniques

Building upon the trained model, effective interaction techniques are crucial for extracting meaningful responses from LLMs. Just as children need appropriate prompting and access to reference materials to articulate their knowledge effectively, LLMs require carefully crafted prompts and supplementary information to generate accurate and contextually relevant outputs. Two primary interaction techniques have emerged: prompt engineering and retrieval augmented generation (RAG), each serving distinct yet complementary roles in enhancing LLM performance in transportation applications.

**2.4.1. Prompt Engineering** Prompt engineering serves as the art of communication with LLMs, akin to how teachers carefully phrase questions to elicit specific responses from students. This technique involves crafting inputs that guide models toward desired outputs, especially critical in transportation contexts where precision and safety are paramount.

The practice of prompt engineering varies across application domains, with widely adopted approaches emerging. The simplest form is zero-shot prompting, where a task description is provided without examples or additional context. For instance, Prompt: "Is the following statement true or false: 'The first autonomous vehicle to complete a cross-country trip without human intervention was developed by Carnegie Mellon University in 1995."' Output: "False." Here, the LLM relies solely on its pre-existing knowledge. While sufficient for straightforward tasks, zero-shot prompting often fails for complex tasks like traffic flow prediction or route optimization.

For more advanced use cases, prompts can include a task description, the LLM's role, specific examples, and relevant context to improve responses. Just as teachers provide structured guidance with examples, these elements enhance LLM accuracy. Prompts can also vary structurally, using techniques such as few-shot prompting, CoT prompting (Wei et al., 2022), tree-of-thought prompting (Yao et al., 2024a), and graph-of-thought prompting (Besta et al., 2024). Some less common prompt techniques are listed in the Appendix.

#### **In-Context Learning**

In-context learning (ICL) involves providing demonstrations to LLMs as context (Dong et al., 2024), without requiring parameter updates. Effective ICL depends on selecting and organizing examples (Qin et al., 2023; He et al., 2023). Techniques like reformatting demonstrations with optimization methods (Dong et al., 2022) further enhance performance, making ICL a powerful tool for guiding LLM behavior.

## Few-Shot Prompting

Few-shot prompting provides a few input-output examples to help LLMs recognize patterns and generalize. For instance, in transportation-related tasks, examples like "The train from Boston to New York departs at 4:30 PM and arrives at 7:45 PM." paired with the question "How long is the journey?" and the answer "The journey is 3 hours and 15 minutes." guide the model for further questions. However, as noted by Lu et al., 2021b, the order of examples can significantly influence performance, necessitating careful prompt design.

#### Chain-of-Thought Prompting

Few-shot prompting can fail to reveal the logic behind a task, especially for problems requiring complex reasoning. CoT prompting improves performance by encouraging step-by-step reasoning (Wei et al.,

2022). For example: "Bill has 5 apples. He buys 2 buckets, each containing 6 apples. How many apples does he have?" CoT: "Bill starts with 5 apples and buys 2 buckets, each containing 6 apples. Total is  $5 + 2 \times 6 = 17$ . Output: 17." CoT prompting is particularly effective for tasks like mathematical reasoning and decision-making (Feng et al., 2024).

## **Automated Prompt Engineering**

Automated prompt engineering (APE) reduces reliance on manual trial-and-error by generating and optimizing prompts automatically. In retrieval and reasoning tasks, APE has been shown to outperform manual prompts (e.g., "let's think step by step") (Jin et al., 2024; Sahoo et al., 2024). Introduced by Zhou et al., 2022, APE uses LLMs to search over candidate prompts, framing the task as a black-box optimization problem. Each prompt is scored, and Monte Carlo search methods iteratively refine the best-performing candidates, significantly improving efficiency and accuracy.

2.4.2. Retrieval-Augmented Generation While prompt engineering focuses on how we ask questions, RAG enhances how LLMs access and utilize information, much like students consulting reference materials. RAG enables LLMs to complement their pre-trained knowledge with specific, up-to-date information, improving factual accuracy and reducing hallucinations (Lewis et al., 2020).

RAG consists of two components: a retriever and a generator. The retriever fetches relevant information from an external knowledge source, while the generator combines this retrieved context with the original prompt to produce informed responses. For instance, in transportation tasks requiring domain-specific knowledge, such as traffic regulations or infrastructure updates, RAG retrieves relevant sections from a knowledge base and integrates them into the generation process.

A common implementation involves segmenting the knowledge base into chunks, encoding them into vector representations, and ranking them by similarity to the prompt (e.g., using cosine similarity). The top-ranked chunks are passed to the generator, enabling more accurate, contextually enriched responses.

Advanced variations, such as advanced RAG and modular RAG (Gao et al., 2023), further optimize retrieval and generation. Modular RAG introduces additional tools, such as search modules, to integrate multiple data sources. For example, Wang et al., 2023d demonstrate improved accuracy by combining transportation databases to provide real-time updates on traffic conditions, route changes, and infrastructure.

## 2.5. Evaluation Metrics of LLMs

Just as educators assess children's language development through tests, LLMs require evaluation metrics to gauge their performance. This is especially important for transportation applications, where accuracy and reliability are critical. Evaluations encompass both general language capabilities and domain-specific competencies.

**2.5.1.** General metrics LLMs are assessed using standard metrics that evaluate accuracy, fairness, robustness, and calibration across a variety of tasks (Guo et al., 2017; Wang et al., 2021; Zhu et al., 2023):

## Accuracy:

Accuracy measures how well model outputs match the ground truth. Key metrics include:

- Exact Match: Evaluates whether the generated text matches the reference exactly.
- F1 Score: Combines precision and recall for classification tasks.
- ROUGE Score: Measures overlap between generated and reference texts, commonly used in summarization.

#### Calibration:

Calibration assesses how well a model's confidence aligns with its correctness. Examples include:

- Expected Calibration Error (ECE): Measures the gap between confidence and accuracy.
- Area Under the Curve (AUC): Evaluates performance in coverage and accuracy for selective predictions.

#### Fairness:

Fairness metrics ensure equal treatment across demographic groups. Examples include:

- Demographic Parity Difference: Measures differences in positive outcomes across groups.
- Equalized Odds Difference: Evaluates whether outcomes are consistent across groups, controlling for actual results.

#### Robustness:

Robustness measures a model's resilience to adversarial inputs or changes. Common metrics include:

- Attack Success Rate (ASR): Evaluates vulnerability to adversarial attacks.
- Performance Drop Rate: Quantifies performance degradation after adversarial modifications.
- 2.5.2. Benchmarks for Assessing Specific Tasks In addition to general evaluation metrics, many customized metrics are developed to assess LLMs in specific domains or tasks. These metrics are tailored to the unique demands of particular applications and are used to evaluate capabilities that general metrics might not adequately capture.

Multimodal Task Benchmarks: For evaluating the performance of multimodal large language models (MLLMs), benchmarks like MMBench (Liu et al., 2023) focus on assessing vision-language models by evaluating the model's capability to process and understand both visual and textual data. This includes performance across tasks such as perception and cognition.

Reasoning Tasks: The advanced reasoning benchmark (Sawada et al., 2023) focuses on evaluating LLMs in complex reasoning tasks across multiple domains, pushing models to handle more sophisticated problem-solving. More recently, a few evaluation methods have been proposed to evaluate specific reasoning tasks, such as coding (e.g., SWE bench (Jimenez et al., 2024)) and machine learning engineering (Chan et al., 2024).

Ethics and Bias in LLMs: TRUSTGPT is a customized benchmark designed to test the ethical considerations of LLMs, including metrics related to toxicity, bias, and value alignment (Huang et al., 2023a).

## 2.6. Comparison of Main-stream LLMs

LLMs are transformer-based pre-trained models (PLMs) with tens to hundreds of billions of parameters, demonstrating superior language understanding, generation, and emergent capabilities compared to smaller models. Notable LLM families include GPT, LLaMA, and PaLM (Zhao et al., 2023; Minaee et al., 2024).

# **GPT Family**

Developed by OpenAI, the GPT family includes some of the most widely used and influential models in NLP.

- **GPT-3** (Brown, 2020): With 175B parameters, it introduced emergent capabilities like ICL and excels in diverse tasks.
- ChatGPT : Based on GPT-3.5 and GPT-4, it is optimized for dialogue tasks and widely used for conversational AI.
- **GPT-4** (Achiam et al., 2023): A multimodal model that accepts both text and image inputs, demonstrating exceptional performance in professional exams.

#### LLaMA Family

Released by Meta, LLaMA models are open-source and widely used for research and FT.

- LLaMA (Touvron et al., 2023): Ranges from 7B to 65B parameters and features architectural modifications like SwiGLU activation.
  - LLaMA-2 (Touvron et al., 2023): Includes chat-specific models fine-tuned for dialogue tasks.

#### **PaLM Family**

Developed by Google, PaLM models excel in large-scale learning tasks.

• PaLM (Chowdhery et al., 2023): A 540B-parameter model achieving state-of-the-art results in FSL.

 $<sup>\</sup>parallel$  https://openai.com/blog/chatgpt

• PaLM-2 (Anil et al., 2023): An improved version of PaLM with better multilingual and reasoning capabilities.

#### Other LLMs

- Mistral-7B (Jiang et al., 2023): A smaller, high-performing model that outperforms LLaMA-2-13B in several benchmarks.
- DeepSeek-R1 \*\*: General-purpose LLM optimized for technical reasoning and coding tasks, featuring 7B-128B parameter variants with Mixture of Experts architectures.

There are also a large number of other LLMs. We refer the reader to the Appendix for other LLMs. All the above-mentioned LLMs contribute to the advancement of this field.

# 3. Applications in traffic and transportation research

In this section, we present how state-of-the-art LLM methodologies are applied in traffic and transportation research. While traffic and transportation are closely related, they differ in scope and focus. Traffic research refers to the flow of vehicles and pedestrians and the associated operational aspects within a transportation system, emphasizing how the different entities interact and move. Transportation, on the other hand, encompasses the entire system and infrastructure that enable the movement of people, goods, and services between locations. This section includes all transportation modes (e.g., road, air, and sea), associated facilities (e.g., airports, ports, and stations), policies, and planning strategies.

Before delving into the detailed review of the LLM applications in traffic and transportation research, we first present the organization of our review. The way we identify and select the relevant studies is also described.

## 3.1. Organization of review of applications

The review is structured into three main subsections based on the scope of the selected studies: traffic, transportation, and multi-task applications. As the studies in the first two subsections each deal with one specific task application using LLMs, we put multi-task applications as a separate subsection. In the traffic and transportation subsections, the selected studies are classified mainly by their application domains. Methodologies are discussed within these contexts. Characteristics of articles under the same application domain are summarized at the end of each subsection for clarity and coherence.

Subsection 3.2 describes our criteria for searching and selecting literature.

Subsection 3.3 focuses on studies examining the operational aspects of traffic systems. These include topics such as traffic management, travel behavior, traffic safety, and traffic infrastructure.

 $<sup>^{**} \ \</sup>mathrm{https://www.deepseek.com}$ 

Subsection 3.4 explores broader transportation systems and infrastructure, encompassing research on logistics, supply chain management, and autonomous driving. These studies investigate the movement of goods and people across various modes, including road, air, and sea.

Finally, subsection 3.5 addresses multi-task applications of LLMs that span multiple aspects of traffic and transportation. These works integrate diverse domains, offering insights into complex, multi-faceted challenges.

#### 3.2. Criteria for Literature Search and Selection

This section introduces our literature survey synthesis method and search strategy.

**3.2.1.** Synthesis Method We employ a scoping review approach, which is particularly suited for mapping the existing literature on a broad topic and identifying key concepts, theories, sources of evidence, and research gaps (Pham et al., 2014). Unlike systematic reviews which aim to answer specific research questions through a detailed appraisal of evidence, scoping reviews are designed to provide an overview of the available literature regardless of study quality.

The selection of a scoping review methodology is motivated by several key considerations of LLM applications in traffic and transportation research. First, the field encompasses a broad spectrum of research objectives and problem definitions, making a comprehensive systematic review challenging. Second, the diverse applications have led to highly specialized and customized implementations of LLMs. Third, the absence of standardized performance metrics and universal benchmarks complicates direct comparisons across studies, with some research utilizing traditional methods as benchmarks while others lack comparative frameworks entirely. Furthermore, the cross-disciplinary nature of the topic and its strong industry applications have resulted in research dissemination across various platforms, including academic journals, preprint archives, conference proceedings, and book chapters. Given these characteristics, a scoping review methodology enables us to effectively categorize this diverse body of literature into coherent subtopics, analyze methodological approaches, compile relevant statistics, and identify critical research gaps in the field.

**3.2.2.** Search Strategy A comprehensive literature search was conducted using the Web of Science database, selected for its extensive coverage of peer-reviewed journals, conference proceedings, and industry reports relevant to traffic and transportation research. The search was performed in November 2024, ensuring the inclusion of the most recent advancements and trends in the application of LLMs.

To identify relevant studies, a systematic search was conducted using the Web of Science database. The selection criteria were defined using the following Boolean string: TS=("Prompt Engineering" OR "Large Language Model" OR "Vision Language Model" OR "GPT") AND (TS=("Autonomous driving" OR "Scenario generation" OR "Safety" OR "Traffic" OR "Traffic Forecast" OR "Travel Behavior" OR "Traffic and Signal Control" OR "Pollution" OR "Transportation Emission" OR "Sustainable

Transport" OR "ITS" OR "Intelligent Transport System" OR "Shared Mobility" OR "Emergency Evacuation" OR "Emergency Response" OR "Traffic Simulation" OR "Pedestrian" OR "Modular Vehicle" OR "Vehicle" OR "Modality" OR "Traffic Accident" OR "Transport" OR "Logistics")). We do not restrict the publication period in our search.

3,122 studies are identified from our initial search using the Web of Science database. We further set the categories to only include "Transportation," "Industrial Engineering," "Management," "Engineering Multidisciplinary," and "Computer Science." This further reduces our paper number to 1,187. Afterward, we manually filter papers according to the following criteria:

Relevance: Studies must focus on the application of LLMs or related language models in traffic and transportation contexts.

**Publication Type**: Peer-reviewed journal articles, conference papers, good-quality pre-prints, and reputable industry reports.

Timeframe: Studies published up to November 2024 to capture the most recent advancements.

**Empirical Evidence**: Studies should provide empirical data or case studies demonstrating the application and effectiveness of LLMs in transportation tasks.

Applying those criteria results in 109 studies for subsequent detailed review. We acknowledge that despite our efforts to provide an exhaustive review through detailed search and selection criteria, by no means we can capture every relevant study in this field. Our methodology faces several limitations. Primarily, non-English publications may elude detection in databases like the Web of Science. Additionally, the use of specific jargon can obscure relevant studies; for example, researchers might use terms like "zero-shot learning" rather than "LLM" in their titles. Similarly, in the fields of traffic and transportation, terms like "routing" might be used instead of more general descriptors. The diversity of terminology across these disciplines can lead to omissions. To mitigate these issues, our approach includes leveraging the domain expertise of authors to identify pertinent papers and exploring tracing studies that are cited by the studies that we have selected to uncover further relevant literature.

## 3.3. Traffic Research

Traffic research refers to the dynamic flow and movement of vehicles, pedestrians, and other entities within a transportation network. It focuses on the operational aspects of mobility, including real-time interactions between road users, traffic flow optimization, and accident prevention. Unlike the broader study of transportation, which encompasses infrastructure development and policy, traffic research emphasizes the immediate and localized behavior of entities within existing systems.

In this section, we explore the role of LLMs in transforming traffic systems. Key areas of focus include autonomous driving, accident analysis, emergency management, traffic safety, signal control and traffic forecasting. These applications demonstrate how LLMs can enhance decision-making, improve system efficiency, and address complex challenges in modern traffic networks.

3.3.1. Autonomous Driving AD is a rapidly evolving field that aims to reduce accidents, enhance road safety, and improve mobility (Kuo et al., 2023). AD systems used to rely on rule-based and optimization-based methods, which provided reliable and interpretable results but struggled with complex, real-world scenarios (Yuan et al., 2024; Aksjonov and Kyrki, 2021; Guanetti et al., 2018; Dai et al., 2020). The advent of learning-based methods, such as RL, marked a significant improvement in handling complexity (Yan et al., 2022). However, these methods face challenges in interpretability and in managing the long tail case, i.e., rare, unpredictable events that traditional models often fail to address effectively.

The introduction of LLMs has enabled better reasoning, knowledge accumulation, and adaptability in AD, and is well-suited for addressing long-tail scenarios. Recent surveys (Yang et al., 2023; Li et al., 2024f; Cui et al., 2024b; Zhou et al., 2024) highlight their ability to perform a range of tasks, including perception, decision-making, and human-vehicle interaction. The evolution of methods is shown in Figure 4. Unlike conventional methods, LLMs excel at integrating multimodal inputs (e.g., visual, sensor, and textual data) and reasoning across diverse scenarios, bridging gaps in traditional AD systems. Key components include inputs, models, and tasks as summarized in Figure 5.

State-of-the-art AD frameworks increasingly incorporate LLMs through modular architectures. These systems typically include human instruction, perception, reasoning, reflection, and memory modules to ensure adaptability and safety. For instance, human instruction allows varying levels of intervention, while perception modules gather situational data. Reasoning modules then integrate this information with memory (past experiences) and scenario descriptions to guide decision-making. Reflection evaluates decisions, identifies unsafe outcomes, and updates the memory module for continuous improvement. Frameworks such as *DILU* (Wen et al., 2023), *DRIVELLM* (Cui et al., 2023), and *DriveGPT4* (Xu et al., 2024) exemplify this structure.

Our review focuses on understanding the integration of LLMs into four core components of AD – perception, decision-making, trajectory planning, and vehicle control. Each component plays a fundamental role in enabling AVs to navigate complex traffic environments safely and efficiently.

#### Perception

Perception is the foundation of AD, enabling vehicles to interpret their surroundings and make informed decisions. Perception systems use sensors like LiDAR, cameras, and radar to perform tasks such as object detection, environmental mapping, and situational awareness, ensuring safe operation in dynamic traffic environments.

Traditional rule-based and deep learning (DL)-based methods are effective in structured scenarios but struggle with real-world complexities. They often fail to handle rare or unpredictable events, lack contextual understanding (e.g., why pedestrians gather at a crosswalk), and require extensive retraining to adapt to new situations. LLMs have the ability to process multimodal data, which enables

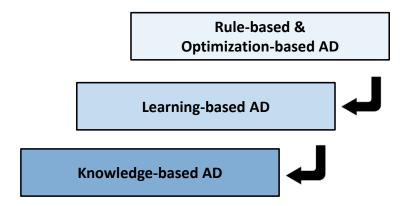


Figure 4 Advancements in Autonomous Driving

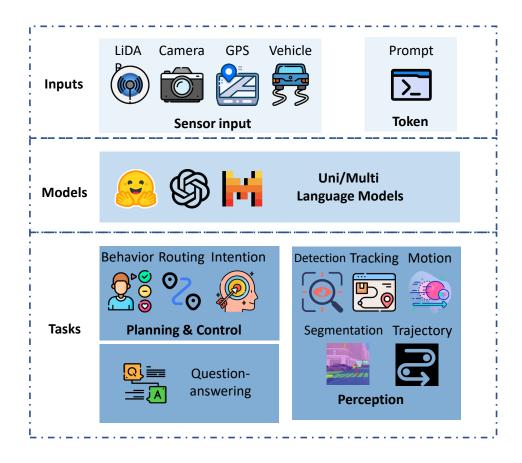


Figure 5 Autonomous Driving Framework (Yang et al., 2023)

a richer understanding of complex scenarios, better generalization, and adaptation with minimal retraining (Wang et al., 2023e). This makes LLMs uniquely suited to enhance perception capabilities in AD.

The first stream of the relevant literature in this field directly applies existing LVLMs with ZSL for recognizing rare and unexpected scenarios. Li et al., 2024b creates a benchmark named *CODA-LM* for assessing the quality of perception in AD with different large vision-language models. In their benchmark, there are three foundational aspects of perception tasks, including general perception, regional perception, and driving suggestions. Unlike prior works that rely on rigid, natural image-based datasets (e.g., SEED-Bench, MMBench), CODA-LM integrates a hierarchical multimodal framework, combining textual and visual data to assess models' reasoning and decision-making in dynamic driving environments. This proposed benchmark enables a fair comparison between current and future versions of LLMs in AD. The second paper in this area, Li et al., 2024e introduces the UnstrPrompt model, which leverages CLIP to independently process image and text features. By utilizing zero-shot semantic segmentation, the model accurately identified drivable areas without requiring paired datasets.

Despite these advancements, current LVLMs like GPT-4V excel in analyzing individual images but lack temporal reasoning and struggle with 3D spatial understanding. For instance, GPT-4V achieves only 50% accuracy in predicting vehicle behaviors, such as moving forward or backward, and performs poorly in identifying speeding vehicles (Sreeram et al., 2024; Wen et al., 2024). To address these limitations, Wang et al., 2024i propose OmniDrive, a 3D multimodal architecture which converts visual inputs into 3D representations, to significantly enhance spatial reasoning and decision-making. By incorporating 3D information, OmniDrive enables vehicles to better interpret spatial relationships and navigate complex environments, marking a crucial step toward more robust and intelligent perception systems.

The second stream of the literature focuses on understanding and predicting the motion of surrounding vehicles. Traditional methods often struggle to model the complex interactions between traffic actors. To address this gap, Lan et al., 2024 introduce Traj-LLM, a model designed to predict the future trajectories of dynamic traffic actors. Traj-LLM leverages observed past trajectories and semantics of surrounding vehicles to predict their future motion without the need for explicit prompt engineering. This approach marks a significant departure from heuristic methods, achieving 7% to 30% improvements over several trajectory prediction benchmarks.

#### Decision-Making using LLM

Decision-making is concerned with the selection of appropriate actions based on input from the perception system. This process entails analyzing the environment, predicting potential outcomes, and determining the optimal course of action while considering factors such as safety, legality, efficiency, and passenger comfort (Azarafza et al., 2024).

A major advancement brought by LLMs in decision-making is the ability to facilitate natural language interaction. Drivers can issue conversational commands, such as "merge into the left lane"

or "adjust speed for the vehicle ahead," and the LLM interprets these commands while analyzing perception data such as vehicle speed, localization, and in-cabin monitoring information. Based on this data, the system determines the feasibility of the action and communicates its reasoning back to the driver in natural language. This capability not only makes decision-making more intuitive for passengers but also enhances situational transparency by explaining why a particular action is safe or unsafe (Cui et al., 2023; Cui et al., 2024a; Xu et al., 2024). By integrating LLMs into the decision-making pipeline, autonomous systems are able to provide passengers with real-time feedback and contextual insights, fostering trust in their operations.

LLM-driven decision-making in autonomous vehicles can address both high-level and low-level actions, which represent two distinct streams of research. High-level actions involve strategic decisions such as left turns, right turns, merging, and lane changes. These actions are often managed through a combination of LLM-based reasoning and traditional rule-based systems, where the LLM provides contextual analysis and generates decisions that adhere to traffic rules and safety protocols (Cui et al., 2023). On the other hand, low-level actions focus on precise vehicle control, such as determining the turning angle or adjusting speed. These actions are approached through end-to-end systems that leverage LLMs trained on general multimodal datasets (e.g., CC3M, WebVid-2M), enabling real-time predictions of control signals directly from video inputs (Xu et al., 2024; Cui et al., 2024a; Wen et al., 2023). By addressing both streams, LLMs bridge the gap between strategic planning and real-time execution, ensuring that the vehicle can respond robustly to dynamic scenarios.

The Memory Module plays a pivotal role in enhancing LLM-driven decision-making by enabling the system to learn from past experiences and continuously improve. This module stores scene descriptions and reasoning processes, allowing the system to recall and apply solutions to long-tail problems—rare or complex scenarios that traditional systems struggle to address (Wen et al., 2023; Fu et al., 2024b). Additionally, the Memory Module records driver preferences, historical actions, maps, and laws, which facilitates personalized decision-making and ensures that the system adapts to individual user behaviors over time. For instance, a driver's typical preferences for merging speeds or following distances can influence how the system interprets and executes commands. Research shows that incorporating both successful experiences and corrected unsafe decisions into the Memory Module significantly enhances the system's ability to learn from mistakes, improving safety and adaptability (Wen et al., 2023; Fu et al., 2024b).

Another critical contribution of LLMs to decision-making is their ability to handle zero-shot reasoning, which enables them to manage novel or unfamiliar situations without prior exposure. By leveraging their broad training on diverse datasets, LLMs can reason through complex scenarios that involve unusual traffic patterns or unexpected obstacles (Cui et al., 2024a). For domain-specific challenges, FT the model with specialized datasets further ensures optimal performance in specific contexts, such as urban intersections or highways (Xu et al., 2024). This dual capability allows LLMs to adapt to dynamic environments while maintaining a high level of reliability.

Safety and interpretability are foundational to decision-making in autonomous driving, and LLMs address these aspects effectively. For instance, OpenAI's GPT-4 has been used as a safety-checking LLM, ensuring that passenger instructions comply with traffic rules and safety protocols (Cui et al., 2023). Beyond safety, LLMs enhance interpretability by providing clear, natural language explanations for their decisions. They articulate why certain actions – such as overtaking or merging – are feasible or not, considering factors like road conditions, traffic density, and vehicle speeds. This transparency is critical in building user trust, especially in safety-critical systems, as passengers can understand the reasoning behind the vehicle's actions (Xu et al., 2024; Fu et al., 2024b).

Despite their strengths, LLMs are not without limitations. A common issue across the literature is hallucination, where the model generates inaccurate or irrelevant outputs. This highlights the importance of robust safety mechanisms and extensive FT to mitigate such errors.

## Trajectory Planning using LLM

Trajectory planning enables vehicles not only to make decisions but also to execute them safely, efficiently, and comfortably. Traditionally, vehicle motion planners have relied on heuristic methods to plan driving trajectories. While these approaches are effective in familiar, structured environments, they often fail in novel or complex scenarios, where adaptability and generalization are crucial. This limitation has spurred a growing interest in leveraging LLMs to enhance trajectory planning.

One of the primary challenges in trajectory planning is generating accurate and adaptable trajectories for the ego-vehicle, which is the vehicle for which the sensors, decision-making algorithms, and driving actions are being studied or implemented. Existing heuristic methods often lack the flexibility to handle diverse and dynamic driving conditions. To address this, Mao et al., 2023 introduces GPT-Driver, an innovative approach that uses language descriptions of coordinate positions to generate driving trajectories. By framing trajectory generation as a language modeling task, GPT-Driver eliminates the reliance on handcrafted heuristics, offering a more generalizable solution for ego-vehicle trajectory planning.

However, trajectory planning does not exist in isolation. It is part of a larger closed-loop system where real-world decisions are continuously refined based on feedback. Recognizing this, Fu et al., 2024a propose LimSim++, a closed-loop simulation platform for deploying multimodal large language models (MLLMs) in AD. LimSim++ takes ego-vehicle trajectory planning to the next level by converting decisions from MLLMs into trajectories and assessing their performance through a built-in evaluation module. The platform introduces a reflection and memory module to iteratively improve decision-making accuracy, enabling MLLMs to learn from past mistakes. This refined reasoning is then used as few-shot instances to further enhance future decision-making, creating a feedback loop that traditional heuristic methods cannot achieve.

Building on the idea of modeling interactions, Xia et al., 2024 propose InteractTraj, a method specifically designed to generate interactive traffic trajectories. Unlike Traj-LLM, which focuses on individual actor predictions, InteractTraj emphasizes the interactions between multiple traffic actors, capturing the dynamic interplay that occurs in real-world driving scenarios. This approach leads to a 16% improvement over baseline methods, showcasing the importance of interaction modeling in trajectory prediction.

The future of trajectory planning lies in the ability to generalize across diverse scenarios, anticipate interactions with surrounding vehicles, and continuously improve through feedback. The integration of language models, as demonstrated by these works, offers a promising path forward for AD innovation.

## Vehicle Control using LLM

Vehicle control is the critical step in autonomous driving, translating high-level decisions into physical actions such as steering, acceleration, and braking. These actions must be precise and smooth to ensure safety and user comfort. However, traditional control systems often rely on rigid, rule-based methods, which can struggle with dynamic or complex scenarios, leading to erratic or uncomfortable vehicle behavior. This challenge has become particularly relevant as LLM-controlled vehicles emerge, where the accuracy of decision-making directly impacts the smoothness and safety of vehicle maneuvers. Recent research has begun to explore the potential of LLMs to address these challenges by introducing more adaptable, human-aligned control systems.

To align vehicle control with human driving preferences, researchers have turned to reinforcement learning from human feedback (RLHF) as a key strategy. RLHF enables LLMs to adapt their outputs based on human evaluation, allowing them to fine-tune steering, speed, and other control parameters in a way that mimics human behavior. Building on this, Cui et al., 2024a leverages GPT-4 to incorporate direct human input into vehicle control. By integrating data from perception, localization, and incabin monitoring, their system refines decision-making accuracy in real time, improving adaptability and creating a smoother, more personalized driving experience.

In addition to aligning with human preferences, efforts are also made toward enabling LLMs to perform end-to-end driving tasks with minimal or no human intervention. Traditional vehicle control systems use modular pipelines where perception, planning, and control are separate components, which often leads to inefficiencies or errors in decision-making. In contrast, LLMs offer the potential to unify these processes by combining reasoning and perception into a single framework. For example, Chen et al., 2024c propose a driving question-answering framework that uses LLMs to generate accurate control actions. By framing control tasks as reasoning problems, their system determines appropriate vehicle maneuvers while maintaining transparency, allowing users to understand the rationale behind each action. This marks a significant step toward autonomous and explainable vehicle control systems.

By combining human feedback, real-time adaptability, and reasoning-driven autonomy, these advancements pave the way for safer, more efficient, and explainable autonomous driving systems. A full list of literature surveyed is presented in Table 3.

3.3.2. LLMs in Travel Behavior and Mobility Prediction Travel behavior is a broad concept encompassing many attributes, including mode choice, travel purpose, destination selection, departure time, route choice, and travel frequency. Understanding these factors enables agencies to enhance transit services and forecast short-term travel conditions while informing long-term transportation infrastructure investments, policy development, and land use planning. For individuals, mode choice predictions integrated into map applications support informed and efficient travel decisions.

Traditional methods like decision trees, support vector machines, and regression models rely on predefined features and structured datasets (Yin et al., 2024a; Koushik et al., 2020). While effective in certain contexts, these methods struggle with the complexity of human travel behavior, especially when dealing with contextual relationships or unstructured data like text and images. They often require extensive feature engineering and lack generalizability in dynamic environments. In contrast, LLMs introduce a shift to travel behavior prediction by integrating both structured and unstructured data. Their ability to process multimodal information and learn contextual relationships enables more precise insights and predictions. LLMs excel in areas such as mode choice prediction, human mobility modeling, and analyzing pedestrian and driver behavior, addressing the limitations of traditional approaches while unlocking new possibilities for intelligent transportation systems.

## **Human Mobility Prediction**

Realistic human mobility prediction serves as key components of many research fields, including transportation, disaster management, and urban planning. Traditional methods of human mobility prediction rely primarily on mathematical modeling, statistical analysis, and classical Machine Learning approaches. These methods are rooted in well-established theories of travel behavior and decision-making, often focusing on structured data inputs like socioeconomic characteristics, travel times, and costs. Among them, the most widely studied methods include gravity model (Pappalardo et al., 2016), discrete choice model (Bierlaire, 1998), activity-based models (Bowman and Ben-Akiva, 2001), agent-based models (Maggi and Vallino, 2016) and machine learning (ML) methods (Toch et al., 2019). These methods for predicting human mobility are based on structured data, extensive feature engineering, and domain-specific assumptions. However, as urban mobility becomes more dynamic and diverse, these methods struggle to handle complex semantic information and unstructured inputs, such as social media updates and sensor-generated textual description, highlighting the need for more adaptable approaches.

Table 3 LLMs in Autonomous Driving

Literature	Model Name	Model Backbone	Input	Output	Modality	Task
Azarafza et al., 2024	-	GPT-4	Road images	Driving action	Image	Driving decision making
Chen et al., 2024c	Driving with LLMs	GPT-3.5, LLaMA-7b	Driving scenario description, object	Driving action	Text	Driving decision making
Cui et al., 2023	DriveLLM	GPT-3.5	Image, location, vehicle state, weather, and experience	Driving action	Text and image	Perception and decision making
Cui et al., 2023 Cui et al., 2024a	-	GPT-4 GPT-4	- Driver command, perception	- Trajectory planning	- Image and sensor data	Prediction, decision making and motion planning
Fu et al., 2024a	${\rm LimSim}{+}{+}$	GPT-3.5, GPT-4, GPT-4V	Road image and video	Trajectory	Image and video	Trajectory planning
Fu et al., 2024b	Drive Like a Human	GPT-3.5	Road environment, vehicle state	Driving decisions	Text	Driving decision making with interpretation and self-reflection
Hu et al., 2023	-	ChatGLM, CLIP, BLIP-2	Road images, human instruction	Navigation reasoning	Image and text	Vehicle navigation
Hussien et al., 2025	GPT-4	Pedestrian and vehicle motion information	Predicted pedestrian actions and reason	Text and data	Behavior understanding of road users	JAAD, PSI and highD
Lan et al., 2024	Traj-llm	LLM	Historical vehicle trajectories, lane information, surrounding information	Trajectory	Spatial-temporal state data	Trajectory prediction
Li et al., 2024d	MiningLLM	GPT-4	Road images	Mining operation	Image and text	Collaborative operations for mining and safety assessment
Li et al., 2024b	-	LLaVA-Llama-3- 8B-v1.1	Vehicle and road information	Predictions of road entities, descriptions of corner cases, and driving suggestions	Image and text	Corner case detection in autonomous driving
Li et al., 2024e	GPT-3.5 with CLIP	Road images and texts describing scene	Segmentation masks with improved accuracy	Image and text	Semantic segmentation and language-guided perception	Talk2Car dataset
Mao et al., 2023	GPT-driver	GPT-3.5	Object images, ego-states, historical trajectories	Predicted trajectories	Text and data	Motion planning, safety compliance, interpreting complex driving scenario and reasoning
Pan et al., 2024a	VLP	-	Images of other vehicles	Action prediction of other vehicles	Image and text	Perception, prediction and motion planning
Sreeram et al., 2024	-	GPT-4V	Sequence of vehicle images	Predicted action of other vehicles	Image, text	Scene understanding
Tanahashi et al., 2023	-	LLaMA 2, GPT 3.5, GPT 4	Road images	Driving decisions	Image and text	Spatial recognition and traffic rules compliance in decision making
Tian et al., 2023	VistaGPT	-	Road images, vehicular states and navigation command	Driving task, waypoints	Image and scalar data	End-to-end autonomous driving system composition and optimization
Wang et al., 2023c	-	GPT-3.5	Road images and human intention	Driving commands (e.g., steering) and trajectories	Image and text	Co-pilot
Wang et al., 2023e	-	GPT-4	Scene description and sensor data	driving task and explanation	Text	enhance safety and interpretability in autonomous driving
Wang et al., 2024g	BEVGPT	GPT	Bird eye view image	Driving decisions	Image	Prediction, decision making and motion planning
Wang et al., 2024j	Drive as veteran	LLaMA-7B	Scenario description	Driving tasks	Text	Driving task
Wang et al., 2024i	OmniDrive	GPT-4	Multiview image	Decision making, planning	Image	Action prediction
Wen et al., 2023	DiLu	GPT-3.5 and GPT4.0	System prompts, textual description, and few-shot experience	Driving action	Text	Decision making
Wen et al., 2024 Xia et al., 2024	- InteractTraj	GPT-4V GPT-4	Image Scenario	Driving action Predicted	Image Text and data	Action prediction Trajectory
Xu et al., 2024	Drivegpt4	GPT-4	description Road videos	trajectories Driving decisions	Video	prediction Driving decision making

To address these challenges, researchers have applied NLP techniques to process unstructured data like textual descriptions and contextual information. NLP models convert text into tokens, enabling the integration of numerical and textual data for mobility prediction. For example, the "SHIFT" framework Xue et al., 2021 incorporates contextual information such as location categories, weather, and temporal data into a unified natural language model, outperforming traditional methods like ARIMA and LSTMs in accuracy and flexibility. Similarly, a BERT-based Masked Language Model (Yang et al., 2024b) is used to predict transportation modes using two million trip records from major Chinese cities, demonstrating its superiority over the traditional MNL approach.

Despite their successes, NLP methods like SHIFT and BERT-based models rely on predefined templates or rigid frameworks for mobility-to-language transformation, limiting their adaptability to novel scenarios or irregular patterns. To overcome these limitations, researchers have increasingly turned to LLMs, which offer greater flexibility, scalability, and reasoning capabilities. Through our literature search of LLMs applied in this area, two streams of research are identified. The first stream explores the potential of using LLMs to process and generate semantic-rich information on mobility patterns similar to the study of NLP. The second stream applies LLMs to make explainable trajectory modeling at the individual level. In both application streams, semantic reasoning, few-shot learning and explanability capabilities of LLMs are used.

In the first stream of research, scholars have applied LLMs to make travel mode choice analysis and predictions. Unlike NLP-based methods that rely on manual annotation or fixed models for specific objectives, LLMs allow for automated inference of travel modes, sentiment analysis, and summarization of reasons behind sentiments in a unified framework. A study by Ruan et al., 2024 has explored the use of LLMs to analyze social media data from tweets for summarization of reasons behind their travel mode choices and has shown good performance.

Additional studies have looked into using LLMs for travel mode prediction. By leveraging prompt engineering, LLMs can transform travel attributes—such as time, cost, and individual preferences—into textual inputs, enabling zero-shot predictions without requiring training data. Mo et al., 2023b demonstrate this approach, showing that LLMs could achieve competitive accuracy compared to traditional MNL and random forest models. However, this study also highlights challenges including reasoning errors and hallucinations, emphasizes the need for improved prompt design. Building on this, Liu et al., 2024d introduce a few-shot learning approach, where a small number of training samples were used to guide the model's understanding of bounded rationality in human decision-making. This approach significantly improves the alignment between model predictions and observed behaviors, underscoring the importance of contextual learning in mobility analysis.

Traditional methods like statistical models, ML, and survey-based approaches often struggle with sparse, imbalanced datasets and fail to adequately capture passenger heterogeneity or the complexities of travel behaviors during delays. To address these challenges, DelayPTC-LLM (Chen et al.,

2024a) leverages advanced reasoning, NSL, and NLP capabilities to analyze delay logs and passenger behavioral data. By integrating prompt engineering and the CoT strategy, DelayPTC-LLM not only predicts travel choices but also explains its reasoning process. Experimental results show that DelayPTC-LLM outperforms traditional models and even standalone LLMs without CoT.

The second stream of research focuses on generating personal mobility trajectories by leveraging historical data and psychological factors, such as attitudes, subjective norms, and perceived behavioral control. For instance, the Chain-of-Planned-Behaviour (CoPB) framework incorporates psychological insights to enable step-by-step reasoning for mobility intention generation, significantly reducing error rates and token consumption compared to pure LLM methods (Shao et al., 2024). Similarly, the LLMob framework integrates self-consistency evaluation to align trajectories with historical data and employs retrieval-augmented strategies to infer motivations from temporal and contextual information (Wang et al., 2024b). LLMob not only generates semantically rich activity patterns but also outperforms state-of-the-art models, such as DeepMove and DiffTraj, by adapting effectively to external factors like pandemics.

The application of LLMs in mobility prediction marks an evolution in the field. Unlike traditional methods, which rely on structured data and domain-specific assumptions, LLMs excel in processing unstructured inputs, integrating diverse contexts, and generating interpretable outputs. Their semantic reasoning and FSL capabilities enable them to adapt to new contexts, making them highly versatile for tasks ranging from travel mode choice prediction to trajectory generation.

# **Pedestrian Motion Prediction**

Pedestrians are the most vulnerable road users. Every day, people walk through urban environments, sharing spaces with motor vehicles at all times and conditions. These interactions expose pedestrians to significant traffic risks. Thus, understanding human behavior is crucial to creating safe and intelligent transportation systems in dynamic urban settings. Traditional methods like rule-based systems and statistical models have been used to study pedestrian and driver behavior, but often fail to capture the complexity and variability of real-world scenarios. Innovative solutions to accurately model complex pedestrian-vehicle interactions are desired to enhance pedestrian safety.

NLP techniques have been used for structured textual data analysis in pedestrian safety-related studies. For example, models like BERT have successfully classified pedestrian maneuver types (Das et al., 2023), showcasing the effectiveness of pre-trained transformers. While NLP provides a foundation for integrating language models into transportation research, LLMs surpass NLP with broader reasoning and adaptability. In pedestrian behavior modeling, LLMs enable more dynamic and realistic simulations. Traditional trajectory planning models, such as kinematic simulations, have often failed to capture the variability of real human movement. Ramesh and Flohr, 2024 demonstrate how the

Flan-T5 model could simulate realistic pedestrian movements by leveraging the reasoning capabilities of LLMs to generate human-like movements. This approach provides a more flexible and reliable foundation for safety assessments and autonomous system development, representing a step forward from static, rule-based simulations.

## **Driver Behavior Analysis**

Driver behavior is a crucial factor in both driving safety and efficiency. From a safety perspective, modeling elements such as driver fatigue and braking performance is essential for reducing risk. In terms of efficiency, decisions regarding cruise speed, lane-changing, and car-following are critical. A better understanding and accurate modeling of these behaviors not only enhances safety and efficiency for human drivers but also strengthens the capabilities of AD systems. In traditional AD, behavior decision-making is typically based on rule-based methods. However, these methods cannot deal with long-tail cases, such as irregular driving behavior. The advent of LLMs shows the potential to address the challenges with powerful generalization and commonsense reasoning abilities, enabling them to infer information from previously unseen scenarios. Chen et al., 2024e address these limitations with the GenFollower model, a prompt-based LLM approach designed to predict car-following behaviors while maintaining interpretability – an essential feature for practical applications. Similarly, Zhang et al., 2024a develop a visual large language model that combines visual and textual data through an FSL approach to analyze distracted driving patterns. This multimodal integration showcases the versatility of LLMs in combining contextual and visual information to tackle safety-critical challenges.

In summary, LLMs have demonstrated their potential applications in pedestrian and driver behavior modeling by enabling realistic pedestrian trajectory simulations and integrating diverse data types to analyze driver behaviors. A full list of the literature surveyed is presented in Table 4.

3.3.3. Traffic Safety Traffic accidents remain a global challenge, causing approximately 1.2 million deaths and 20-50 million severe injuries annually. Traditional approaches in traffic safety rely on statistical methods for crash frequency and severity analysis, offering essential insights but with limitations. They depend heavily on structured datasets like crash statistics and vehicle counts, leaving unstructured data sources – such as police reports, incident logs, and social media – largely untapped. These methods lack the ability to capture nuanced, context-specific information and often fail to model complex interactions among factors like driver behavior, road conditions, and environmental influences, resulting in incomplete understanding and less effective preventive measures (Xu et al., 2025).

NLP has emerged as a tool to address these challenges by converting unstructured text into structured formats, enabling more comprehensive analysis. For instance, NLP has been used to transform crash reports into spatial data for more accurate accident location detection (Wang et al., 2017). Social media updates, analyzed using NLP techniques, have also been employed for real-time traffic

Table 4 Travel Behavior Prediction Using LLMs

	Table 4 Travel Deliavior Frediction Osing Leivis								
Literature	Model Backbone	Input	Output	Modality	Task	Data Source			
Chen et al., 2024d	ChatGPT	Student prompts	Explanations, code	Text	Inquiry-based learning	Interaction logs			
Chen et al., 2024e	GPT-4	Natural language prompts	Car-following predictions	Text	Driver behavior analysis	None			
Chen et al., 2024a	GPT-4	Event logs and passenger travel choice datasets	Predicted passenger travel choices during metro delays	Text	Passenger travel choice prediction	Shenzhen Metro AFC data			
Liang et al., 2024	GPT-4	Event descriptions and mobility data	Mobility patterns	Text	Human mobility prediction	Public events data			
Liu et al., 2024d	GPT-3.5/GPT-4	Designed prompts	Predictions, reasoning	Text	Behavioral prediction	Translated dataset variables			
Mo et al., 2023a	GPT-3.5	Contextual prompts	Mobility predictions	Text	Human mobility prediction	None			
Ramesh and Flohr, 2024	Flan-T5-Base	Text descriptions	Pedestrian motion trajectories	Text	Pedestrian behavior simulation	Simulation datasets			
Ruan et al., 2024	$\begin{array}{c} \mathrm{GPT}\text{-} \\ 3.5/\mathrm{Llama2} \end{array}$	Social media data	Travel mode predictions	Text	Travel mode prediction	Social media platforms			
Shao et al., 2024	LLaMA3-8B	Intention generation prompts	High-quality predictions	Text	Intention generation	Tencent and China Mobile data			
Wang et al., 2024b	GPT-3.5- turbo	Historical data and contextual prompts	Mobility activities	Text	Human mobility prediction	None			
Zhang et al., 2024a	LLaMA	Images and text cues	Driver distraction classification	Image and text	Driver behavior analysis	Driver images			
Zhao et al., 2024b	LLaVA	Image-text pairs	Driving behavior strategies	Image and text	Driver decision- making	nuScenes dataset			

prediction (Sampath and Supriya, 2023). By bridging structured and unstructured data, NLP enhances ITS, enabling more responsive traffic management (Ali et al., 2021; Wan et al., 2020). LLMs build on NLP's foundation, offering unique advantages for accident analysis, multimodal integration, and proactive traffic prediction, as explored in the following subsections.

## **Accident Analysis**

While NLP systems have provided some progress through the extraction of structured insights from textual data, their reliance on predefined rules and limited adaptability have constrained their effectiveness. LLMs change how accident analysis is conducted by automating the extraction and interpretation of unstructured textual data, such as accident reports and police logs. For instance,

models like TrafficSafetyGPT have shown the ability to process vast volumes of textual accident data, extracting critical details such as causes, severity, and contributing factors with high accuracy (Zheng et al., 2023b). By FT general-purpose LLMs with domain-specific knowledge, such as government guidelines and historical crash data, these models streamline the process of data extraction and summarization, significantly reducing human effort and the potential for error. This capability enables researchers and policymakers to uncover deeper insights into accident causes and mechanisms, which were previously inaccessible with traditional methods or simpler NLP systems.

Beyond data extraction, LLMs enable the generation of synthetic scenarios to address the challenges posed by rare, high-impact accidents. Scenario engineering, a novel application of LLMs, allows researchers to simulate diverse crash scenarios, including edge cases such as vehicle trajectory deviations or lane-change conflicts, which are rarely captured in real-world datasets (Chang et al., 2024). By augmenting existing datasets with these synthetic scenarios, LLMs enhance the robustness of safety research and enable the testing of preventive measures under a wider range of conditions.

LLMs also excel in multidimensional accident analysis by integrating textual and visual inputs. VLMs, which combine the capabilities of LLMs with computer vision, have been applied to analyze crash scenes in unprecedented detail. For example, recent studies have demonstrated how LLMs, paired with techniques like segment extraction and dynamic prompts, can generate granular descriptions of accident contexts, including pedestrian behavior, vehicle dynamics, and environmental conditions (Xuan et al., 2024). Similarly, V2X-perception architectures integrate data from panoramic multi-camera views and road devices to construct a comprehensive understanding of traffic environments for accidents (Wang et al., 2023b). These advancements enable a level of detail and context in accident analysis that far exceeds the capabilities of traditional methods or stand-alone NLP systems.

# Proactive Traffic Prediction and Safety Management

LLMs have demonstrated a potential in predicting high-risk traffic conditions and accident hotspots. By analyzing historical accident records, real-time traffic updates, and environmental data, these models can identify patterns and correlations that elude traditional methods. For example, GPT-based models have been applied to predict accident types and traffic flow disruptions with remarkable precision, enabling traffic authorities to implement proactive measures such as adjusting traffic signals or deploying resources to high-risk areas (Bäumler and Prokop, 2024; Wang, 2024). These predictions are not only more accurate than those generated by traditional methods but also provide actionable insights for preemptive traffic management.

Furthermore, LLMs enhance traffic crash response planning by prioritizing high-risk scenarios and improving communication with first responders. Techniques such as CoT reasoning and prompt engineering (PE) enable LLMs to analyze crash scenarios and severity outcomes in real time, providing detailed recommendations for response strategies (Zhen et al., 2024). This capability improves the

speed and effectiveness of accident responses, minimizing the impact of accidents and saving lives (Fan et al., 2024).

In summary, LLMs can provide deeper insights into traffic accidents, have real-time capabilities, and facilitate proactive interventions. LLMs are paving the way for safer and more efficient road systems. A full list of the literature surveyed is presented in Table 5.

**3.3.4.** Emergency Management Transportation systems heavily rely on precise, real-time information for effective disaster response. However, there's a notable gap in quickly obtaining detailed disaster information, such as the extent and location of an event. Traditional tools like remote sensing lack the needed detail. For social media platforms, while they can provide instant data, they often contain excessive irrelevant information. Moreover, traditional data analysis methods struggle to address the complex, multi-dimensional nature of disasters.

LLMs offer a solution to address the limitations of some traditional approaches in transportation emergency management. Unlike traditional models, LLMs excel in context-sensitive reasoning, multi-modal data integration, and the real-time extraction of actionable information. By leveraging techniques like CoT reasoning and prompt engineering, LLMs can provide transparent, step-by-step analysis and adapt to specific emergency management tasks. In this section, we introduce three streams of research identified in our literature search, namely real-time information processing and decision support, communication and overcoming language barriers, and risk assessment and prediction.

#### Real-Time Information Processing and Decision Support

Effective disaster response hinges on the ability to process and analyze large amounts of fragmented, real-time information – something human operators often struggle with. LLMs address this gap by synthesizing unstructured data from diverse sources, such as social media posts, news reports, and emergency calls. These sources may contain critical details like the location and type of emergency, which LLMs can extract to provide actionable instructions for dispatchers (Otal and Canbaz, 2024).

During emergencies, information is noisy and evolves rapidly. LLMs help filter irrelevant data, prioritize critical updates, and create a comprehensive picture of the crisis. This capability enhances collaboration between human responders and AI systems, enabling faster resource mobilization and reducing response delays.

Despite their strengths, LLMs are prone to hallucination – an unacceptable limitation in high-stakes scenarios. To address this, frameworks like E-KELL integrate knowledge graphs with LLMs, structuring verified emergency-related data for improved reliability. Using a prompt chain mechanism, E-KELL guides LLMs step-by-step through logical reasoning, ensuring outputs align with regulations and factual data (Chen et al., 2023). This structured approach significantly reduces hallucinations, making LLMs more dependable for emergency management.

Table 5 Travel Safety Analysis Using LLMs

		Table 5	ravel Safety Analy	SIS USING LLIVIS		
Literature	Model Backbone	Input	Output	Modality	Task	Data Source
Zheng et al., 2023b	GPT (version unknown)	Textual accident reports	Accident-based responses	Text	Multisensory safety analysis	CA DMV, NHTSA, OSM
Chang et al., 2024	GPT-4	Scenario tokens, instruction tokens, experience tokens, etc. from naturalistic driving datasets	Generated driving scenarios (vehicle trajectories, lane changes, and risk analysis)	Multimodal (vehicle trajectories, semantic un- derstanding, and text)	Scenario generation for risky driving conditions	HighD dataset
Xuan et al., 2024	Qwen-VL	Traffic video segments	Traffic safety descriptions (pedestrians, vehicle behavior, road conditions, contextual information)	Text and video	Traffic safety description and analysis	Woven Traffic Safety (WTS) and BDD100K datasets
Wang et al., 2023b	GPT-4V (in V2X)	Mutli-camera images	3D object detection, BEV maps, trajecotry detection, and safety assessments	Text and video	Accident analysis and prevention	DeepAccident dataset
Bäumler and Prokop, 2024	BERT	Textual accident descriptions	Accident type	Text	Classify traffic accidents	German federal states accident dataset
Wang, 2024	GPT-4	Language queries related to traffic conditions	Traffic advisory reports	Text	Provide traffic advisories	Inductive loop detectors
Zhen et al., 2024	GPT-3.5- turbo	Structured tabular data on various traffic crash attributes	Predictions on crash severity	Text and data	Crash severity inference	CrashStats dataset
Fan et al., 2024	LLaMA-2	Crash events dataset (general information, infrastructure details, event descriptions, and unit information)	Injury and severity predictions	Text, data and image	Injury and severity predictions and classification	Highway Safety Information System (HSIS)

# Communication and Overcoming Language Barriers

Language diversity in multicultural and multilingual communities poses a significant challenge during emergencies. Traditional solutions, such as pre-trained translation models or static multilingual

databases, lack the flexibility to handle nuanced, real-time interactions. LLMs fill this gap with advanced multilingual support and natural language understanding capabilities.

For example, LLMs can translate emergency calls in real time, generate follow-up questions for dispatchers, and adapt messages to the linguistic and cultural context of the affected population (Jiang, 2024; Otal and Canbaz, 2024). As such, LLMs act as "virtual institutional memories," learning from past interactions to improve future responses. Unlike rigid translation models, LLMs dynamically adjust to evolving communication needs, ensuring that critical information reaches all individuals during a crisis.

### Risk Assessment and Prediction

Proactive risk assessment is essential for mitigating the impact of disasters, yet traditional methods relying on static models or historical data often fail to account for real-time changes or integrate diverse datasets. LLMs excel in identifying patterns, analyzing past emergencies, and predicting emerging risks by processing data from sources such as scientific reports, weather forecasts, and community feedback (Jiang, 2024).

For instance, LLMs can dynamically integrate real-time meteorological data with historical patterns to detect evolving risks during natural disasters. This enables emergency managers to anticipate threats and implement mitigation strategies, thereby reducing the disaster impacts. LLMs outperform traditional systems by effectively synthesizing diverse data streams, which are often siloed in conventional approaches (Jiang, 2024).

In summary, LLMs assist emergency response by offering dynamic and intelligent solutions with their powerful ability of multi-source summarization and analysis. A full list of the selected literature is presented in Table 6.

3.3.5. Traffic Forecasting Traffic forecasting is essential for improving transportation systems by enabling better traffic management, reducing congestion, and optimizing infrastructure efficiency (Liu and Meidani, 2024a; Ma et al., 2015). Accurate predictions of traffic support informed decision-making for both transportation agencies and users, whether for route planning or emergency response (Liu and Meidani, 2024b). Key metrics such as traffic volume, speed, travel time, and congestion levels can all benefit from the insights provided by traffic data. The application of NLP and LLMs in traffic flow forecasting helps to incorporate textual data and predict spatial-temporal traffic flow series.

Recent advances have seen the integration of NLP into traffic flow prediction by extracting insights from textual data, such as social media and news reports, to complement traditional traffic and weather datasets. For instance, Jin et al., 2021 integrates large-scale textural road information and weather information with BERT to predict long-term traffic flow. Furthermore, Essien et al., 2021 and Yan et al., 2025 demonstrate how combining tweet-based information with traffic and weather data

Table 6 Emergency Management Using LLMs

Literature	Model Backbone	Input	Output	Modality	Task	Data Source
Chen et al., 2023	ChatGLM-6b and GPT-3.5	Structured queries related to emergency management scenarios	Precise, actionable guidance for decision making	Text	Emergency decision support	Official data in China
Jiang, 2024	GPT-3.5	Textual data from diverse sources (historic, scientific reports, etc.)	Processed insights, predictions, etc.	Text	Real-time information processing	Historic emergency data, scientific literature, real-time situational updates, etc.
Otal and Canbaz, 2024	LLaMA2	Emergency call data, social media messages, etc.	Classification and interpretation of emergency data	Text and voice	Classification of emergency	Turkey's Emergency- Disaster Messages Dataset
Yin et al., 2024b	LLaMA-2	Disaster- related social media posts	Multi-label classifications	Text	Extract critical information from disaster-related posts	CrisisBench dataset

improves prediction accuracy. Tsai et al., 2022 incorporate social media features into long-term traffic forecasting models. These approaches highlight the potential of NLP to enhance traffic predictions by providing richer, context-aware datasets that address real-world complexities.

While LLMs are generally not designed as a tool for forecasting purposes, the existing research has seen applications in traffic flow prediction as well as reasoning of traffic flow, as described below.

#### **Traffic Flow Prediction**

Traffic flow prediction models aim to forecast traffic volumes and speeds across spatial networks and varying time horizons. Compared to NLP models, LLMs provide a unique advantage as they inherently encode real-world knowledge (e.g., "heavy rain increases taxi demand"), enabling them to generalize to rare or unseen events (e.g., concerts and extreme weather) that lack historical data. LLMs are fed with text information, and their final hidden layer outputs (or pooled outputs) are used as embeddings. These embeddings are fused with historical traffic data and fed into traditional spatiotemporal models (e.g., STGCN and diffusion convolutional recurrent neural network (DCRNN)). Using this method, Huang, 2024 demonstrate how LLMs could represent text-based regional information as nodes in a traffic network, enriching the spatial-temporal embeddings and improving the prediction accuracy.

Low-rank adaptation (LoRA) fine-tuning techniques are also used to make LLMs more efficient for traffic flow prediction, even with limited labeled data. Ren et al., 2024 demonstrate that LoRA FT preserves pre-trained knowledge while enhancing the model's ability to extract temporal and spatial patterns, ensuring high prediction accuracy with minimal computational overhead.

Literature Model Input Output Modality Task **Data Source** Backbone LLaMA2 Traffic flow Guo et al., Traffic flow Explanation of Text and LargeST 2024series + graphlink traffic flow data reasoning dataset, OpenStreet, NOAA Gebre et al., GPT-4 Traffic flow Explanation of Text and Traffic flow NGSIM 2024density + link traffic data reasoning dataset prompt density Huang, 2024 GPT (version Traffic flow Future traffic Text and Traffic flow NYC bike series + flow data prediction share data not mentioned) semantic information Liu et al., GPT-2, Traffic flow NYCTaxi, Future traffic Text and Traffic flow 2024bLLaMA2 CHBike series + graph flow data prediction Ren et al., GPT-2 Traffic flow Future traffic Text and Traffic flow PeMS dataset 2024 series + Graph data prediction GPT-4 Network-wide Text and Traffic flow Network-wide Wang et al., SQL queries + interpretations 2024amobility info + data reasoning mobility database prompt GPT-4 & Prompts for Python code Texts Generate codes MATSim-Ying et al., 2024Phi-3-mini writing codes for congestion generated pricing synthetic computation travel data Zhang et al., **GPT-3.5** Missing traffic Imputed traffic Text and Traffic data PeMS dataset 2024bdata query data data imputation query

Table 7 Traffic Forecasting

## Traffic Flow Reasoning

Beyond prediction and imputation, LLMs excel in reasoning and providing actionable insights for traffic management. Traditional models often focus solely on forecasting traffic conditions, leaving the interpretation of results and decision-making to human experts. LLMs, however, go a step further by enabling contextual understanding, root cause analysis, and actionable recommendations for alleviating traffic issues.

For instance, large-scale traffic flow reasoning can benefit from the interpretive capabilities of LLMs. These models can process human prompts to generate SQL queries, interpret traffic flow predictions, and explain the underlying causes of congestion or bottlenecks. Using FSL and CoT prompts, Guo et al., 2024 and Wang et al., 2024a demonstrate how LLMs can align traffic flow predictions with natural language explanations, providing interpretability and enhancing trust in model outputs.

Similar approaches can also be applied for human mobility analysis, addressing challenges like integrating unstructured textual data (e.g., concert details or artist popularity) with historical human mobility patterns. By employing CoT prompting, these models generate step-by-step reasoning for human mobility predictions, with improved transparency and interpretability in the case study of public events (Liang et al., 2023).

Another advancement involves interacting physics-informed models with LLMs, allowing users to query specific traffic conditions and receive human-like explanations. Gebre et al., 2024 use GPT-4 to combine traffic flow reasoning with physical models, enabling more accurate and explainable analyses of traffic density and flow patterns.

In summary, LLMs advance real-time traffic reasoning and management, by playing three key roles: (i) interpreting human prompts and SQL queries or code to extract relevant data; (ii) extracting temporal and semantic information to capture the correlation between traffic flows; and (iii) interpreting the traffic flow prediction results and suggesting ways to alleviate bottlenecks. Apart from the aforementioned literature, two additional studies are worth noting. Zhang et al., 2024b study the use of an LLM for querying the traffic speed imputation system. Ying et al., 2024 and Zhang et al., 2024b study the decision support aspect of congestion pricing with LLMs. They are not included in the above review as we consider their applications of LLMs indirectly related to the research topic. The key studies in this field are summarized in Table 7.

3.3.6. Traffic Signal Control Traditional adaptive traffic signal control (TSC) struggles with flexibility and generalization, particularly in unfamiliar or dynamic scenarios, and thus fails to adapt effectively. For example, rule-based systems, while effective under stable conditions, are not capable of adequately handling sudden changes, such as emergency vehicle arrivals or sensor failures (Pang et al., 2024b; Wang et al., 2024c). RL-based methods, although capable of learning from real-time data, can suffer from overfitting to specific conditions or a lack of flexibility when faced with unforeseen circumstances (Gregurić et al., 2020; Chu et al., 2021).

To overcome these limitations, integrating LLMs like GPT-4 into TSC frameworks is becoming increasingly prevalent. These models enhance decision-making processes by leveraging their extensive knowledge and reasoning capabilities (Pang et al., 2024b; Wang et al., 2024c). For instance, the iLLM-TSC (Integration of RL and LLM for TSC) framework combines RL's capacity for learning traffic control policies from real-time data and making decisions with the LLM's ability to evaluate these decisions to verify their reasonableness (Pang et al., 2024b). This integration not only helps in filling the gaps left by RL models – such as missing state information or unconsidered events – but also increases the robustness of TSC systems under diverse conditions like communication degradation (Pang et al., 2024b; Villarreal et al., 2023).

Another focus in the field is the adaptation and fine-tuning of LLMs like GPT-4 specifically for TSC tasks. Models like LA-Light (Wang et al., 2024d) and LightGPT (Lai et al., 2023) demonstrate frameworks that incorporate LLMs as central agents in their decision-making processes, allowing the traffic signal systems to modify their strategies in real-time based on evolving traffic conditions. These models go beyond generic language processing to directly interact with real-time traffic environments, providing context-aware control strategies (Masri et al., 2024). Such specialization is critical because

generalist LLMs, while powerful, may not always accurately interpret traffic-specific inputs unless they are trained on highly relevant data.

In conclusion, integrating LLMs into traffic signal control systems addresses the limitations of traditional methods by enhancing flexibility, generalization, and decision-making under dynamic conditions. By combining the strengths of RL and LLMs, frameworks like iLLM-TSC improve robustness and adaptability, even in complex or unforeseen scenarios. The full list of the literature is provided in Table 8.

	Table 0 Travel Signal Control Osing LLIVIS						
Literature	Model Backbone	Input	Output	Modality	Task	Data Source	
Pang et al., 2024b	GPT-4	Real-time traffic data and scenario descriptions	Refined signal control decisions	Text and data	Optimize traffic signal control	Traffic simulation data	
Wang et al., 2024c	GPT-4	Static and dynamic traffic data	Optimized traffic signal phase	Text and data	Optimize traffic signal control	Shanghai traffic data and traffic simulation data	
Villarreal et al., 2023	GPT-4	Natural language queries	Suggested state spaces and reward functions for RL tasks	Text	Assist in defining RL components for mixed traffic control	Traffic simulation data	
Lai et al., 2023	LightGPT	Real-time traffic data and task descriptions	Traffic signal control decisions	Text and data	Optimize traffic signal control	Jinan and Hangzhou traffic data	
Masri et al., 2024	GPT-4o-mini	Real-time traffic data and scenario descriptions	Traffic signal control recom- mendations	Text and data	Optimize traffic signal control and improve mixed traffic safety	Generated data from GPT-40-mini	

Table 8 Travel Signal Control Using LLMs

3.3.7. Traffic Simulation Traffic simulation is the process of using computational models to represent and analyze the movement of vehicles and pedestrians within a transportation network. This powerful tool enables researchers and engineer to study traffic behavior under diverse conditions and scenarios. By simulating traffic dynamics, it helps explain traffic patterns (e.g., Xu and Gayah, 2023) and evaluate the potential impacts of traffic control and management strategies (e.g., Yu et al., 2020) prior to real-world implementation.

To perform a traffic simulation, a traffic scenario must be constructed first. The key elements of a traffic scenario include the network (e.g., geometry, road type, and lane configuration), traffic demand (e.g., origin-destination trips), and traffic infrastructure (e.g., traffic signals) (Lopez et al., 2018). Defining these components accurately is crucial for creating a realistic traffic scenario. However,

this process is often complex, as these elements originate from diverse data sources, making data integration and post-processing both challenging and time-intensive.

In recent years, a few studies have emerged that utilize LLMs for intuitive and efficient creation of complex scenarios for traffic simulation. For example, Tan et al., 2023 introduce a language-conditioned model, LCTGen, which utilizes an LLM to transform natural language descriptions of driving scenarios into simulation scenarios that include both the initial states and motions of traffic actors (e.g., vehicles and pedestrians). Similarly, Zhong et al., 2023 combine an LLM with a scene-level diffusion model to bridge user queries and traffic generation. By translating linguistic commands into differentiable loss functions, the proposed model guides the simulation to produce realistic and query-compliant multi-agent traffic interactions such as car following and lane changing.

Güzay et al., 2023 extends the application of LLMs beyond simple scenario generation. In this study, LLMs are employed to convert natural language prompts describing traffic scenarios – such as road layouts, intersections, vehicle types, and traffic conditions – into simulation parameters in "XML" format, directly compatible with the traffic simulator "Simulation Of Urban Mobility" (Lopez et al., 2018) and ready to be simulated.

In summary, LLMs aid traffic simulation by simplifying the creation of realistic and complex scenarios through NLP. LLM-driven approaches enhance the efficiency and adaptability of traffic simulation, supporting more effective traffic management and planning as summarized in Table 9.

Literature	Model Backbone	Input	Output	Modality	Task	Data Source
Tan et al., 2023	GPT-4	Traffic scenario descriptions	Traffic scenarios with vehicle parameters	Text	Generate realistic traffic scenarios	Traffic simulation data
Zhong et al., 2023	GPT-4	Traffic scenario descriptions and traffic data	Traffic scenarios with vehicle trajectories	Text and data	Generate realistic traffic scenarios	Vehicle trajectory data
Güzay et al., 2023	GPT-4	Traffic scenario descriptions	Traffic simulation files in XML format	Text	Generate realistic traffic scenarios and simulation files	Waymo Open Dataset

Table 9 Travel Simulation Using LLMs

**3.3.8.** Road Network Generation The generation of road networks is critical for a wide range of applications, including traffic simulation, autonomous navigation systems, and urban planning. Accurate and efficient road network models are essential for tasks such as optimizing traffic flow, planning transportation infrastructure, and supporting real-time navigation. Traditional methods for generating road networks, such as segmentation-based techniques and manual modeling with traffic simulation software, have been widely used but suffer from significant limitations.

Segmentation-based techniques rely on satellite or aerial imagery and involve generating binary segmentation masks to distinguish between roads and non-road areas. These techniques demand extensive labeled datasets, require time-intensive pre-processing, and are sensitive to image quality. On the other hand, manual modeling with traffic simulation software requires significant human intervention to input road layouts, design intersections, and configure traffic parameters. This process is labor-intensive, prone to errors, and requires domain expertise.

LLMs can generate road networks without these deficiencies, offering advantages such as reduced dependence on labeled datasets, automation of manual tasks, and the ability to process multimodal inputs like text, images, and hand-drawn maps. Their reasoning and recognition capabilities enable faster, more accurate, and cost-effective road network generation.

Through our literature search, we have identified two relevant works in this field: NavGPT introduced by Rasal and Boddhu, 2024 and network generation AI (NGAI) proposed by Chen et al., 2024b. These studies highlight the potential of LLMs in generating road networks and demonstrate innovative approaches that address the shortcomings of traditional methods.

NavGPT is a multi-modal LLM designed to generate navigable road networks directly from aerial images (Rasal and Boddhu, 2024). It takes raw aerial imagery as input and outputs detailed road networks in a recognized format. Unlike traditional segmentation-based methods, NavGPT bypasses the need for binary segmentation masks by aligning visual features from aerial images with language-based outputs. NavGPT builds upon the MiniGPT-4 architecture, leveraging frozen pre-trained components for both vision and language. The model is fine-tuned using a novel training methodology to identify road geometries. It requires only one A100 GPU and approximately 26 hours for retraining, making it lightweight and cost-effective.

NGAI takes a step further to support multimodal inputs, including text, satellite images, and hand-drawn maps, and outputs road networks (Chen et al., 2024b). NGAI employs a U-Net model for satellite image segmentation and advanced corner detection algorithms to extract road geometries. Unlike NavGPT, NGAI relies on pre-trained models for image processing but integrates them with LLMs through the LangChain framework. This combination allows NGAI to autonomously select and invoke plugins for modeling tasks, ensuring flexibility and adaptability for various user queries.

The use of multimodal LLMs for road network generation simplifies workflows, reduces dependence on labeled datasets, and enables efficient processing of diverse input formats. This area of research remains limited as shown in Table 10. Future research is expected to focus on improving the accuracy of image-based recognition, enhancing the adaptability of LLMs to varied traffic scenarios, and expanding their capabilities to handle dynamic and large-scale transportation systems.

### 3.4. Transportation Research

Transportation encompasses the broad movement of people, goods, and services across land, air, and water. Transportation research involves designing, planning, and management of infrastructure,

	Table 10 Road Network Generation					
Literature	Model Name	Model Backbone	Inputs	Outputs	Modality	Data Source
Rasal and Boddhu, 2024	NavGPT	Mini-GPT4	Satellite images and texts	Detailed road networks	Text and image	Satellite images
Chen et al., 2024b	NGAI	GPT (version not specified)	Texts and hand-drawn images	Detailed road networks	Text and image	Open Street Map

Table 10 Road Network Generation

policies, and systems to ensure efficient and sustainable mobility. Unlike traffic, which focuses on localized operational dynamics, transportation research analyzes large-scale networks, long-term planning, and multimodal integration. In this section, we examine how LLMs help enhance operational efficiency, optimize resource allocation, and provide intelligent solutions to tackle the multifaceted challenges of transportation systems.

**3.4.1.** Aviation and Air Traffic Control Aviation and air traffic control (ATC) require precision, real-time decision-making, and smooth human-machine collaboration. These fields grapple with numerous challenges, such as managing dense airspace, directing uncrewed aerial vehicles (UAVs), accommodating noisy environments, and dealing with diverse accents while adhering to stringent regulations.

While traditional methods have advanced speech recognition, predictive modeling, and navigation systems, they sometimes fail to handle the complexity and variability of these areas, particularly with unstructured data or when dynamic responses are crucial. LLMs have proven valuable assets in overcoming these limitations. In this section, we introduce a few research areas, including strategic decision-making in ATC, situational awareness and regulatory compliance in ATC, situational awareness of UAVs, as well as autonomous flight control.

## Strategic Decision-Making in ATC

Strategic decision-making in ATC focuses on proactive, long-term planning to manage traffic flows, mitigate demand-capacity imbalances, and address recurring constraints in the National Airspace System. Such decisions are critical in ensuring air traffic operations' overall efficiency and reliability. Strategic decisions include implementing ground delay programs (GDPs) to manage flight arrival demand at capacity-constrained airports, rerouting flights to avoid restricted airspace or severe weather, and recalling similar past events to inform current decision-making. These tasks require traffic managers to analyze complex data, such as weather forecasts, operational constraints, and historical traffic patterns, to optimize traffic flow while minimizing delays.

Traditional methods for supporting strategic decisions in ATC rely heavily on manual processes, requiring traffic managers to sift through vast, unstructured datasets or rely on experience and intuition. For example, identifying historical GDP patterns or recalling similar traffic scenarios is often

a time-intensive and cognitively demanding task, compounded by the need to filter and extract relevant information from loosely organized data sources. While existing tools, such as the flight schedule monitor (FSM), offer some degree of automation, they are limited in their ability to integrate diverse data types, generate summaries, or provide actionable insights tailored to specific queries.

Abdulhak et al., 2024 introduce an LLM-driven conversational agent that is explicitly designed to address traditional methods' limitations in managing strategic air traffic flow. By FT LLMs on a 23-year dataset of over 80,000 GDP issuances, revisions, and cancellations, CHATATC provides a powerful tool for summarizing historical GDP patterns, extracting insights, and answering context-specific queries in natural language. Unlike traditional methods, CHATATC automates labor-intensive data retrieval and summarization, enabling traffic managers to focus on unique challenges rather than repetitive, routine tasks. The tool also reduces the barriers to training new traffic managers by offering accessible insights into historical patterns, enhancing their understanding of strategic traffic flow management.

#### Situational Awareness in ATC

Situational awareness in ATC refers to the ability of controllers to perceive, understand, and predict the status of air traffic and environmental conditions to ensure safe and efficient operations. A significant challenge in training AI for situational awareness is the lack of anomaly datasets. Fox et al., 2024 address this using GPT-3.5 to generate synthetic data sets, including regular and anomalous air traffic conversations, using FS. These datasets provide diverse training samples for a variational auto-encoder (VAE), which learns a latent representation of "normal" communication patterns. The VAE detects "off-nominal" (anomalous) scenarios by measuring reconstruction loss and identifying communications that deviate from typical patterns. This approach effectively processes unstructured natural language data while overcoming dataset scarcity, offering a robust framework to enhance anomaly detection and situational awareness in civil aviation.

### Regulatory Compliance in ATC

Regulatory compliance involves adhering to established aviation rules, procedures, and standards to maintain safety, prevent conflicts, and support the orderly flow of air traffic within controlled airspace. Regulatory compliance is very text-intensive and has a strong cognitive burden for traffic managers. LLMs have been applied to automate the classification and simplification of regulatory texts, making them more accessible to air traffic controllers. For example, GPT-3 has been used to classify and summarize air traffic flow management (ATFM) regulations, improving comprehension and accountability (Jarry et al., 2024). These tools reduce the cognitive burden, allowing controllers to focus on operational tasks and ensuring safer, more efficient air traffic management.

#### Situational Awareness of UAVs

UAVs increasingly rely on semantic scene understanding to navigate complex environments and communicate effectively with human operators. Traditional UAV navigation systems, while robust in structured settings, often struggle in dynamic or unstructured environments, such as disaster zones or crowded urban areas. LVLMs enable UAVs to process visual and textual data, giving them a contextual understanding of their surroundings. For example, LVLM-driven frameworks can generate semantically rich real-time descriptions of environmental characteristics by combining object detection output (for instance, from YOLOv7) with advanced natural language generation capabilities (for example, GPT-3) (De Curtò et al., 2023). By providing detailed, human-readable descriptions, LLM-enhanced UAVs improve operators' situational awareness and reduce cognitive load, enabling faster and more informed decision-making.

## **Autonomous Flight Control**

In air combat scenarios, high precision, adaptability, and responsiveness are crucial. Traditional flight control methods fall into two categories: model-based approaches and model-free approaches, such as DRL. Model-based methods rely on accurate mathematical modeling which is often limited by uncertainties. DRL offers a model-free framework capable of adaptive control through agent-environment interactions, making it a promising technique to tackle complex and dynamic flight control tasks. However, standalone DRL has limitations such as sparse rewards, low sample efficiency, and slow convergence, which hinder its effectiveness in high-dimensional, six-degree-of-freedom (6-DOF) flight control tasks.

LLMs provide contextual knowledge and logical reasoning capabilities that complement DRL's trial-and-error learning paradigm. For example, in 6-DOF flight control, LLMs serve as a guide mechanism during the training process, improving the quality of agent-environment interactions by injecting domain knowledge and timely feedback into the learning loop (Yang et al., 2024a). This integration allows the intelligent flight controller (IFC) to address sparse reward issues by evaluating agent actions against predefined criteria derived from domain-specific knowledge bases, such as flight operation manuals. LLMs can accelerate the learning process and improve sample efficiency by rejecting suboptimal actions and guiding the agent toward more promising strategies.

Another notable contribution of LLM-guided DRL is its ability to handle the feasibility of complex tactical maneuvers, such as looping, immelmann turn, and split S-flight actions that are often required in air combat. LLMs enhance the training process of DRL by providing a structured evaluation of action feasibility. For example, during early training episodes, LLMs evaluate the agent's control actions (e.g., roll, pitch, yaw, and throttle commands) concerning the aircraft's current state and target goals, rejecting actions that deviate from the desired trajectory. This structured feedback

reduces exploration inefficiencies and accelerates convergence, enabling the IFC to achieve precise control over flight attitudes and perform highly adaptive maneuvers. By integrating logical reasoning and domain knowledge into the learning process, LLMs address the key limitations of DRL, resulting in reduced reliance on exhaustive sample exploration and improved training efficiency.

The intersection of LLMs, RL, and other AI technologies with aviation and ATC heralds a new era of intelligence, automation, and efficiency. These technologies support strategic decision-making, enhance autonomous flight control, and advance UAV situational awareness capabilities. Despite the aforementioned research works, this area remains largely unexplored. A summary of the existing studies is shown in Table 11.

Table 11 Aviation and Air Traffic Control

Literature	Model Backbone	Input	Output	Modality	Task	Data Source
Abdulhak et al., 2024	GPT-4	Historical air traffic data	ATFM suggestions	Text	Conversational ATFM support	Historical ATC data
Carranza et al., 2023	Custom LLM	Text prompts	UAV responses	Text	UAV communication	Custom dataset
De Curtò et al., 2023	GPT-4	Scene descriptions	Semantic scene understanding	Image and text	Scene analysis	UAV scene dataset
Fox et al., 2024	GPT-3.5	Air traffic conversations	Reduces the controller's cognitive load	Text and data	Measuring communications	VAE detects "off-nominal" (anomalous) scenarios
Jarry et al., 2024	GPT-3	ATFM regulations	Categorized regulations	Text	Regulation analysis	ATFM regulatory data
Yang et al., 2024a	ChatGLM-6B	$\begin{array}{l} {\rm State\ variables} \\ {\rm +\ commands} \end{array}$	Flight control actions	Text and data	RL for 6-DOF control	Flight simulations

ATFM: Air Traffic Flow Management

ATC: Air Traffic Control 6-DOF: Six Degrees of Freedom

**3.4.2.** Maritime Transportation Maritime transportation plays a vital role in global trade and economy, connecting nations and facilitating the movement of goods across the world. As the industry continues to evolve and face new challenges, researchers are exploring innovative solutions to enhance efficiency, safety, and sustainability in various aspects of maritime operations. Through our literature search, we have identified three key areas of research that applied LLMs: satellite image detection of ships, navigation and path finding, and unmanned ships.

## Satellite Image Detection

Monitoring the status and locations of vessels is crucial for efficient port operations. Traditionally, this has been achieved through remote sensing technologies that provide operators with visualizations of vessel placements. This process has the potential to be automated thanks to the advancements in

language models, which permit integrating and interpreting data from multiple sources into humanreadable text. The Popeye model proposed by Zhang et al., 2024d integrate visual perception with the generalization capabilities of LLMs. By aligning visual features with language features and employing a cross-domain joint training strategy, the Popeye model achieves superior zero-shot performance on ship interpretation tasks.

## **Navigation and Path Finding**

Enhanced vessel location monitoring serves as a stepping stone to improve maritime navigation, overcoming limitations of traditional methods that rely on human expertise, fragmented data, and manual processing of variables like weather, tides, and traffic patterns. These traditional approaches face challenges in handling dynamic maritime complexities – such as growing vessel sizes, fluctuating conditions, and data silos – leading to inefficient routes, higher fuel consumption, and safety risks.

To address these challenges, researchers have turned to NLP and LLMs as powerful tools for enhancing vessel navigation and path planning (Wang et al., 2024e). KUNPENG is a comprehensive model in this field that combines NLP, LLMs, and other AI techniques. It enables the real-time processing and analysis of vast amounts of heterogeneous data, including weather conditions, vessel performance, and traffic patterns. This allows for the generation of optimal route recommendations, the early detection of potential hazards, and the coordination of multi-vessel operations, all of which contribute to enhanced safety, efficiency, and sustainability in the maritime industry.

### **Unmanned Ships**

The emerging field of unmanned ship systems (USSs) promises enhanced safety and efficiency in maritime operations by eliminating the need for human crews, thus reducing the risks associated with human error. However, the absence of onboard personnel poses challenges in detecting operational anomalies. To address this, recent research by Li et al., 2024a has developed a sophisticated anomaly detection framework that integrates LLMs with a Bi-LSTM model. This approach preprocesses operational data to generate vectorized representations, which are then used to detect anomalies in real time. It identifies deviations from established patterns, enabling prompt responses to potential issues. This method has shown superior performance over traditional machine learning techniques, significantly boosting the survivability and reliability of autonomous maritime systems.

The application of LLMs and other advanced AI techniques in maritime transportation has the potential to improve various aspects of the industry. From satellite image detection of ships to navigation and path finding, and to the development of unmanned ships, LLMs are enabling more efficient, accurate, and data-driven decision-making processes. Nonetheless, the existing research remains scant, as shown in Table 12, suggesting that the area of maritime transportation remains fertile for future exploration of LLM applications.

				•		
Literature	Model Backbone	Input	Output	Modality	Task	Data Source
Li et al., 2024a	BERT& GPT	Vectorised log data	Real-time anomaly	Text	Anomaly detection	Simulated data
Wang et al., 2024e	Not mentioned	sensor data, satellite data, navigation data, meterological data	Autonomous navigation instructions	Text and image	Intelligent maritime navigation	Maritime knowledge base
Zhang et al., 2024d	LLaMA	Remote sensing imaginery	Horizontal bounding boxes	Text and image	Ship detection	Satellite images

Table 12 Maritime Transportation

VHF: Very High Frequency

VTS: Vessel Traffic Service

AIS: Automatic Identification System Ship I/O: Ship input/output list

**3.4.3.** Supply Chain Management There has been a broad spectrum of research in the field of supply chain management. While supply chain management encompasses a wide range of topics, this subsection focuses specifically on research related to transportation and logistics. This area of research involves handling complex networks, managing dynamic decision-making processes, and extracting actionable insights from unstructured data.

According to the research agenda provided by Dhara and Barba, 2024 and Aguero and Nelson, 2024, LLMs have the potential to be applied in several aspects of supply chain management, including knowledge management, demand forecasting, customer service, contract analysis, and quality control and maintenance. Among these aspects, knowledge management and demand forecasting are particularly relevant to research in the field of transportation logistics. Knowledge management involves automating data extraction, information summarization, and providing access to vast amounts of structured and unstructured data for risk analysis and decision-making (AlMahri et al., 2024). Demand forecasting utilizes real-time market data, consumer behavior, and historical trends to predict future demands and provide optimization (Quan and Liu, 2024).

## Knowledge Management and Risk Identification

Modern manufacturers often source raw materials from suppliers across multiple tiers, making supply chain visibility crucial for risk management and decision-making. However, obtaining and managing knowledge beyond tier 1 suppliers can be challenging. AlMahri et al., 2024 proposes a framework that leverages LLMs to facilitate smoother knowledge extraction and management from unstructured sources. The framework collects data and information about suppliers from unstructured web resources and applies ZSL for named entity recognition (NER). NER identifies and classifies entities such as company names, locations, and product types within large volumes of text. Subsequently,

relation extraction (RE) is employed to delineate the relationships between these entities, forming the edges of the knowledge graph that illustrate the flow of materials, information, and finances across the supply chain.

The purpose of knowledge extraction is to serve risk identification in the supply chain. Shahsavari et al., 2024 applt LLM+Bayesian Networks approach for the probabilistic reasoning of risky events. This approach extracts information from news feeds of events and datasets, feeding it into Bayesian Networks to model the probabilities of events that lead to significant supply chain risks based on causal relationships. This method enables merchants to better identify potential events that may cause significant disruptions to the supply chain. Shahsavari et al., 2024 successfully tests this method in the transportation sector of Victoria, Australia in 2021, identifying four contributing events to supply chain disruptions: rising COVID cases, vaccination mandates, construction worker strikes, and a bridge blockage. The method provides detailed probabilities of each contributing event causing disruptions.

While historical data offers valuable insights into past risks, historical data may not account for novel or unprecedented scenarios. To address this limitation, simulation modeling has emerged as a complementary approach for risk identification. According to Jackson et al., 2024, LLMs can generate complex Python code for logistics simulations using appropriately designed prompts. These models enable researchers to simulate hypothetical scenarios and assess their potential impacts on supply chains.

### **Demand Forecasting and Inventory Decisions**

With the simulation model in place, inventory decisions can be made. Quan and Liu, 2024 address inventory decisions using LLMs. The purpose of the research is to replace traditional regression models with LLMs. The authors feed the LLM agent with historical demand information, inventory levels, and other essential state features, requesting the LLM to output an order quantity. It is found that this method works fairly well under simple demand scenarios, offering better explainability than deep models. However, this method underperforms regression models when it comes to more complicated scenarios.

Li et al., 2023 propose a more powerful application of LLMs in modeling what-if scenarios in supply chains. Instead of directly asking LLMs to output results, they design an agent named 'OptiGuide' that automatically interacts with optimization models from the backend to derive results. The interaction with the LLM within OptiGuide involves inputting queries in natural language, which the LLM then translates into optimization code. This code is executed by traditional optimization solvers that interact with databases to fetch necessary data and compute outcomes. The results are then converted back into natural language by the LLM, providing users with actionable insights and detailed explanations of the optimization results. OptiGuide represents a significant step forward in

making complex supply chain optimizations accessible to a broader range of business users without requiring them to have specialized knowledge in optimization algorithms or machine learning.

Despite the aforementioned advantages, most industry interviewees indicate that LLMs should serve an assistive role rather than taking full autonomy (Dhara and Barba, 2024). Users must also be cautious about data privacy, the accuracy of responses, and the operational dependence on LLMs. The future of supply chain management with LLMs needs to take an integrated approach where technology complements human skills rather than replacing them.

LLMs offer unique advantages to improve supply chain management practices by enhancing efficiency, reducing costs, minimizing risks, and improving service delivery. However, they must also be used carefully to mitigate the associated risks. The full list of the literature is provided in Table 13.

Table 13 Supply Chain Management

Literature	Model	Input	Output	Modality	Task	Data Source
	Backbone					
AlMahri et al., 2024	GPT-4	EV manufacturers info	Knowledge graph	Text & quantity	Find relationships of suppliers	Wikipedia
Chen et al., 2024f	GPT-3.5 & 4	Agent states	Movement	Text and quantity	Collaborative boxes	None
Jackson et al., 2024	GPT-3	Inventory simulation prompt	Python code	Text	Generating simulation code	None
Li et al., 2023	GPT-3.5 & 4	If-then questions	New solutions to changes	Text	Implement optimisation and interpret solutions	Open source benchmarks
Quan and Liu, 2024	GPT-4	Stage, state information of supplier	Order quantity	Text and quantity	Inventory decisions	Generated data
Shahsavari et al., 2024	GPT-3.5	News	Probabilities of contributing events	Text	Identify contributing events and risks	Google news
Zhao et al., 2024a	GPT-3.5 & open-source pre-trained models	News, merchant and supplier information	Risk information	Text	Label risks given news	Tier 1 suppliers, Google news
Aguero and Nelson, 2024			Research ag	genda paper		
Schöpper and Kersten, 2021	Review paper					
Dhara and Barba, 2024	Research agenda paper					
Wang et al., 2024h			Research ag	genda paper		
Zhou et al., 2021			Research ag	genda paper		

EV: Electric Vehicle

**3.4.4. Public Transportation** When LLMs are applied to public transportation systems, three research areas are identified: passenger demand forecasting, bus holding control, and infrastructure analysis.

## Passenger Demand Forecasting

Accurate demand prediction is vital for optimizing resource allocation and service reliability, especially under irregular conditions like delays or disruptions. Conventional methods, such as ARIMA and LSTM-based models, rely on structured numerical data (e.g., historical ridership) but struggle to adapt to dynamic, unstructured variables like real-time delays or weather events. For instance, while GC-LSTM incorporates spatial dependencies, it fails to contextualize how sudden delays propagate through a network, leading to inaccurate forecasts. Huang, 2024 tackles this by reformulating numerical and spatial data into natural language prompts (e.g., "A 15-minute delay occurred at Station X; predict passenger flow at Station Y"). The proposed LLM-based framework employs CoT reasoning, enabling step-by-step analysis of delay impacts and passenger behavior. This approach significantly outperforms traditional models in accuracy during irregular scenarios, offering interpretable predictions that aid transit agencies in proactive scheduling.

### **Bus Holding Control**

Bus holding, i.e., delaying departures to maintain schedule adherence, is a key strategy for reducing congestion and improving service regularity. Traditional model-based methods use rigid algorithms to predict bus states and demand, but they lack flexibility in dynamic environments. RL emerged as a promising alternative by framing holding decisions as a reward-maximization problem. However, RL's reliance on manually designed reward functions (e.g., penalizing deviations from schedules) limits its adaptability and requires labor-intensive tuning. Yu et al., 2024 integrate LLMs into RL to automate reward function design. Here, LLMs interpret real-time operational data (e.g., passenger load and traffic) and generate context-aware rewards (e.g., prioritizing crowded buses for priority dispatch). This hybrid paradigm reduces human intervention, accelerates policy convergence by 30%, and improves on-time performance by 15% in simulations. The LLM's ability to parse unstructured data (e.g., driver reports) further enhances decision-making in edge cases, such as accidents or road closures.

## Infrastructure Analysis

Bus stop accessibility and quality directly impact user experience and equity. Traditional evaluations rely on computer vision tools like YOLO for detecting specific features (e.g., wheelchair ramps). While effective for isolated tasks, these methods lack contextual reasoning – for example, recognizing

that a ramp obscured by debris is non-functional. Oliveira et al., 2024 proposes an LLM-driven framework that synthesizes multimodal data (street-level images, maintenance logs) to generate holistic assessments. The model identifies physical defects (e.g., cracked pavements) and correlates them with accessibility metrics (e.g., ADA compliance), providing actionable insights like, "Tactile paving at Stop 12 is 80% worn, posing risks to visually impaired passengers." This approach scales evaluations across entire networks, reducing manual inspection costs by 40% while prioritizing high-impact repairs.

In summary, LLMs demonstrate great potential in enhancing public transportation systems across three areas: passenger demand forecasting, bus holding control, and infrastructure analysis. The advancement of LLMs helps improve prediction accuracy, streamline decision-making, and optimize resource allocation, aiding transit agencies in providing more reliable and equitable services. A summary of the reviewed literature is presented in Table 14.

Literature Model Inputs Outputs Modality Data Source Backbone Huang et al., GPT-4 Text and data AFC data from Delay, historical Passenger flow 2024bpassenger flow predictions Shenzhen Metro data, adjacency matrix State-action Yu et al., 2024 GPT-4 Optimized Text and data Bus data from data reward function Beijing INACITY Oliveira et al., LLaVA Street-level Qualitative and Text and image 2024 images, quantitative platform evaluation assessments criteria

Table 14 Public Transportation

### 3.5. Multi-task Applications

Many LLMs are versatile and can cover a range of traffic and transportation tasks. This section focuses on reviewing LLM works that span multiple aspects of traffic and transportation research. According to Zhang et al., 2024e, the various tasks that LLMs cope with can be categorized into three main areas: memorization, understanding, and application as shown in Figure 6.

Memorization refers to an LLM's ability to recall answers to specific questions. Syed et al., 2024 evaluate this capability using datasets from TransportBench, which consists of undergraduate-level exam questions covering various traffic- and transportation-related topics, such as transportation economics, driver characteristics, vehicle motion, road geometry design, traffic flow/control, transportation planning, utility/modal split, transportation networks, and public transit systems. The authors employed a zero-shot prompting strategy, directly inputting the original problem descriptions into web-version LLMs. The accuracy rates range from 40% to 70%, demonstrating the LLMs' ability to memorize and recall transportation-related information. Besides zero-shot learning, LLMs also show the ability to retrieve information from databases as needed. The ReBoostSQL framework, introduced by Sui et al.,

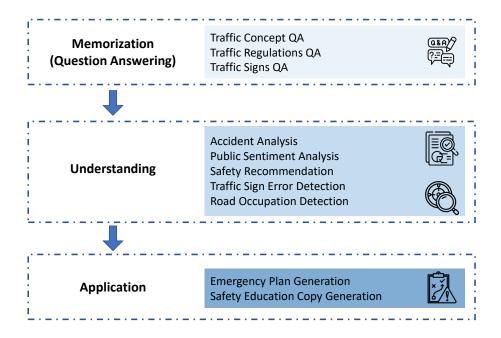


Figure 6 Capabilities of LLMs in the field of traffic and transportation research

2023, exemplifies the model's prowess in converting natural language queries into executable SQL commands. This capability is particularly beneficial in complex business environments where traffic data is stored across large, intricate databases. By enhancing query accuracy and linking these queries to specific database schemas, ReBoostSQL improves the efficiency of data retrieval and ensures the relevance and precision of the results.

Moving beyond memorization, understanding involves an LLM's capacity to analyze text and data and make recommendations based on its findings. TransGPT, a model fine-tuned with a vast amount of textual data from various transportation-related sources, exemplifies this capability. The model's training data includes books, reports, documents, websites, driving tests, traffic signs, landmarks, and corpora. Wang et al., 2024f showcases the potential applications of TransGPT in traffic analysis and modeling, such as generating synthetic traffic scenarios, explaining traffic phenomena, answering traffic-related questions, providing traffic recommendations, and generating traffic reports. These applications highlight the model's ability to understand and interpret transportation-related information.

The third category, application, refers to an LLM's ability to generate traffic operational plans and strategies. The innovative Open-TI model, introduced by Da et al., 2024, aims to create a Turing Indistinguishable level of traffic intelligence. This model integrates the theoretical strengths of LLMs with practical traffic management needs, marking a shift towards operational applications that focus on real-time traffic simulation and management. Open-TI processes natural language inputs

and generates detailed traffic analyses and simulations, representing a significant step towards creating intelligent transportation systems that can interact seamlessly with human operators and adapt dynamically to changing traffic conditions.

The most comprehensive model is TrafficGPT developed by Zhang et al., 2024c, which provides a framework combining the foundational strengths of LLMs with Traffic Foundation Models to enhance traffic management. TrafficGPT facilitates a deeper understanding of traffic data, enabling more nuanced and effective traffic control strategies. The framework uses natural language inputs from users to guide the execution of various traffic-related tasks through a sequence of steps involving understanding, planning, and execution using traffic foundation Models. The full list of the literature surveyed in this subsection is provided in Table 15.

Table 15 Multi-task Applications

Literature	Model Name	Model Backbone	Modality	Task	Data Source
Da et al., 2024	Open-ti	Pre-trained traffic LLMs	Text and data	Traffic data analysis, simulation, optimization and control	Traffic datasets, maps, real-time traffic data
Sui et al., 2023	N/A	GPT-4	Text and SQL	Text-to-SQL, schema linking SQL generation, SQL boosting	CTtraffic and CompanyZ
Syed et al., 2024	None	GPT-4V	Text and images	Traffic event recognition, analysis and reporting	Open-source datasets
Tian et al., 2023	None	GPT-4, Claude 3.5, Gemini 1.5, Llama	Text	Solving transportation engineering questions	TransportBench dataset
Wang et al., 2024f	TransGPT	ChatGLM, VisualGLM	Text and images	Answering transportation-related questions	Traffic engineering documents, examinations papers
Zhang et al., 2024e	TransportationGame	e 16 LLMs	Text and images	Memorization, understanding, application	Examination papers, news, gov websites, images
Zhang et al., 2024c	TrafficGPT	GPT-4	Text and images	Traffic management and decision support	Various traffic data sources

In conclusion, the surveyed literature highlights the versatility, effectiveness, and remarkable potential of LLMs in transforming various aspects of the transportation domain. As research in this field progresses, it is evident that the integration of LLMs with domain-specific models and frameworks

will play a crucial role in making traffic and transportation systems more intelligent, efficient, and adaptable.

### 4. Current Trends and Future Directions

Our literature review has uncovered a variety of LLM methodologies and applications in traffic and transportation research. Yet significant challenges remain in addressing the associated problems. In this section, we provide a statistical analysis of the studies examined, to highlight the prevailing trends and explore the potential avenues for future research.

### 4.1. Statistical Analysis

Figure 7 presents a temporal distribution of LLM usage in transportation research from 2023 to 2025. We observe a moderate level of adoption in 2023, where GPT-4 accounts for the largest share among the models used, followed by GPT-3.5, ChatGLM, and unspecified version of GPT. This initial wave likely reflects early experimentation as researchers and practitioners assessed the feasibility of integrating LLMs into diverse transportation problems. In 2024, there is an increase in overall LLM usage, with GPT-4 as the most significant increase—surpassing all other models in both absolute and relative terms. LLaMA and "Other models" also show substantial growth, indicating a rapidly diversifying research ecosystem. The jump in 2024 could be attributed to the broader release and improved accessibility of newer models, as well as increased recognition of the potential benefits of language-based approaches for tasks such as policy formulation, demand forecasting, and real-time system management. As this paper is finished at the beginning of 2025, papers in this year are mostly not included. We believe that there will be an increasing trend in the number of papers in 2025.

Figure 8 illustrates the mapping of LLMs to various transportation applications, capturing the breadth of how different models have been applied in this domain. The applications include traffic signal control, traffic forecasting, autonomous driving, aviation and air traffic control, supply chain management, travel behavior prediction, and several others. We observe that GPT-4 is prominently featured across a wide range of these applications, likely due to its enhanced reasoning capabilities and broader availability after its introduction. GPT-3.5 is also widely adopted, reflecting its early presence and established user base. Meanwhile, ChatGLM, LLaMA, and other models each find more specialized applications. Again, if a particular model is used only once, it is categorized under "Others" to maintain clarity in the mapping. This mapping diagram shows that certain application areas, such as autonomous driving and travel safety analysis, attract the widest variety of LLMs, suggesting that these tasks may require flexible language reasoning, scenario analysis, or large-scale data handling. By contrast, more specialized tasks such as travel simulation or emergency management appear to cluster around a smaller set of preferred models.

Figure 9 provides another perspective by illustrating the proportion of LLM usage across major transportation application categories. Here, AD (26.9%) emerges as the most represented application,

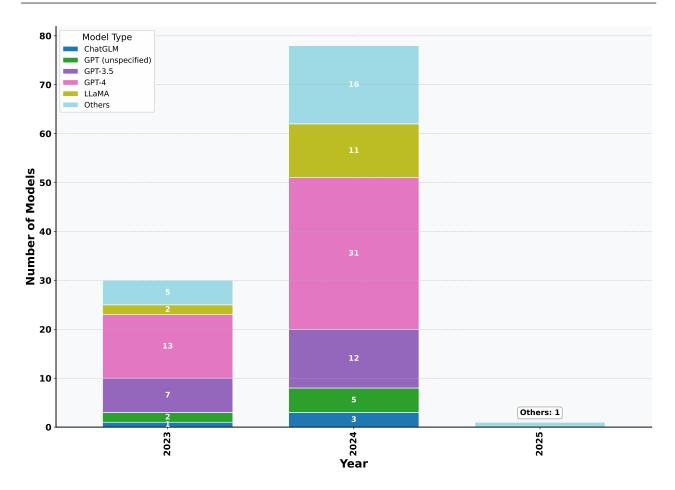


Figure 7 Number of Publications by Year

reflecting the strong interest in leveraging LLMs for tasks such as decision-making in autonomous vehicles and real-time navigation support. Travel Behavior Prediction (14.0%) ranks second, followed by Traffic Forecasting (8.6%) and Travel Safety Analysis (8.6%). Multi-task Applications and Supply Chain Management each account for 7.5%, suggesting growing diversity in how LLMs are deployed. Meanwhile, Aviation & ATC (6.5%) and TSC (5.4%) occupy the mid-range, while Emergency Management (4.3%), Travel Simulation (3.2%), Public Transportation (2.2%), Maritime Transportation (2.2%), and Road Network Generation (2.2%) form smaller but noteworthy slices. Taken together, these distributions indicate that certain areas—particularly those involving real-time or highly dynamic tasks—are more likely to incorporate LLMs, whereas emerging domains may still be experimenting with smaller-scale or pilot deployments.

Overall, these figures reveal two primary trends. First, GPT-based models (GPT-3.5 and GPT-4) maintain a dominant presence across a wide spectrum of transportation research topics. Second, while other LLMs such as ChatGLM and LLaMA are not as universally deployed, their increase in specialized domains suggests that researchers are exploring more diverse solutions to address emerging traffic and transportation challenges. Future studies may benefit from a closer examination of how

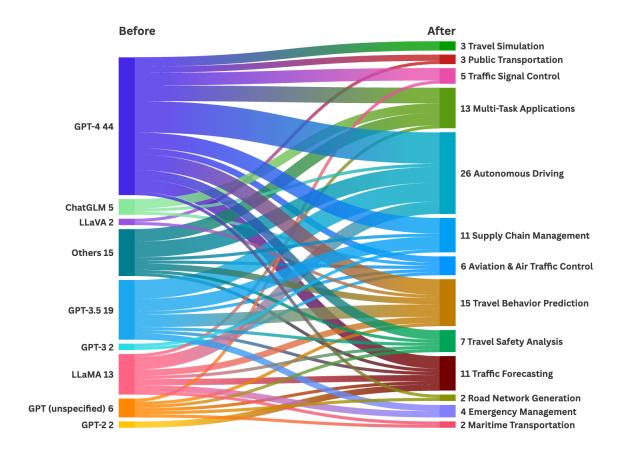


Figure 8 Mapping of LLM Models to Applications

specific model features – such as FT protocols, context-window lengths, or multimodal capabilities – translate into performance gains in particular transportation applications.

### 4.2. Future Directions for LLM Methodologies

### LLMs for Problem Formulation and Solver Integration

LLMs can process natural language descriptions of transportation and traffic problems, translate them into formal mathematical models (e.g., linear programming, integer programming), and interface with backend solvers to find solutions. This integration significantly lowers the barrier for non-experts to utilize OR techniques, by shifting the burden of model construction to the LLM. Through our literature search, we have only identified a relevant paper in supply chain management (Li et al., 2023). A general LLM-integrated solver named ORLM has been proposed, as demonstrated in the methodology section of this paper, but a field-specific model remains lacking.

Relevant methods have the potential to be applied to a variety of traffic and transportation challenges. For example, in traffic signal optimization, natural language queries such as "Optimize green light timings to reduce congestion during peak hours" can be translated into formal linear programming models, which can then be solved to minimize delays or maximize traffic flow efficiency. Similarly,

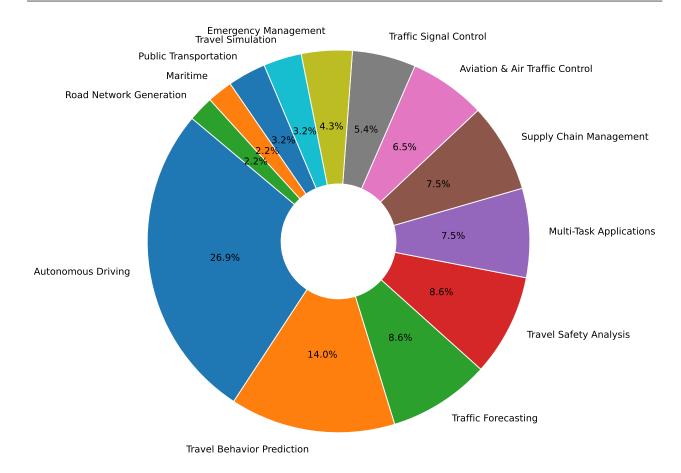


Figure 9 Applications Percentage

in transit scheduling, LLMs can process inputs like "Create a bus schedule to maximize coverage while minimizing idle time for the fleet" and construct integer programming models to generate optimized schedules. Another promising application is freight route optimization, where LLMs can handle descriptions such as "Find the cheapest and fastest way to deliver goods from warehouse A to multiple destinations" by building transportation models, such as network flow or mixed-integer programming models, and interfacing with solvers to identify the most efficient routes. These examples highlight how LLMs can lower the barrier for non-experts to use advanced OR techniques, enabling more accessible and effective decision-making.

## **Automatic Design of Heuristics**

Traditional heuristics (and meta-heuristic) methods are applied to solve a variety of traffic and transportation problems, such as signal control, delivery and scheduling. Traditional manual heuristic design requires significant domain expertise and time, while automatic heuristic design methods such as genetic programming cannot produce new heuristics from scratch (Liu et al., 2024c). LLMs have the

potential to address these limitations by simultaneously adjusting heuristic parameters and exploring new heuristics.

Liu et al., 2024c introduce evolution of heuristics (EoH), a novel framework that combines LLMs and evolutionary computation (EC) to automate the design of optimization heuristics. Unlike traditional methods that manually craft heuristics or rely on NNs, EoH iteratively evolves both the thoughts (natural language descriptions) and codes (executable implementations) of heuristics. It uses LLMs to generate and refine heuristics based on prompt strategies. The proposed framework not only outperforms hand-crafted heuristics but also surpasses other automatic heuristic design methods, while requiring significantly fewer computational resources.

This method can potentially be applied to traffic and transportation systems by designing heuristics for complex optimization problems such as traffic signal control, vehicle routing, dynamic ride-sharing, and public transit scheduling. For example, in vehicle routing, EoH could evolve heuristics to dynamically allocate vehicles to customers in real time, considering factors like traffic congestion, fuel minimization, and service time. Similarly, for traffic signal optimization, EoH can design adaptive heuristics to adjust signal timings dynamically to reduce congestion and improve traffic flow. The ability of EoH to combine linguistic reasoning (thoughts) with code generation enables the generation of domain-specific heuristics for traffic systems that are adaptive and efficient, particularly in scenarios where real-time decisions are crucial.

### Prompt Optimization with Operational Research Techniques

As shown in the survey in the previous section, efficient prompt techniques such as CoT greatly enhance the accuracy of outputs from LLMs. Prompts can be optimized not only with human experience, but also with mathematical optimization methods. Unlike continuous domains where gradient descent is applicable, prompt optimization is inherently discrete.

Pioneering works in this field have used this method to address adversarial prompts—inputs deliberately crafted to bypass a language model's safety mechanisms and generate harmful outputs (Thompson and Sklar, 2024). The authors introduce a framework called fluent student-teacher redteaming (FLRT), which leverages fine-tuned LLMs. In this framework, objectives are defined to both maximize attack success rates (ASR) and minimize token repetition, thereby preserving natural fluency as shown in Equation (7). At the same time, constraints such as maintaining a reasonable prompt length and meeting fluency criteria are imposed to shape the overall optimization as shown in Equation (8).

$$\min_{\text{prompt}} \text{ Discrepancy}(\text{LLM output}(\text{prompt}), \text{target})$$
(7)

subject to:

Fluency(prompt) 
$$\geq$$
 threshold. (8)

To solve this discrete problem, heuristic methods, including greedy coordinate gradient (GCG) and the BEAST algorithm, are employed. These methods iteratively adjust tokens, propose new sequences, and assess their performance, ultimately generating more effective prompts.

Prompt optimization techniques are relatively novel and have not been applied in any literature in our field throughout our literature search. They have broader applications in tasks like scheduling, logistics, and decision-making, where fluency corresponds to interpretability in operational contexts. For instance, in traffic management, optimized prompts could guide AI systems to generate adaptive routing suggestions during congestion. Similarly, in logistics, multi-task optimization of delivery routes could be achieved by leveraging the same principles described in FLRT. By integrating operations research methodologies, prompt optimization becomes a powerful tool to balance competing objectives like efficiency, safety, and fluency while navigating the constraints of discrete, complex systems. This area remains limited through our literature search.

## Retrieval-Augmented Generation

RAG enhances LLMs by integrating a retrieval mechanism that fetches external information dynamically during the generation process. Instead of relying solely on the LLM's pre-trained knowledge, which is static and limited to its training data, RAG dynamically retrieves relevant data from external sources such as text corpora, databases, or APIs. This dynamic retrieval capability enables LLMs to provide up-to-date and accurate information, reducing the likelihood of hallucinations (incorrect or fabricated outputs). Additionally, RAG allows LLMs to handle domain-specific tasks without requiring extensive FT, making LLMs more versatile and cost-effective.

The RAG framework consists of two key components. The first is the retriever, which fetches relevant external data based on the input query. The second is the generator, which uses the retrieved data to synthesize coherent, contextually informed responses. Together, these components create a system that combines the strengths of retrieval-based systems with the generative capabilities of LLMs.

Relevant techniques have been used in travel behavior studies (Wang et al., 2024b) and for AD (Hussien et al., 2025). Application opportunities also abound in other areas. For example, in transportation, an RAG model can retrieve real-time metro delay logs and combine them with its reasoning capabilities to generate actionable insights for passenger flow management. This integration of external, real-time data with the model's reasoning ability makes RAG particularly valuable for dynamic, real-world applications.

### **Knowledge Graphs**

KGs, which represent entities and their relationships in a structured format, complement LLMs by providing a source of grounded and interconnected knowledge. This integration is achieved through methods such as embedding KGs into vector spaces using graph neural networks, enabling LLMs to query KGs during inference via RAG, or using KGs as external reasoning modules alongside LLMs. By combining unstructured generative capabilities with structured relational data, LLMs are capable of generating more accurate, interpretable, and contextually relevant outputs. Through our literature search, only a few works that apply KGs for regulatory compliance can be found, in the areas of emergency traffic management (Chen et al., 2023), supply chain management (AlMahri et al., 2024), and autonomous driving (Hussien et al., 2025).

The combination of KGs and LLMs has the potential to open up significant possibilities for optimizing operations and enhancing decision-making. For emergency traffic management, KGs can encode relationships between road networks, traffic incidents, weather conditions, and historical congestion data. LLMs can leverage the information to provide dynamic traffic insights, such as predicting congestion hotspots or suggesting alternate routes. For supply chain management, KGs capture the relationships between warehouses, delivery hubs, and road networks, to allow LLMs to optimize delivery routes, manage inventory, and mitigate risks associated with disruptions.

Autonomous driving may further benefit from KG-LLM integration by enabling context-aware decision-making. KGs can represent the relationships between road entities, traffic conditions, and pedestrian behavior, while LLMs reason over this data to generate safer navigation strategies. Moreover, public transportation systems may use KGs to integrate real-time transit schedules, passenger feedback, and sensor data, allowing LLMs to generate personalized travel recommendations and demand forecasts. These capabilities contribute to improving the system operational efficiency, reducing traffic delays, and enhancing user experiences for passengers.

## Large Vision-Language Models

LVLMs can perform complex multimodal tasks such as image captioning, visual question answering (VQA), text-to-image generation, and scene understanding. These models are trained on large datasets of paired visual and textual information, allowing them to understand and align visual features with linguistic concepts. One of the key advantages of LVLMs is their ability to work in few-shot or zero-shot settings, where they can handle new tasks with minimal or no additional training. This adaptability, combined with their ability to retrieve and interpret cross-modal information, makes LVLMs suitable for a wide range of applications. They also excel at improving accessibility by generating descriptive text for visual data, enabling visually impaired individuals to better understand visual content. Furthermore, LVLMs facilitate natural human-computer interactions, where users can interact with AI systems using both text and images.

Through our literature search, a few existing studies are identified that apply this technique in the perception of autonomous driving (Wang et al., 2023e; Li et al., 2024b), maritime transportation (Zhang et al., 2024d), travel behavior analysis (Zhao et al., 2024b; Zhang et al., 2024a), and scenario engineering of accident analysis (Xuan et al., 2024).

Beyond these uses, LVLMs has potential for other applications as well. For example, in traffic management, LVLMs could process live camera feeds to deliver text updates on road conditions, congestion, or crashes, while addressing queries like "Why is Main Street blocked?" by merging real-time visuals with language skills. In public transit, they could analyze video of passenger flow at stations, suggesting actions like "Station X is overcrowded; add more trains." Likewise, in logistics and supply chain management, LVLMs could streamline warehousing and freight by monitoring goods, overseeing loading processes, and verifying safety compliance via video, linking visual insights to actionable language outputs. These possibilities highlight LVLMs' game-changing impact across varied domains.

## Light and Specialized LLMs

We need lightweight and specialized LLMs to overcome the significant challenges of deploying traditional, resource-intensive models in real-world applications. Standard LLMs, with their massive parameter counts and high computational demands, are typically confined to high-performance cloud servers, making them unsuitable for mobile devices, offline environments, and applications requiring low energy consumption and high privacy. Lightweight models address these limitations by optimizing for lower latency, reduced memory usage, and energy efficiency, enabling their deployment on resource-constrained devices like smartphones, smartwatches, and compact medical tools. These advancements not only expand accessibility but also reduce operational costs and environmental impact, paving the way for broader adoption across various industries (Misra, 2024).

Additionally, specialized LLMs are crucial for delivering high performance within specific use cases while maintaining efficiency. By employing innovative techniques such as quantization, ternary value representation (e.g., BitNet b1.58), and shared compression layers, these models achieve impressive results with limited computational resources. For instance, models like MiniCPM-V and MobileLLM demonstrate the ability to perform complex tasks, such as image and video processing, on smaller devices (Balani, 2024). This specialization enables the integration of advanced AI capabilities into vehicles, airplanes or any other means of mobility.

In our literature search, a few specialized LLMs are identified for traffic data analysis, simulation, optimization, and control (Da et al., 2024; Zhang et al., 2024c) as well as for answering transport-related questions (Wang et al., 2024f). However, the number of specialized models remains small, highlighting the need for more targeted research and development in this area.

### Mixture of Experts

Mixture of experts (MoE) is an ML technique that combines the strengths of multiple specialized models, called experts, to solve complex tasks. In MOE, each expert is trained to specialize in a specific aspect of the problem, while a gating mechanism (typically another model) determines which

expert(s) to activate for a given input. A classic example of success is DeepSeek-V3, which applied MoE method to vastly improve computational efficiency, where only a small subset of experts (specialized sub-networks) is activated for each input (Liu et al., 2024a).

At the time of conducting our literature review, we find that MoE is rarely used in traffic and transportation applications. On the other hand, MoE can be applied, for instance, with one expert specializing in urban traffic flow prediction while another expert handling freight logistics optimization. This expert specialization reduces computational costs by ensuring that only the most relevant parts of an LLM are used, making MoE ideal for resource-constrained environments. Furthermore, the gating mechanism in MoE dynamically selects the most appropriate experts based on input, enabling an LLM to handle diverse transportation tasks with precision and efficiency.

For further specialization, ensemble methods can be employed with multiple lightweight models combined and each optimizing a specific task. For example, one model could focus on detecting traffic incidents, while a second model predicts their impacts and a third optimizes route suggestions. These models can work collaboratively to provide comprehensive solutions, ensuring that the system remains versatile and capable of addressing complex and interrelated transportation challenges.

## Hybrid Online and Onboard LLMs

Hybrid online and onboard LLMs combine the adaptability of cloud-based continuous learning with the reliability of real-time, localized decision-making. While most LLM applications in autonomous driving and traffic systems focus on either centralized, online models or isolated, onboard systems, hybrid approaches are often overlooked due to the technical complexity and the challenges of synchronizing online and onboard functionalities. However, hybrid models are uniquely positioned to address the demands of transportation systems, where both long-term adaptability and immediate responsiveness are essential.

Throughout our literature search, we have not found any applications of a hybrid approach in traffic and transportation research despite multiple potentials. For example, in traffic management, online LLMs analyze global trends and predict long-term patterns, while onboard LLMs optimize local traffic flows in real time. For AVs, onboard systems handle immediate sensor data (e.g., obstacle detection or navigation), while online systems refine driving algorithms by learning from global edge cases. In personalized navigation, online LLMs provide route recommendations based on user preferences, while onboard systems adapt to real-time inputs and hazards. Hybrid systems also enhance traffic safety by combining predictive online insights with onboard hazard detection and improve urban planning by integrating high-level design insights with real-world monitoring through embedded onboard models.

## Hybrid Reinforcement Learning and LLM

RL is a very popular technique nowadays for real-time optimization. However, as mentioned earlier the integration of RL and LLMs is still in its infancy in traffic and transportation research. In our literature search, we have only identified papers in autonomous flight control (Yang et al., 2024a) and bus holding control (Yu et al., 2024).

One approach of hybridizing RL and LLM is to use an LLM as the task planner and RL as the task executor: an LLM breaks down high-level goals into actionable sub-tasks, while RL optimizes their execution. For instance, in traffic management, an LLM might propose rerouting during congestion, and an RL agent adjusts signal timings accordingly. In AVs, LLMs could translate instructions like "avoid construction zones" into actions refined by RL.

Another approach of hybridization is to use LLMs to enhance RL state representation in multimodal settings, such as traffic systems with textual incident reports, sensor data, and camera feeds. By creating a unified, semantically rich environment model, LLMs enable RL agents to make better-informed decisions, e.g., interpreting "accident on Interstate Highway 101" to guide resource allocation. In logistics, LLMs could simulate constraints like delivery delays, aiding RL in real-time routing adjustments. Similarly, in human-AI systems like ride-sharing, LLMs manage user preferences while DRL optimizes fleet distribution. This synergy of LLMs and RL holds significant promise across diverse application areas.

## Integrating 3D Reconstruction with LLMs

AD systems often struggle with scene understanding when relying solely on static images (Sreeram et al., 2024; Wen et al., 2024). Significant challenges exist in interpreting complex, dynamic environments. To address the challenges, the method of integrating 3D reconstruction with LLMs involves using 3D reconstruction techniques to generate detailed spatial data from various sensors, including cameras, LiDAR, and radar (Wang et al., 2024i). The spatial data is then transformed into a format that LLMs can process, such as text or structured data, enabling LLMs to interpret and reason about the driving environment in a more nuanced way.

Looking ahead, several key areas require further exploration to fully realize the potential of integrating 3D reconstruction with LLMs in autonomous driving. One direction could involve training LLMs to predict and narrate future 3D scene changes, for example anticipating a pedestrian's path from reconstructed trajectories to enhance proactive navigation. Another idea is integrating LLMs with real-time 3D semantic mapping to enable vehicles to query and reason about unseen areas (e.g., "What's around the corner?") using reconstructed spatial context. A third possibility is developing LLMs that optimize 3D reconstruction itself by prioritizing key environmental features (e.g., obstacles over scenery), to reduce the computational load while maintaining accuracy for driving tasks.

## 4.3. Future Directions for LLM Applications

### **Ethical Considerations**

Ethical considerations involve evaluating the moral implications and societal impacts of deploying LLMs in traffic and transportation systems, focusing on issues like safety, fairness, privacy, and equity. These considerations address risks such as LLMs generating inaccurate outputs (e.g., misreading traffic signs), making life-or-death decisions in crash scenarios, or mishandling sensitive user data, ensuring that technology aligns with human values. Studying these ethical aspects offers significant advantages: it fosters the creation of safer, more reliable systems by identifying and mitigating risks like algorithmic bias, enhances public trust through transparency, and ensures compliance with legal standards, reducing liability while promoting equitable access for diverse populations, including those with disabilities or in underserved regions.

Throughout our literature search, we have not identified any paper that focuses on ethical considerations of applying LLMs in traffic and transportation research. Future research should prioritize developing ethical decision-making frameworks for LLMs in autonomous vehicles, particularly to address fairness and accountability in unavoidable crash scenarios, drawing inspiration from debates like the "trolley problem" in philosophy. Another critical direction is designing privacy-preserving techniques for managing sensitive transportation data, ensuring a balance between system utility and user rights, such as anonymizing location data while maintaining functionality. Additionally, studies should focus on reducing biases in LLMs to promote equity across diverse communities and investigate human-AI interaction to prevent over-reliance, ensuring drivers retain situational awareness and intervention capabilities in emergencies. These efforts will be essential for responsibly integrating LLMs into transportation systems.

### Cross-cultural Adaptability

Many current LLMs are developed and trained using datasets that are biased toward specific languages, cultures, or regions, leading to limited effectiveness in global or multicultural contexts. For example, traffic optimization systems may fail to account for regional driving behaviors, informal transportation networks, or cultural preferences for specific modes of travel, thereby reducing their applicability in diverse environments. Neglecting these differences can result in inequitable access to transportation benefits, diminished trust in AI systems, and a lack of global scalability for solutions. Addressing cross-cultural adaptability is vital to ensure inclusivity, fairness, and the success of AI-driven transportation systems in a globalized world.

Existing methods have applied LLMs directly for translation in emergency dispatch scenarios, which are considered indirectly related to traffic and transportation (Jiang, 2024; Otal and Canbaz, 2024). Applications that merit further research include real-time assistance systems for public transportation

in multilingual or low-literacy contexts, culturally sensitive hazard alert systems, and multimodal transportation integration in regions with informal systems like minibuses or tuk-tuks. Additionally, AV interfaces and virtual assistants must be adapted to accommodate diverse languages, accents, and communication norms. Research on sustainability applications, such as promoting eco-friendly transportation modes, is also critical to understanding how cultural values and attitudes toward the environment influence user behavior. These areas represent opportunities to design more inclusive and context-aware transportation systems.

To conduct relevant research, existing LLM methodologies can be leveraged by focusing on FT models with culturally diverse datasets that represent a range of languages, dialects, and regional contexts. Researchers can also employ transfer learning to adapt pre-trained models to specific cultural settings, ensuring the inclusion of underrepresented populations. Comparative studies can be conducted by deploying LLM-powered applications in different cultural contexts and analyzing variations in user interactions, preferences, and outcomes. Moreover, participatory approaches such as co-designing applications with local stakeholders can help identify cultural nuances and tailor solutions accordingly.

## Resilience and Adaptability

Resilience ensures that systems can withstand disruptions like data outages, unexpected traffic incidents, or evolving user needs, while adaptability enables systems to adjust to new conditions, such as emerging transportation technologies or policy changes. These challenges are frequently neglected because much research focuses on the immediate functionality or performance of LLMs without considering their robustness in dynamic real-world environments. However, transportation systems are inherently complex, involving unpredictable events, diverse user behaviors, and rapidly changing external factors. Failing to address resilience and adaptability can lead to unreliable systems, decreased trust among users, and limited scalability across different contexts or regions.

Existing research has studied resilience in the fields of traffic safety (Bäumler and Prokop, 2024; Wang, 2024), supply chain (Shahsavari et al., 2024) and ATC (Abdulhak et al., 2024). Future research can explore applications such as adaptive traffic management systems, which dynamically respond to unforeseen disruptions, or personalized navigation tools that adjust to changing user preferences and real-time data. Another promising area is emergency response planning, where resilient LLM-powered systems can process live data during disasters to facilitate evacuations. Additionally, public transportation could benefit from research on operation adaptability to predicted demand during unusual events, such as festivals or extreme weather. These applications can address critical needs in real-world transportation systems while improving their overall reliability and performance.

To conduct relevant research on these challenges, existing LLM methodologies can be adapted by incorporating real-time feedback loops, multi-modal training (e.g., integrating text, sensor, and geospatial data), and domain-specific fine-tuning. Researchers could simulate real-world disruptions, such as data outages or sudden traffic pattern changes, to test and improve system resilience. For adaptability, continual learning techniques can be employed, allowing LLMs to update their parameters as new data becomes available or as transportation systems evolve. Collaboration with transportation experts to curate high-quality datasets and define evaluation metrics for resilience and adaptability will ensure that the research aligns with practical needs.

### **Standard Metrics**

Quantitatively assessing the impact of LLMs requires the use of well-defined methodologies and key performance indicators. Current research employs a variety of datasets and evaluation metrics, often tailored to specific studies, making it difficult to compare the performance of different LLM-based models. For example, traffic prediction models can be assessed using error metrics like mean absolute error or root mean squared Error (Huang, 2024; Ren et al., 2024), while incident detection systems can be measured based on classification accuracy, precision, and recall (Zheng et al., 2023b). Each task requires scholars to carefully consider and assess evaluation criteria.

In addition to task-specific metrics, operational efficiency and economic impact are important considerations. LLMs can be evaluated based on their ability to reduce costs, such as operational expenses for traffic management systems, or to generate savings in fuel consumption and travel time through better route planning. For applications involving human interactions, usability metrics like user satisfaction scores and task completion rates provide insights into how effectively LLMs assist users in making decisions or accessing information.

Additionally, the dynamic nature of transportation systems complicates the establishment of consistent benchmarks, as real-world conditions often vary significantly. Long-term impacts, such as reductions in accidents, emissions, or system-wide costs, are harder to measure and require continuous monitoring to fully assess the effectiveness of LLM-powered solutions. Therefore, the quantitative evaluation of LLMs in transportation and traffic systems must consider a broad spectrum of metrics.

### **Urban Delivery**

In the realm of urban delivery and logistics, several pressing challenges persist, such as optimizing dynamic routing, managing customer interactions, and enhancing last-mile delivery efficiency. These challenges are compounded by the complexity of urban environments, making traditional solutions less effective and highlighting the need for innovative approaches.

To date, our literature search indicates that no studies have explored the application of LLMs to urban delivery, which has a few promising research directions. First, integrating LLMs with real-time spatial-temporal data offers a promising avenue for creating dynamic, context-aware routing systems, potentially cutting delivery times and environmental impact. Second, LLMs can process real-time traffic data, including traffic patterns, weather conditions, and local events, to assist in dynamic route

optimization. Third, research could investigate how LLMs might generate augmented instructions for last-mile delivery, aiding couriers in navigating urban complexities—such as apartment layouts or alternative drop-off points—while aligning with customer preferences, thus streamlining operations and enhancing delivery performance.

### Road Design

While we have identified papers that use multi-modal LLMs for generating road networks in recognized computer formats based on aerial images or even hand-drawn graphs (Rasal and Boddhu, 2024; Chen et al., 2024b), the design of roads remains an open area for exploration. Future research could extend the applications of LLMs beyond network generation to address more complex and context-sensitive road design challenges. Below, we outline several potential research directions:

- Economic and Demographic Factors-Based Road Design: LLMs could be developed to design roads based on a combination of economic and demographic factors. By integrating the relevant multimodal data, an LLM could propose road layouts optimized for cost-effectiveness and equitable access. For instance, an LLM could identify underserved areas and suggest new road designs that connect isolated populations to urban hubs, markets, or essential services. Integrating geospatial and economic data with the LLM reasoning capabilities could lead to more inclusive and economically sustainable road planning.
- Road Bottleneck Identification and Mitigation: Current research does not address the use of LLMs for identifying and mitigating road bottlenecks. Future work could explore how LLMs can analyze traffic flow data, accident reports, and congestion patterns to pinpoint bottlenecks in road networks. By combining this analysis with generative capabilities, LLMs could propose redesigns or expansions of problematic road segments to improve traffic efficiency. This could be particularly valuable for dynamic traffic management systems aiming to adapt road designs in real-time based on evolving traffic conditions.
- Integration of Environmental and Climate Factors: Road design often neglects the critical impact of environmental and climate factors. LLMs could integrate environmental data, such as flood zones, heat maps, and greenhouse gas emission levels, to propose climate-resilient road designs. For example, they could suggest elevated roads in flood-prone areas or optimize layouts to reduce vehicle emissions. By incorporating sustainability considerations, LLMs could contribute to the development of environmentally conscious transportation systems.
- Integration with Real-Time Traffic Systems: Road design is often treated as a static problem, but real-world traffic systems are highly dynamic. Future LLMs could integrate with real-time traffic monitoring systems to propose adaptive road designs and layouts. For example, they could suggest temporary road expansions, one-way systems, or dedicated bus lanes during peak hours, providing a flexible approach to urban traffic management.

- Exploration of Lightweight and Cost-Effective Models: Building on the lightweight training methods demonstrated in Rasal and Boddhu, 2024, future research could focus on developing low-resource LLMs for road design. This would make advanced road planning tools accessible to low-income regions or smaller municipalities that lack the computational resources to deploy large-scale AI systems.
- Multimodal Road Design Optimization: Current LLM applications focus on generating networks. Future research could explore optimizing road designs for multimodal transportation systems. This could include integrating data from public transit networks, freight logistics, and pedestrian pathways to propose road layouts that balance the needs of different transportation modes. LLMs could also consider the trade-offs between infrastructure costs, travel times, and environmental impacts.

### End-to-end Autonomous Vehicle

The integration of LLMs into end-to-end autonomous driving frameworks has attracted significant attention due to their ability to unify perception, prediction, and planning into a single model. Inspired by the research of Zhu et al., 2024, we propose two research directions where LLMs can play a role in end-to-end autonomous driving.

First, end-to-end AD frameworks, such as Tesla's full self-driving (FSD), rely heavily on diverse and high-quality datasets to train models capable of directly generating motion plans. However, the collection and annotation of such data are resource-intensive, with rare and critical events—like extreme weather conditions or sudden pedestrian crossings—often underrepresented in datasets. To address this, strategies such as simulation-based data augmentation and active learning can improve dataset efficiency while reducing the dependency on costly real-world data collection. Nonetheless, balancing cost, scenario diversity, and realism remains an ongoing challenge that needs further exploration.

Second, interpretability is a crucial factor for establishing trust in end-to-end AD systems, as their black-box nature limits transparency in decision-making. Techniques such as attention mechanisms and explainable AI (XAI) tools offer ways to shed light on the reasoning behind model predictions, while intermediate representations can enhance transparency by breaking down the decision-making process into interpretable stages. Recent advancements, such as vision-language planning (VLP), leverage LLMs to generate natural language explanations for motion planning decisions, providing an additional layer of interpretability. However, striking a balance between interpretability and system performance continues to be a significant challenge (Pan et al., 2024a).

#### Interaction with Other Road Users

In autonomous driving, effectively managing interactions with road users like pedestrians, cyclists, and other vehicles is a pivotal challenge that directly influences safety and traffic flow, such as interpreting a pedestrian's wave to cross, understanding a car's flashing lights signaling a turn, or reacting

to a cyclist's sudden swerve. These issues are critical because AVs must quickly and accurately decipher these often non-verbal, dynamic cues to avoid collisions and ensure smooth navigation.

Although there are a number of research outputs that analyze trajectories of other road users (Lan et al., 2024; Wang et al., 2024i), research on finer granularity remains absent. Future research could significantly advance this field by integrating LLMs with multimodal systems, such as combining them with computer vision to simultaneously analyze verbal warnings and visual gestures, thereby building a richer understanding of road user intentions. Another promising direction involves training LLMs on contextual and historical data to predict behaviors – like anticipating a pedestrian's crossing based on subtle cues or past patterns – enhancing proactive decision-making. Additionally, equipping LLMs to enable vehicles to communicate outwardly, such as broadcasting a message to yield or acknowledging a cyclist's signal via text or sound, could foster clearer, safer interactions in shared traffic spaces, ultimately elevating the adaptability and trustworthiness of AD systems.

## Travel Behavior Modeling

In the research area of travel behavior, LLMs complement traditional methods by excelling at processing vast, unstructured datasets—think social media posts or open-ended survey responses—where conventional models falter. Their strength lies in uncovering nuanced patterns and adapting to multi-modal data, offering a richer understanding of travel behavior. For instance, studies like Ruan et al., 2024 showcase LLMs predicting travel mode choices from social media, while Chen et al., 2024a and Shao et al., 2024 leverage them for public transit and historical data analysis using CoT prompting.

Despite these advancements, several research gaps remain that future studies should address to enhance LLMs' utility in travel behavior modeling. One critical gap is the limited integration of real-time, multi-modal data. Current models excel with historical or static datasets but struggle to adapt to dynamic urban conditions, such as sudden traffic incidents or weather shifts. Future research should focus on developing LLMs that fuse live traffic feeds, weather data, and visual inputs (e.g., from traffic cameras) using techniques like continual learning and transfer learning. For example, in metro delay scenarios, an LLM could combine real-time delay logs, passenger sentiment from social media, and station camera feeds to predict adaptive travel choices, improving upon DelayPTC-LLM's (Chen et al., 2024a) static approach. Similarly, for pedestrian safety, integrating live sensor data from smart crossings with LLM reasoning could enhance real-time collision risk predictions, addressing limitations in the simulations of Radford et al., 2021. Another gap is explainability, particularly in safety-critical applications. Enhancing methods like attention visualization or counterfactual reasoning could make LLM predictions—such as a driver's lane-changing decision or a pedestrian's crossing intent—more transparent, aiding AV decision-making and urban planning.

# 4.4. Strengths and weaknesses of LLMs

By leveraging pre-trained capability and adaptability, LLMs bring significant advantages to traffic and transportation research. The following advantages across different applications to traffic and transportation applications.

Versatility with Pre-Trained Models: LLMs offer significant versatility in traffic and transportation applications due to their pre-trained capabilities. These models can be employed in various scenarios with ZSL or FSL, meaning they can perform tasks without needing extensive retraining on new data. For instance, an LLM can be used to analyze traffic reports, understand road conditions, or even generate summaries of transportation policies without specific training for each task. This flexibility is particularly beneficial in the dynamic field of transportation, where new challenges and data types, such as real-time traffic updates or incident reports, emerge frequently.

Adaptability through Fine-Tuning: Fine-tuning LLMs allows them to be tailored to specific transportation contexts, enhancing their performance in those areas. For example, a transportation agency could fine-tune an LLM on historical traffic data to better predict congestion patterns or optimize public transit schedules. This adaptability ensures that the model is not only general but also highly effective in specific, localized scenarios, such as managing traffic in a particular city or region. Studies have highlighted the effectiveness of fine-tuning in tasks like route optimization, where LLMs can learn from local traffic patterns to improve decision-making, aligning with the need for precision in urban transportation systems.

Multimodal Integration: LLMs can integrate with visual data, providing a more comprehensive understanding of transportation scenarios that involve both text and images. In transportation, this could mean analyzing images from traffic cameras alongside textual reports to better understand and respond to traffic incidents or road conditions. Visual language models, which combine the capabilities of LLMs with image processing, can help in tasks like automatic incident detection or monitoring road maintenance needs. Recent research, such as the use of VAD-LLaMA for traffic anomaly detection, demonstrates how LLMs with visual data can enhance real-time responsiveness in autonomous driving systems, improving safety and efficiency.

Handling Unforeseen and Long-Tail Scenarios: The broad knowledge base of LLMs enables them to understand and respond to unusual or rare situations in transportation. Whether it's dealing with extreme weather conditions, unusual traffic patterns due to events, or new types of vehicles, LLMs can provide insights or predictions based on their general understanding of language and world knowledge. This is crucial for transportation systems to be resilient and adaptive to unpredictable events, such as sudden road closures or emergency evacuations. Their ability to generate unseen scenarios also supports scenario planning, as seen in studies using LLMs for traffic management at urban intersections, where they handle mixed traffic conditions effectively.

Generating Synthetic Datasets: LLMs can create artificial data for training other models or for simulation purposes, which is particularly useful in transportation where real data might be limited or sensitive. For example, synthetic data can be used to train models for rare events like accidents or to simulate different traffic scenarios for testing new control algorithms. This capability helps improve the robustness and safety of transportation systems without relying solely on real-world data, which might be scarce or biased. Research has explored LLMs generating synthetic datasets for traffic flow forecasting, enhancing model training in data-scarce environments.

Language Translation and Adaptation: Transportation is a global issue, and different regions have different languages. LLMs can translate information or adapt to different language contexts, which is helpful for international transportation systems or for providing multilingual information to users. This ensures that transportation information and services are accessible to a diverse range of users, enhancing inclusivity and efficiency in global transportation networks.

Rich Embeddings: The embeddings from LLMs contain nuanced information that traditional NLP methods might not capture, providing deeper insights into transportation data. For instance, these embeddings can help understand the sentiment of public feedback on transportation services or in identifying patterns in traffic-related text data that are not immediately obvious. This can lead to better decision-making and more effective communication strategies in transportation management, such as analyzing social media for real-time traffic sentiment, which traditional NLP might overlook.

Apart from the above strengths, LLMs do have weaknesses. Addressing the weaknesses is essential for ethical and effective use of LLMs. However, this will require further advancements in LLM development and training processes. The major weaknesses include:

High Computational Requirements: Training and running LLMs require significant computational power, which can be a barrier for some transportation agencies, especially those with limited budgets or in regions with poor Internet connectivity. Real-time applications in transportation, such as traffic signal control or dynamic routing, demand low latency and high processing speeds, which LLMs may not always meet efficiently. This can limit their practical implementation in time-critical transportation systems, particularly in developing regions where resources are constrained.

Privacy: Privacy is another critical concern when using LLMs in transportation, as detailed in Das et al., 2025. These models are trained on vast amounts of data, which may include sensitive information about individuals' travel habits, such as commuting patterns or frequent destinations, raising risks of personally identifiable information (PII) leakage. There is a risk that LLMs could inadvertently reveal or misuse this data, leading to privacy breaches, especially in cloud-based systems where data might be shared with third parties. Transportation agencies must ensure that any data used in training or fine-tuning LLMs is handled in accordance with privacy regulations, such as GDPR, and steps are taken to prevent data leakage or unauthorized access, which is crucial for maintaining public trust and compliance.

AI Hallucinations: LLMs can produce outputs that are factually incorrect or entirely made up, leading to misinformation or wrong decisions, such as suggesting non-existent road closures or incorrect

detour routes. In transportation, such errors could have serious consequences, from misdirecting traffic to providing inaccurate safety information, potentially endangering lives.

**Up-to-date Data**: LLMs are usually trained on historical data, which may not reflect current traffic conditions, new infrastructure like recently built highways, or regulatory changes such as updated speed limits. LLMs risk providing outdated or irrelevant recommendations, which could compromise transportation efficiency and safety, as seen in Zhang et al., 2024f.

Lack of Uniform Benchmark: The lack of a uniform benchmark for evaluating LLMs in transportation, especially in generating unseen scenarios and synthetic datasets, is a significant hurdle, as highlighted in Syed et al., 2024. Without standardized evaluation methods, it is difficult to compare the performance of different LLMs or to assess how they stack up against traditional approaches based on human expert-defined rules, such as traffic simulation models. Developing such benchmarks is crucial to drive progress in this area and ensure that LLMs are effectively enhancing transportation systems, particularly for tasks like predicting rare events or creating training data for other AI models.

Numerical Ability: While these models excel in processing and generating text, their performance with numerical data, such as traffic counts, travel times, or fuel consumption estimates, may not be as robust as that of traditional numerical models or specialized AI models designed for such tasks. For instance, predicting exact traffic volumes during peak hours or predicting demand for supply chain applications requires precise numerical reasoning, which LLMs might struggle with compared to statistical models.

Trustworthiness and Safety: LLMs, with their probabilistic outputs and potential for errors or biases, may not always meet the stringent safety requirements of such systems, as they can hallucinate or provide inconsistent results under varying conditions. Besides, LLMs are vulnerable to prompt injection, jailbreak attacks, and data poisoning, which can manipulate their responses or make them reveal sensitive information (Yao et al., 2024b).

#### 5. Conclusion

In this paper, we offer a comprehensive overview of LLM methodologies and their applications within traffic and transportation research. We categorize these studies by their specific use cases – ranging from urban logistics to AD interactions – while critically examining the associated challenges. LLMs emerge as potent tools for overcoming the shortcomings of conventional approaches, excelling in managing long-tail scenarios, integrating text with diverse data types, and enhancing system adaptability. Despite these advancements, substantial research opportunities persist, both in refining the necessary methodologies and expanding their practical applications, promising further innovation in this rapidly evolving field.

# References

- Abdulhak, Sinan, Wayne Hubbard, Karthik Gopalakrishnan and Max Z Li (2024). "CHATATC: Large Language Model-Driven Conversational Agents for Supporting Strategic Air Traffic Flow Management". In: arXiv preprint arXiv:2402.14850.
- Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. (2023). "Gpt-4 technical report". In: arXiv preprint arXiv:2303.08774.
- Aguero, David and Scott D Nelson (2024). "The potential application of large language models in pharmaceutical supply chain management". In: *The Journal of Pediatric Pharmacology and Therapeutics* 29.2, pp. 200–205.
- Ainslie, Joshua, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón and Sumit Sanghai (2023). "Gqa: Training generalized multi-query transformer models from multi-head check-points". In: arXiv preprint arXiv:2305.13245.
- Aksjonov, Andrei and Ville Kyrki (2021). "Rule-based decision-making system for autonomous vehicles at intersections with mixed traffic environment". In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). IEEE, pp. 660–666.
- Ali, Farman, Amjad Ali, Muhammad Imran, Rizwan Ali Naqvi, Muhammad Hameed Siddiqi and Kyung-Sup Kwak (2021). "Traffic accident detection and condition analysis based on social networking data". In: Accident Analysis & Prevention 151, p. 105973.
- AlMahri, Sara, Liming Xu and Alexandra Brintrup (2024). "Enhancing Supply Chain Visibility with Knowledge Graphs and Large Language Models". In: arXiv preprint arXiv:2408.07705.
- Anil, Rohan, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. (2023). "Palm 2 technical report". In: arXiv preprint arXiv:2305.10403.
- Azarafza, Mehdi, Mojtaba Nayyeri, Charles Steinmetz, Steffen Staab and Achim Rettberg (2024). "Hybrid Reasoning Based on Large Language Models for Autonomous Car Driving". In: arXiv preprint arXiv:2402.13602.
- Balani, Navveen (2024). Future of Large Language Models: Generalized, Specialized, and Orchestrator Models. URL: https://navveenbalani.medium.com/future-of-large-language-models-generalized-specialized-and-orchestrator-models-c229aa60f593.
- Bañón, Marta, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, et al. (2020). "ParaCrawl: Web-scale acquisition of parallel corpora". In: Association for Computational Linguistics (ACL).
- Bäumler, Maximilian and Günther Prokop (2024). "Predicting the type of road traffic accident for test scenario generation". In: *IEEE Access*.

Systems.

- Besta, Maciej, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. (2024). "Graph of thoughts: Solving elaborate problems with large language models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 16, pp. 17682–17690.
- Bierlaire, Michel (1998). "Discrete choice models". In: Operations research and decision aid methodologies in traffic and transportation management. Springer, pp. 203–227.
- Bowman, John L and Moshe E Ben-Akiva (2001). "Activity-based disaggregate travel demand model system with activity schedules". In: *Transportation research part a: policy and practice* 35.1, pp. 1–28.
- Brown, Tom B (2020). "Language models are few-shot learners". In: arXiv preprint arXiv:2005.14165. Cao, Yuji, Huan Zhao, Yuheng Cheng, Ting Shu, Yue Chen, Guolong Liu, Gaoqi Liang, Junhua Zhao, Jinyue Yan and Yun Li (2024). "Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods". In: IEEE Transactions on Neural Networks and Learning
- Carranza, Jose Martinez, Delia Irazu Hernandez-Farias, Leticia Oyuki Rojas-Perez and Aldrich Alfredo Cabrera Ponce (2023). "Why do I need to speak to my drone?" In: 14th ANNUAL INTERNATIONAL MICRO AIR VEHICLE CONFERENCE AND COMPETITION, IMAV2023-7.
- Chan, Jun Shern, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng and Aleksander Mądry (2024). MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. arXiv: 2410.07095 [cs.CL]. URL: https://arxiv.org/abs/2410.07095.
- Chang, Cheng, Siqi Wang, Jiawei Zhang, Jingwei Ge and Li Li (2024). "LLMScenario: Large Language Model Driven Scenario Generation". In: *IEEE Transactions on Systems*, Man, and Cybernetics: Systems, pp. 1–14. DOI: 10.1109/TSMC.2024.3392930.
- Chen, Chen, Yuxin He, Hao Wang, Jingjing Chen and Qin Luo (2024a). "Delayptc-llm: Metro passenger travel choice prediction under train delays with large language models". In: arXiv preprint arXiv:2410.00052.
- Chen, Jiajing, Weihang Xu, Haiming Cao, Zihuan Xu, Yu Zhang, Zhao Zhang and Siyao Zhang (2024b). "Multimodal Road Network Generation Based on Large Language Model". In: arXiv preprint arXiv:2404.06227.
- Chen, Long, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund and Jamie Shotton (2024c). "Driving with llms: Fusing object-level vector modality for explainable autonomous driving". In: 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 14093–14100.
- Chen, Mark, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. (2021). "Evaluating large language models trained on code". In: arXiv preprint arXiv:2107.03374.

- Chen, Minze, Zhenxiang Tao, Weitong Tang, Tingxin Qin, Rui Yang and Chunli Zhu (Nov. 2023). Enhancing Emergency Decision-making with Knowledge Graphs and Large Language Models. arXiv: 2311.08732 [cs]. (Visited on 08/29/2024).
- Chen, Qiang, Hung-Cheng Chen and Yu-Liang Lin (Mar. 2024d). "ChatGPT-powered Inquiry-based Learning Model of Training for Intelligent Car Racing Competition". In: Sensors and Materials 36.3, p. 1147. ISSN: 0914-4935, 2435-0869. DOI: 10.18494/SAM4726. (Visited on 09/12/2024).
- Chen, Xianda, Mingxing Peng, PakHin Tiu, Yuanfei Wu, Junjie Chen, Meixin Zhu and Xinhu Zheng (July 2024e). GenFollower: Enhancing Car-Following Prediction with Large Language Models. (Visited on 09/10/2024).
- Chen, Yongchao, Jacob Arkin, Yang Zhang, Nicholas Roy and Chuchu Fan (2024f). "Scalable multirobot collaboration with large language models: Centralized or decentralized systems?" In: 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 4311–4317.
- Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. (2023). "Palm: Scaling language modeling with pathways". In: *Journal of Machine Learning Research* 24.240, pp. 1–113.
- Chu, Kai-Fung, Albert YS Lam and Victor OK Li (2021). "Traffic signal control using end-to-end off-policy deep reinforcement learning". In: *IEEE Transactions on Intelligent Transportation Systems* 23.7, pp. 7184–7195.
- Conover, Mike, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia and Reynold Xin (2023). "Free dolly: Introducing the world's first truly open instruction-tuned llm". In: Company Blog of Databricks.
- Cui, Can, Yunsheng Ma, Xu Cao, Wenqian Ye and Ziran Wang (2024a). "Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 902–909.
- Cui, Can, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. (2024b). "A survey on multimodal large language models for autonomous driving". In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 958–979.
- Cui, Yaodong, Shucheng Huang, Jiaming Zhong, Zhenan Liu, Yutong Wang, Chen Sun, Bai Li, Xiao Wang and Amir Khajepour (2023). "Drivellm: Charting the path toward full autonomous driving with large language models". In: *IEEE Transactions on Intelligent Vehicles*.
- Da, Longchao, Kuanru Liou, Tiejin Chen, Xuesong Zhou, Xiangyong Luo, Yezhou Yang and Hua Wei (2024). "Open-ti: Open traffic intelligence with augmented language model". In: *International Journal of Machine Learning and Cybernetics*, pp. 1–26.

- Dai, Zhuang, Xiaoyue Cathy Liu, Xi Chen and Xiaolei Ma (2020). "Joint optimization of scheduling and capacity for mixed traffic with autonomous and human-driven buses: A dynamic programming approach". In: *Transportation Research Part C: Emerging Technologies* 114, pp. 598–619.
- Das, Badhan Chandra, M Hadi Amini and Yanzhao Wu (2025). "Security and privacy challenges of large language models: A survey". In: *ACM Computing Surveys* 57.6, pp. 1–39.
- Das, Subasish, Amir Hossein Oliaee, Minh Le, Michael P. Pratt and Jason Wu (July 2023). "Classifying Pedestrian Maneuver Types Using the Advanced Language Model". In: *Transportation Research Record: Journal of the Transportation Research Board* 2677.7, pp. 599–611. ISSN: 0361-1981, 2169-4052. DOI: 10.1177/03611981231155187. (Visited on 10/03/2024).
- De Curtò, J, I De Zarza and Carlos T Calafate (2023). "Semantic scene understanding with large language models on unmanned aerial vehicles". In: *Drones* 7.2, p. 114.
- Devlin, Jacob (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: arXiv preprint arXiv:1810.04805.
- Dhara, Sasank and Sebastian Delgado Barba (July 2024). "Large Language Models in Supply Chain Management". Advisor: Prof. Nizar Abdelkafi. Tesi di Laurea Magistrale in Management Engineering. Milan, Italy: Politecnico di Milano.
- Ding, Ning, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun and Bowen Zhou (2023). "Enhancing chat language models by scaling high-quality instructional conversations". In: arXiv preprint arXiv:2305.14233.
- Dong, Qingxiu, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. (2022). "A survey on in-context learning". In: arXiv preprint arXiv:2301.00234.
- Dong, Qingxiu, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li and Zhifang Sui (2024). A Survey on In-context Learning. arXiv: 2301.00234 [cs.CL]. URL: https://arxiv.org/abs/2301.00234.
- Dubois, Yann, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang and Tatsunori B Hashimoto (2024). "Alpacafarm: A simulation framework for methods that learn from human feedback". In: *Advances in Neural Information Processing Systems* 36.
- Essien, Aniekan, Ilias Petrounias, Pedro Sampaio and Sandra Sampaio (2021). "A deep-learning model for urban traffic flow prediction with traffic events mined from twitter". In: World Wide Web 24.4, pp. 1345–1368.
- Fan, Zhiwen, Pu Wang, Yang Zhao, Yibo Zhao, Boris Ivanovic, Zhangyang Wang, Marco Pavone and Hao Frank Yang (2024). "Learning Traffic Crashes as Language: Datasets, Benchmarks, and What-if Causal Analyses". In: arXiv preprint arXiv:2406.10789. DOI: 10.48550/arXiv.2406.10789.
- Feng, Chao, Xinyu Zhang and Zichu Fei (2023). "Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs". In: arXiv preprint arXiv:2309.03118.

- Feng, Guhao, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He and Liwei Wang (2024). "Towards revealing the mystery behind chain of thought: a theoretical perspective". In: Advances in Neural Information Processing Systems 36.
- Fox, Kevin L, Kevin R Niewoehner, Mark Rahmes, Josiah Wong and Rahul Razdan (2024). "Leverage Large Language Models For Enhanced Aviation Safety". In: 2024 Integrated Communications, Navigation and Surveillance Conference (ICNS). IEEE, pp. 1–11.
- Fu, Daocheng, Wenjie Lei, Licheng Wen, Pinlong Cai, Song Mao, Min Dou, Botian Shi and Yu Qiao (2024a). "LimSim++: A Closed-Loop Platform for Deploying Multimodal LLMs in Autonomous Driving". In: arXiv preprint arXiv:2402.01246.
- Fu, Daocheng, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi and Yu Qiao (2024b). "Drive like a human: Rethinking autonomous driving with large language models". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 910–919.
- Fu, Yao, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng and Tushar Khot (2023). "Chain-of-Thought Hub: A Continuous Effort to Measure Large Language Models' Reasoning Performance". In: arXiv preprint arXiv:2305.17306.
- Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun and Haofen Wang (2023). "Retrieval-augmented generation for large language models: A survey". In: arXiv preprint arXiv:2312.10997.
- Gebre, Tewodros Syum, Leila Hashemi-Beni, Eden Tsehaye Wasehun and Freda Elikem Dorbu (2024). "AI-Integrated Traffic Information System: A Synergistic Approach of Physics Informed Neural Network and GPT-4 for Traffic Estimation and Real-Time Assistance". In: *IEEE Access*.
- Gianpaolo, Ghiani, Laporte Gilbert and Musmanno Roberto (2013). Introduction to Logistics Systems Management.
- Gregurić, Martin, Miroslav Vujić, Charalampos Alexopoulos and Mladen Miletić (2020). "Application of deep reinforcement learning in traffic signal control: An overview and impact of open traffic data". In: *Applied Sciences* 10.11, p. 4011.
- Guanetti, Jacopo, Yeojun Kim and Francesco Borrelli (2018). "Control of connected and automated vehicles: State of the art and future challenges". In: *Annual reviews in control* 45, pp. 18–40.
- Guo, Chuan, Geoff Pleiss, Yu Sun and Kilian Q. Weinberger (Aug. 2017). "On Calibration of Modern Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1321–1330. URL: https://proceedings.mlr.press/v70/guo17a.html.
- Guo, Xusen, Qiming Zhang, Mingxing Peng, Meixin Zhua, et al. (2024). "Explainable Traffic Flow Prediction with Large Language Models". In: arXiv preprint arXiv:2404.02937.
- Gupta, Prakhar, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi and Jeffrey P Bigham (2022). "InstructDial: Improving zero and few-shot generalization in dialogue through instruction tuning". In: arXiv preprint arXiv:2205.12673.

- Güzay, Çağrı, Ege Özdemir and Yahya Kara (2023). "A Generative AI-driven Application: Use of Large Language Models for Traffic Scenario Generation". In: 2023 14th International Conference on Electrical and Electronics Engineering (ELECO). IEEE, pp. 1–6.
- Hadi, Muhammad Usman, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. (2023). "A survey on large language models: Applications, challenges, limitations, and practical usage". In: *Authorea Preprints*.
- He, Jiabang, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu and Heng Tao Shen (2023). "Icl-d3ie: In-context learning with diverse demonstrations updating for document information exli2024automatedli2024automatedli2024automatedtraction". In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 19485–19494.
- Health, Stanford Medicine Children's (2025). Age-Appropriate Speech and Language Milestones. Accessed: 2025-03-07. URL: https://www.stanfordchildrens.org/en/topic/default?id=age-appropriate-speech-and-language-milestones-90-P02170.
- Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan and Sylvain Gelly (2019). "Parameter-efficient transfer learning for NLP". In: *International conference on machine learning*. PMLR, pp. 2790–2799.
- Hu, Edward J, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang and Weizhu Chen (2021). "Lora: Low-rank adaptation of large language models". In: arXiv preprint arXiv:2106.09685.
- Hu, Yaqi, Dongyuan Ou, Xiaoxu Wang and Rong Yu (2023). "Enabling Vision-and-Language Navigation for Intelligent Connected Vehicles Using Large Pre-Trained Models". In: 2023 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics). IEEE, pp. 390–396.
- Huang, Chenyu, Zhengyang Tang, Dongdong Ge, Shixi Hu, Ruoqing Jiang, Benyou Wang, Zizhuo Wang and Xin Zheng (2024a). "ORLM: A Customizable Framework in Training Large Models for Automated Optimization Modeling". In: arXiv e-prints, arXiv-2405.
- Huang, Ping, Yuxin He, Hao Wang, Jingjing Chen and Qin Luo (2024b). "A Prompt Refinement-based Large Language Model for Metro Passenger Flow Forecasting under Delay Conditions". In: arXiv preprint arXiv:2410.15111.
- Huang, Xiannan (2024). "Enhancing Traffic Prediction with Textual Data Using Large Language Models". In: arXiv preprint arXiv:2405.06719.
- Huang, Yue, Qihui Zhang, Lichao Sun, et al. (2023a). "Trustgpt: A benchmark for trustworthy and responsible large language models". In: arXiv preprint arXiv:2306.11507.
- Huang, Yunpeng, Jingwei Xu, Junyu Lai, Zixu Jiang, Taolue Chen, Zenan Li, Yuan Yao, Xiaoxing Ma, Lijuan Yang, Hao Chen, et al. (2023b). "Advancing transformer architecture in long-context large language models: A comprehensive survey". In: arXiv preprint arXiv:2311.12351.

- Hussien, Mohamed Manzour, Angie Nataly Melo, Augusto Luis Ballardini, Carlota Salinas Maldonado, Rubén Izquierdo and Miguel Angel Sotelo (2025). "Rag-based explainable prediction of road users behaviors for automated driving using knowledge graphs and large language models". In: Expert Systems with Applications 265, p. 125914.
- Jackson, Ilya, Maria Jesus Saenz and Dmitry Ivanov (2024). "From natural language to simulations: applying AI to automate simulation modelling of logistics systems". In: *International Journal of Production Research* 62.4, pp. 1434–1457.
- Jarry, Gabriel, Philippe Very and Ramon Dalmau (2024). "The Effectiveness of Large Language Models for Textual Analysis in Air Transportation". In: EasyChair preprints.
- Ji, Jiaming, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang and Yaodong Yang (2024). "Beavertails: Towards improved safety alignment of llm via a human-preference dataset". In: Advances in Neural Information Processing Systems 36.
- Jiang, Albert Q, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. (2023). "Mistral 7B". In: arXiv preprint arXiv:2310.06825.
- Jiang, Yue (May 2024). "The Applications of Large Language Models in Emergency Management".
  In: 2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). Chongqing, China: IEEE, pp. 199–202. ISBN: 9798350316537. DOI: 10.1109/IMCEC59810.2024.10575031. (Visited on 08/29/2024).
- Jimenez, Carlos E, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press and Karthik R Narasimhan (2024). "SWE-bench: Can Language Models Resolve Real-world Github Issues?" In: The Twelfth International Conference on Learning Representations. URL: https://openreview.net/forum?id=VTF8yNQM66.
- Jin, Can, Hongwu Peng, Shiyu Zhao, Zhenting Wang, Wujiang Xu, Ligong Han, Jiahui Zhao, Kai Zhong, Sanguthevar Rajasekaran and Dimitris N Metaxas (2024). "APEER: Automatic Prompt Engineering Enhances Large Language Model Reranking". In: arXiv preprint arXiv:2406.14449.
- Jin, KyoHoon, JeongA Wi, EunJu Lee, ShinJin Kang, SooKyun Kim and YoungBin Kim (2021). "TrafficBERT: Pre-trained model with large-scale data for long-range traffic flow forecasting". In: Expert Systems with Applications 186, p. 115738.
- Jurafsky, Daniel and James H. Martin (2025). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models. 3rd. Online manuscript released January 12, 2025. URL: https://web.stanford.edu/~jurafsky/slp3/.
- Kalyan, Katikapalli Subramanyam, Ajit Rajasekharan and Sivanesan Sangeetha (2021). "Ammus: A survey of transformer-based pretrained models in natural language processing". In: arXiv preprint arXiv:2108.05542.

- Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu and Dario Amodei (2020). "Scaling laws for neural language models". In: arXiv preprint arXiv:2001.08361.
- Kocetkov, Denis, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, et al. (2022). "The stack: 3 tb of permissively licensed source code". In: arXiv preprint arXiv:2211.15533.
- Koushik, Anil NP, M Manoj and N Nezamuddin (2020). "Machine learning applications in activity-travel behaviour research: a review". In: *Transport reviews* 40.3, pp. 288–311.
- Kuo, Yong-Hong, Janny MY Leung and Yimo Yan (2023). "Public transport for smart cities: Recent innovations and future challenges". In: European Journal of Operational Research 306.3, pp. 1001– 1026.
- Lai, Siqi, Zhao Xu, Weijia Zhang, Hao Liu and Hui Xiong (2023). "Large language models as traffic signal control agents: Capacity and opportunity". In: arXiv preprint arXiv:2312.16044.
- Lan, Zhengxing, Lingshan Liu, Bo Fan, Yisheng Lv, Yilong Ren and Zhiyong Cui (2024). "Traj-llm: A new exploration for empowering trajectory prediction with pre-trained large language models". In: *IEEE Transactions on Intelligent Vehicles*.
- Le Scao, Teven, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. (2023). "Bloom: A 176b-parameter open-access multilingual language model". In.
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov and Luke Zettlemoyer (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv: 1910.13461 [cs.CL]. URL: https://arxiv.org/abs/1910.13461.
- Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. (2020). "Retrieval-augmented generation for knowledge-intensive nlp tasks". In: Advances in Neural Information Processing Systems 33, pp. 9459–9474.
- Li, Beibin, Konstantina Mellou, Bo Zhang, Jeevan Pathuri and Ishai Menache (2023). "Large language models for supply chain optimization". In: arXiv preprint arXiv:2307.03875.
- Li, Dun, Huan Wang and Yan-Fu Li (2024a). "Robust Anomaly Detection In Unmanned Ship Systems Based On Large Language Models". In.
- Li, Yanze, Wenhua Zhang, Kai Chen, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. (2024b). "Automated evaluation of large vision-language models on self-driving corner cases". In: arXiv preprint arXiv:2404.10595.
- Li, Yihao, Ru Zhang and Jianyi Liu (2024c). "An enhanced prompt-based LLM reasoning scheme via knowledge graph-integrated collaboration". In: *International Conference on Artificial Neural Networks*. Springer, pp. 251–265.

- Li, Yuchen, Luxi Li, Zizhang Wu, Zhenshan Bing, Yunfeng Ai, Bin Tian, Zhe Xuanyuan, Alois Christian Knoll and Long Chen (2024d). "Miningllm: Towards mining 5.0 via large language models in autonomous driving and smart mining". In: *IEEE Transactions on Intelligent Vehicles*.
- Li, Yuchen, Luxi Li, Zizhang Wu, Zhenshan Bing, Zhe Xuanyuan, Alois Christian Knoll and Long Chen (2024e). "UnstrPrompt: Large Language Model Prompt for Driving in Unstructured Scenarios". In: *IEEE Journal of Radio Frequency Identification* 8, pp. 367–375. ISSN: 2469-7281, 2469-729X. DOI: 10.1109/JRFID.2024.3367975. (Visited on 09/10/2024).
- Li, Yun, Kai Katsumata, Ehsan Javanmardi and Manabu Tsukada (2024f). "Large Language Models for Human-like Autonomous Driving: A Survey". In: arXiv preprint arXiv:2407.19280.
- Liang, Yicong, Di Zou, Haoran Xie and Fu Lee Wang (2023). "Exploring the potential of using ChatGPT in physics education". In: Smart Learning Environments 10.1, p. 52.
- Liang, Yuebing, Yichao Liu, Xiaohan Wang and Zhan Zhao (Sept. 2024). "Exploring Large Language Models for Human Mobility Prediction under Public Events". In: Computers, Environment and Urban Systems 112, 102153Provide the following details only based on this paper: Model backbone, Fine tuning, Input, Output, Computation Resources, Modality, Task, Learning, Data Source, lessons learned Please be concise. ISSN: 01989715. DOI: 10.1016/j.compenvurbsys.2024.102153. (Visited on 09/05/2024).
- Liu, Aixin, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. (2024a). "Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model". In: arXiv preprint arXiv:2405.04434.
- Liu, Chenxi, Sun Yang, Qianxiong Xu, Zhishuai Li, Cheng Long, Ziyue Li and Rui Zhao (2024b). "Spatial-temporal large language model for traffic prediction". In: arXiv preprint arXiv:2401.10134.
- Liu, Fei, Xialiang Tong, Mingxuan Yuan, Xi Lin, Fu Luo, Zhenkun Wang, Zhichao Lu and Qingfu Zhang (2024c). "Evolution of heuristics: Towards efficient automatic algorithm design using large language model". In: arXiv preprint arXiv:2401.02051.
- Liu, Tianming, Manzi Li and Yafeng Yin (2024d). "Can Large Language Models Capture Human Travel Behavior? Evidence and Insights on Mode Choice". In: Evidence and Insights on Mode Choice (August 26, 2024).
- Liu, Tong and Hadi Meidani (2024a). "End-to-end heterogeneous graph neural networks for traffic assignment". In: *Transportation Research Part C: Emerging Technologies* 165, p. 104695.
- (2024b). "Heterogeneous Graph Sequence Neural Networks for Dynamic Traffic Assignment". In: arXiv preprint arXiv:2408.04131.
- Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: arXiv preprint arXiv:1907.11692.

- Liu, Yuan, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. (2023). "Mmbench: Is your multi-modal model an all-around player?" In: arXiv preprint arXiv:2307.06281.
- Longpre, Shayne, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. (2023). "The flan collection: Designing data and methods for effective instruction tuning". In: *International Conference on Machine Learning*. PMLR, pp. 22631–22648.
- Lopez, Pablo Alvarez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner and Evamarie Wießner (2018). "Microscopic Traffic Simulation using SUMO". In: The 21st IEEE International Conference on Intelligent Transportation Systems. IEEE. URL: https://elib.dlr.de/124092/.
- Lu, Pan, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark and Ashwin Kalyan (2022). "Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning". In: arXiv preprint arXiv:2209.14610.
- Lu, Shuai, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, et al. (2021a). "Codexglue: A machine learning benchmark dataset for code understanding and generation". In: arXiv preprint arXiv:2102.04664.
- Lu, Yao, Max Bartolo, Alastair Moore, Sebastian Riedel and Pontus Stenetorp (2021b). "Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity". In: arXiv preprint arXiv:2104.08786.
- Ma, Xiaolei, Zhimin Tao, Yinhai Wang, Haiyang Yu and Yunpeng Wang (2015). "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data". In: *Transportation Research Part C: Emerging Technologies* 54, pp. 187–197.
- Ma, Yecheng Jason, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan and Anima Anandkumar (2023). "Eureka: Human-level reward design via coding large language models". In: arXiv preprint arXiv:2310.12931.
- Maggi, Elena and Elena Vallino (2016). "Understanding urban mobility and the impact of public policies: The role of the agent-based models". In: Research in Transportation Economics 55, pp. 50–59.
- Mao, Jiageng, Yuxi Qian, Junjie Ye, Hang Zhao and Yue Wang (2023). "Gpt-driver: Learning to drive with gpt". In: arXiv preprint arXiv:2310.01415.
- Masri, Sari, Huthaifa I Ashqar and Mohammed Elhenawy (2024). "Leveraging Large Language Models (LLMs) for Traffic Management at Urban Intersections: The Case of Mixed Traffic Scenarios". In: arXiv preprint arXiv:2408.00948.
- Minaee, Shervin, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain and Jianfeng Gao (2024). "Large language models: A survey". In: arXiv preprint arXiv:2402.06196.

- Misra, Kushagra (2024). Era of Light Weight LLM Models: Innovations and Mobile Possibilities. URL: https://medium.com/@kushagramisra10/era-of-light-weight-llm-models-innovations-and-mobile-possibilities-8634064e0b09.
- Mo, Baichuan, Hanyong Xu, Dingyi Zhuang, Ruoyun Ma, Xiaotong Guo and Jinhua Zhao (2023a). Large Language Models for Travel Behavior Prediction. DOI: 10.48550/ARXIV.2312.00819. (Visited on 09/06/2024).
- (2023b). "Large language models for travel behavior prediction". In: arXiv preprint arXiv:2312.00819.
- Nakano, Reiichiro, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. (2021). "Webgpt: Browser-assisted question-answering with human feedback". In: arXiv preprint arXiv:2112.09332.
- Nijkamp, Erik, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese and Caiming Xiong (2022). "Codegen: An open large language model for code with multi-turn program synthesis". In: arXiv preprint arXiv:2203.13474.
- Nottingham, Kolby, Prithviraj Ammanabrolu, Alane Suhr, Yejin Choi, Hannaneh Hajishirzi, Sameer Singh and Roy Fox (2023). "Do embodied agents dream of pixelated sheep: Embodied decision making using language guided world modelling". In: *International Conference on Machine Learning*. PMLR, pp. 26311–26325.
- Oliveira, A, Mateus Espadoto, Roberto Hirata Jr, R Damaceno and Roberto M Cesar Jr (2024). "Towards a Method for Evaluating Bus Stop Infrastructure with Street Level Images and Large Language Models". In: *Proceedings*.
- Otal, Hakan T. and M. Abdullah Canbaz (June 2024). "LLM-Assisted Crisis Management: Building Advanced LLM Platforms for Effective Emergency Response and Public Collaboration". In: 2024 IEEE Conference on Artificial Intelligence (CAI), pp. 851–859. DOI: 10.1109/CAI59869.2024. 00159. arXiv: 2402.10908 [cs]. (Visited on 08/29/2024).
- Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. (2022). "Training language models to follow instructions with human feedback". In: *Advances in neural information processing systems* 35, pp. 27730–27744.
- Pan, Chenbin, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar and Liu Ren (2024a). "VLP: Vision Language Planning for Autonomous Driving". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14760–14769.
- Pan, Shirui, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang and Xindong Wu (2024b). "Unifying large language models and knowledge graphs: A roadmap". In: *IEEE Transactions on Knowledge and Data Engineering*.

- Pang, Aoyu, Maonan Wang, Man-On Pun, Chung Shue Chen and Xi Xiong (2024a). "iLLM-TSC: Integration reinforcement learning and large language model for traffic signal control policy improvement". In: arXiv preprint arXiv:2407.06025.
- (2024b). "iLLM-TSC: Integration reinforcement learning and large language model for traffic signal control policy improvement". In: arXiv preprint arXiv:2407.06025.
- Pappalardo, Luca, Salvatore Rinzivillo and Filippo Simini (2016). "Human mobility modelling: exploration and preferential return meet the gravity model". In: *Procedia Computer Science* 83, pp. 934–939.
- Penedo, Guilherme, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei and Julien Launay (2023). "The Refined-Web dataset for Falcon LLM: outperforming curated corpora with web data, and web data only". In: arXiv preprint arXiv:2306.01116.
- Peng, Baolin, Chunyuan Li, Pengcheng He, Michel Galley and Jianfeng Gao (2023). "Instruction tuning with gpt-4". In: arXiv preprint arXiv:2304.03277.
- Pham, Mai T, Andrijana Rajić, Judy D Greig, Jan M Sargeant, Andrew Papadopoulos and Scott A McEwen (2014). "A scoping review of scoping reviews: advancing the approach and enhancing the consistency". In: *Research synthesis methods* 5.4, pp. 371–385.
- Prabhod, Kummaragunta Joel (2023). "Advanced Techniques in Reinforcement Learning and Deep Learning for Autonomous Vehicle Navigation: Integrating Large Language Models for Real-Time Decision Making". In: *Journal of AI-Assisted Scientific Discovery* 3.1, pp. 1–20.
- Qin, Chengwei, Aston Zhang, Chen Chen, Anirudh Dagar and Wenming Ye (2023). "In-context learning with iterative demonstration selection". In: arXiv preprint arXiv:2310.09881.
- Quan, Yinzhu and Zefang Liu (2024). "InvAgent: A Large Language Model based Multi-Agent System for Inventory Management in Supply Chains". In: arXiv preprint arXiv:2407.11384.
- Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger and Ilya Sutskever (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv: 2103.00020 [cs.CV]. URL: https://arxiv.org/abs/2103.00020.
- Radford, Alec, Karthik Narasimhan, Tim Salimans and Ilya Sutskever (2018). "Improving Language Understanding by Generative Pre-Training". In: *OpenAI blog*.
- Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li and Peter J Liu (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *Journal of machine learning research* 21.140, pp. 1–67.
- Ramesh, Mohan and Fabian B. Flohr (June 2024). "Walk-the-Talk: LLM Driven Pedestrian Motion Generation". In: 2024 IEEE Intelligent Vehicles Symposium (IV). Jeju Island, Korea, Republic of: IEEE, pp. 3057–3062. ISBN: 9798350348811. DOI: 10.1109/IV55156.2024.10588860. (Visited on 10/03/2024).

- Rasal, Sumedh and Sanjay Kumar Boddhu (2024). "Beyond segmentation: Road network generation with multi-modal llms". In: *Science and Information Conference*. Springer, pp. 308–315.
- Ren, Yilong, Yue Chen, Shuai Liu, Boyue Wang, Haiyang Yu and Zhiyong Cui (2024). "TPLLM: A traffic prediction framework based on pretrained large language models". In: arXiv preprint arXiv:2403.02221.
- Ruan, Kangrui, Xinyang Wang and Xuan Di (2024). "From twitter to reasoner: Understand mobility travel modes and sentiment using large language models". In: arXiv preprint arXiv:2411.02666.
- Sahoo, Pranab, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal and Aman Chadha (2024). "A systematic survey of prompt engineering in large language models: Techniques and applications". In: arXiv preprint arXiv:2402.07927.
- Sampath, Koyyalagunta Krishna and M Supriya (2023). "Traffic Prediction in Indian Cities from Twitter Data Using Deep Learning and Word Embedding Models". In: *International Conference on Multi-disciplinary Trends in Artificial Intelligence*. Springer, pp. 671–682.
- Sanh, Victor, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. (2021). "Multitask prompted training enables zero-shot task generalization". In: arXiv preprint arXiv:2110.08207.
- Sawada, Tomohiro, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J Nay, Kshitij Gupta and Aran Komatsuzaki (2023). "Arb: Advanced reasoning benchmark for large language models". In: arXiv preprint arXiv:2307.13692.
- Schöpper, Henning and Wolfgang Kersten (2021). "Using natural language processing for supply chain mapping: a systematic review of current approaches". In: 5th international conference on computational linguistics and intelligent systems (COLINS 2021). 5. RWTH Aachen, pp. 71–86.
- Shahsavari, Maryam, Omar Khadeer Hussain, Morteza Saberi and Pankaj Sharma (2024). "Empowering Supply chains Resilience: LLMs-Powered BN for Proactive Supply Chain Risk Identification". In.
- Shao, Chenyang, Fengli Xu, Bingbing Fan, Jingtao Ding, Yuan Yuan, Meng Wang and Yong Li (2024). "Chain-of-planned-behaviour workflow elicits few-shot mobility generation in LLMs". In: arXiv preprint arXiv:2402.09836.
- Shazeer, Noam (2019). "Fast transformer decoding: One write-head is all you need". In:  $arXiv\ preprint\ arXiv:1911.02150$ .
- Singhal, Karan, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. (2023a). "Large language models encode clinical knowledge". In: *Nature* 620.7972, pp. 172–180.
- Singhal, Karan, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. (2023b). "Towards expert-level medical question answering with large language models". In: arXiv preprint arXiv:2305.09617.

- Sreeram, Shiva, Tsun-Hsuan Wang, Alaa Maalouf, Guy Rosman, Sertac Karaman and Daniela Rus (2024). "Probing Multimodal LLMs as World Models for Driving". In: arXiv preprint arXiv:2405.05956.
- Sui, Guanghu, Zhishuai Li, Ziyue Li, Sun Yang, Jingqing Ruan, Hangyu Mao and Rui Zhao (2023). "Reboost Large Language Model-based Text-to-SQL, Text-to-Python, and Text-to-Function—with Real Applications in Traffic Domain". In: arXiv preprint arXiv:2310.18752.
- Syed, Usman, Ethan Light, Xingang Guo, Huan Zhang, Lianhui Qin, Yanfeng Ouyang and Bin Hu (2024). "Benchmarking the Capabilities of Large Language Models in Transportation System Engineering: Accuracy, Consistency, and Reasoning Behaviors". In: arXiv preprint arXiv:2408.08302.
- Tan, Shuhan, Boris Ivanovic, Xinshuo Weng, Marco Pavone and Philipp Kraehenbuehl (2023). "Language conditioned traffic generation". In: arXiv preprint arXiv:2307.07947.
- Tanahashi, Kotaro, Yuichi Inoue, Yu Yamaguchi, Hidetatsu Yaginuma, Daiki Shiotsuka, Hiroyuki Shimatani, Kohei Iwamasa, Yoshiaki Inoue, Takafumi Yamaguchi, Koki Igari, et al. (2023). "Evaluation of large language models for decision making in autonomous driving". In: arXiv preprint arXiv:2312.06351.
- Taori, R, I Gulrajani, T Zhang, Y Dubois, X Li, C Guestrin, P Liang and TB Hashimoto Stanford Alpaca (2023a). An Instruction-Following LLaMA Model.
- Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang and Tatsunori B Hashimoto (2023b). "Alpaca: A strong, replicable instruction-following model". In: Stanford Center for Research on Foundation Models. https://crfm. stanford.edu/2023/03/13/alpaca. html 3.6, p. 7.
- Thompson, T Ben and Michael Sklar (2024). "FLRT: Fluent Student-Teacher Redteaming". In: arXiv preprint arXiv:2407.17447.
- Tian, Yonglin, Xuan Li, Hui Zhang, Chen Zhao, Bai Li, Xiao Wang and Fei-Yue Wang (2023). "VistaGPT: Generative parallel transformers for vehicles with intelligent systems for transport automation". In: *IEEE Transactions on Intelligent Vehicles*.
- Toch, Eran, Boaz Lerner, Eyal Ben-Zion and Irad Ben-Gal (2019). "Analyzing large-scale human mobility data: a survey of machine learning methods and applications". In: *Knowledge and Information Systems* 58, pp. 501–523.
- Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. (2023). "Llama 2: Open foundation and fine-tuned chat models". In: arXiv preprint arXiv:2307.09288.
- Tsai, Meng-Ju, Zhiyong Cui, Hao Yang, Cole Kopca, Sophie Tien and Yinhai Wang (2022). "Traffictwitter transformer: A nature language processing-joined framework for network-wide traffic forecasting". In: arXiv preprint arXiv:2206.11078.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser and Illia Polosukhin (2017). "Attention is All You Need". In: pp. 5998–6008.

- Villarreal, Michael, Bibek Poudel and Weizi Li (2023). "Can chatgpt enable its? the case of mixed traffic control via reinforcement learning". In: pp. 3749–3755.
- Wan, Xiangpeng, Michael C Lucic, Hakim Ghazzai and Yehia Massoud (2020). "Empowering real-time traffic reporting systems with nlp-processed social media data". In: *IEEE Open Journal of Intelligent Transportation Systems* 1, pp. 159–175.
- Wang, Bingzhang (2024). "Real-Time Data Informed Traffic Analytics Framework Powered by Large Language Model (LLM)". MA thesis. University of Washington.
- Wang, Bingzhang, Muhammad Monjurul Karim, Chenxi Liu, Yinhai Wang, et al. (2024a). "Traffic Performance GPT (TP-GPT): Real-Time Data Informed Intelligent ChatBot for Transportation Surveillance and Management". In: arXiv preprint arXiv:2405.03076.
- Wang, Boxin, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah and Bo Li (2021). "Adversarial glue: A multi-task benchmark for robustness evaluation of language models". In: arXiv preprint arXiv:2111.02840.
- Wang, Guan, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song and Yang Liu (2023a). "Open-chat: Advancing open-source language models with mixed-quality data". In: arXiv preprint arXiv:2309.11235.
- Wang, Jiawei, Renhe Jiang, Chuang Yang, Zengqing Wu, Makoto Onizuka, Ryosuke Shibasaki, Noboru Koshizuka and Chuan Xiao (2024b). Large Language Models as Urban Residents: An LLM Agent Framework for Personal Mobility Generation. DOI: 10.48550/ARXIV.2402.14744. (Visited on 09/06/2024).
- Wang, Lening, Yilong Ren, Han Jiang, Pinlong Cai, Daocheng Fu, Tianqi Wang, Zhiyong Cui, Haiyang Yu, Xuesong Wang, Hanchu Zhou, Helai Huang and Yinhai Wang (2023b). "AccidentGPT: Accident Analysis and Prevention from V2X Environmental Perception with Multi-modal Large Model". In: arXiv preprint arXiv:2312.13156. DOI: 10.48550/arXiv.2312.13156.
- Wang, Maonan, Aoyu Pang, Yuheng Kan, Man-On Pun, Chung Shue Chen and Bo Huang (2024c). "LLM-assisted light: Leveraging large language model capabilities for human-mimetic traffic signal control in complex urban environments". In: arXiv preprint arXiv:2403.08337.
- (2024d). "LLM-assisted light: Leveraging large language model capabilities for human-mimetic traffic signal control in complex urban environments". In: arXiv preprint arXiv:2403.08337.
- Wang, Naiyao, Tongbang Jiang, Ye Wang, Shaoyang Qiu, Bo Zhang, Xinqiang Xie, Munan Li, Chunliu Wang, Yiyang Wang, Hongxiang Ren, et al. (2024e). "KUNPENG: An Embodied Large Model for Intelligent Maritime". In: arXiv preprint arXiv:2407.09048.
- Wang, Peng, Xiang Wei, Fangxu Hu and Wenjuan Han (2024f). "Transgpt: Multi-modal generative pre-trained transformer for transportation". In: arXiv preprint arXiv:2402.07233.
- Wang, Pengqin, Meixin Zhu, Xinhu Zheng, Hongliang Lu, Hui Zhong, Xianda Chen, Shaojie Shen, Xuesong Wang, Yinhai Wang and Fei-Yue Wang (2024g). "BEVGPT: Generative Pre-trained Foundation Model for Autonomous Driving Prediction, Decision-Making, and Planning". In: *IEEE Transactions on Intelligent Vehicles*.

- Wang, Shanshan, Honghui Dong, Yue Zhou, Limin Jia and Yong Qin (2017). "Exploring traffic accident locations from natural language based on spatial information retrieval". In: 2017 29th Chinese Control And Decision Conference (CCDC). IEEE, pp. 3490–3495.
- Wang, Shenao, Yanjie Zhao, Xinyi Hou and Haoyu Wang (2024h). "Large language model supply chain: A research agenda". In: arXiv preprint arXiv:2404.12736.
- Wang, Shihao, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li and Jose M Alvarez (2024i). "OmniDrive: A Holistic LLM-Agent Framework for Autonomous Driving with 3D Perception, Reasoning and Planning". In: arXiv preprint arXiv:2405.01533.
- Wang, Shiyi, Yuxuan Zhu, Zhiheng Li, Yutong Wang, Li Li and Zhengbing He (2023c). "ChatGPT as your vehicle co-pilot: An initial attempt". In: *IEEE Transactions on Intelligent Vehicles*.
- Wang, Xintao, Qianwen Yang, Yongting Qiu, Jiaqing Liang, Qianyu He, Zhouhong Gu, Yanghua Xiao and Wei Wang (2023d). "Knowledgpt: Enhancing large language models with retrieval and storage access on knowledge bases". In: arXiv preprint arXiv:2308.11761.
- Wang, Yixuan, Ruochen Jiao, Chengtian Lang, Sinong Simon Zhan, Chao Huang, Zhaoran Wang, Zhuoran Yang and Qi Zhu (2023e). "Empowering autonomous driving with large language models: A safety perspective". In: arXiv preprint arXiv:2312.00812.
- Wang, Yizhong, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi and Hannaneh Hajishirzi (2022). "Self-instruct: Aligning language models with self-generated instructions". In: arXiv preprint arXiv:2212.10560.
- Wang, Yujin, Zhaoyan Huang, Quanfeng Liu, Yutong Zheng, Jinlong Hong, Junyi Chen, Lu Xiong, Bingzhao Gao and Hong Chen (2024j). "Drive as Veteran: Fine-tuning of an Onboard Large Language Model for Highway Autonomous Driving". In: 2024 IEEE Intelligent Vehicles Symposium (IV). IEEE, pp. 502–508.
- Wei, Jason, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai and Quoc V Le (2021). "Finetuned language models are zero-shot learners". In: arXiv preprint arXiv:2109.01652.
- Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. (2022). "Chain-of-thought prompting elicits reasoning in large language models". In:

  Advances in neural information processing systems 35, pp. 24824–24837.
- Wen, Licheng, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He and Yu Qiao (2023). "Dilu: A knowledge-driven approach to autonomous driving with large language models". In: arXiv preprint arXiv:2309.16292.
- Wen, Licheng, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, MA Tao, Yingxuan Li, XU Linran, Dengke Shang, et al. (2024). "On the Road with GPT-4V (ision): Explorations of Utilizing Visual-Language Model as Autonomous Driving Agent". In: ICLR 2024 Workshop on Large Language Model (LLM) Agents.

- Xia, Junkai, Chenxin Xu, Qingyao Xu, Chen Xie, Yanfeng Wang and Siheng Chen (2024). "Language-Driven Interactive Traffic Trajectory Generation". In: arXiv preprint arXiv:2405.15388.
- Xu, Guanhao, Jianfei Chen, Zejiang Wang, Anye Zhou, Max Schrader, Joshua Bittle and Yunli Shao (2025). "Enhancing Traffic Safety Analysis with Digital Twin Technology: Integrating Vehicle Dynamics and Environmental Factors into Microscopic Traffic Simulation". In: arXiv preprint arXiv:2502.09561.
- Xu, Guanhao and Vikash V Gayah (2023). "Non-unimodal and non-concave relationships in the network Macroscopic Fundamental Diagram caused by hierarchical streets". In: *Transportation Research Part B: Methodological* 173, pp. 203–227.
- Xu, Guohai, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, Xing Gao, Jitao Sang, Rong Zhang, et al. (2023). "Cvalues: Measuring the values of chinese large language models from safety to responsibility". In: arXiv preprint arXiv:2307.09705.
- Xu, Zhenhua, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li and Hengshuang Zhao (2024). "Drivegpt4: Interpretable end-to-end autonomous driving via large language model". In: *IEEE Robotics and Automation Letters*.
- Xuan, Khai Trinh, Khoi Nguyen Nguyen, Bach Hoang Ngo, Vu Dinh Xuan, Minh-Hung An and Quang-Vinh Dinh (June 2024). "Divide and Conquer Boosting for Enhanced Traffic Safety Description and Analysis with Large Vision Language Model". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 7046–7055.
- Xue, Hao, Flora D. Salim, Yongli Ren and Charles L. A. Clarke (Dec. 2021). Translating Human Mobility Forecasting through Natural Language Generation. DOI: 10.48550/arXiv.2112.11481. arXiv: 2112.11481 [cs]. (Visited on 09/09/2024).
- Xue, L (2020). "mt5: A massively multilingual pre-trained text-to-text transformer". In: arXiv preprint arXiv:2010.11934.
- Yan, Yimo, Andy HF Chow, Chin Pang Ho, Yong-Hong Kuo, Qihao Wu and Chengshuo Ying (2022). "Reinforcement learning for logistics and supply chain management: Methodologies, state of the art, and future opportunities". In: *Transportation Research Part E: Logistics and Transportation Review* 162, p. 102712.
- Yan, Yimo, Songyi Cui, Jiahui Liu, Yaping Zhao, Bodong Zhou and Yong-Hong Kuo (2025). "Multi-modal fusion for large-scale traffic prediction with heterogeneous retentive networks". In: *Information Fusion* 114, p. 102695.
- Yang, Menglong, Yanqiao Han, Yang Ren and Weizheng Li (2024a). "Large Language Model Guided Reinforcement Learning Based 6 Degree-of-Freedom Flight Control". In: *IEEE Access*.
- Yang, Ying, Wei Zhang, Hongyi Lin, Yang Liu and Xiaobo Qu (June 2024b). "Applying Masked Language Model for Transport Mode Choice Behavior Prediction". In: *Transportation Research Part A: Policy and Practice* 184, p. 104074. ISSN: 09658564. DOI: 10.1016/j.tra.2024.104074. (Visited on 09/05/2024).

- Yang, Zhenjie, Xiaosong Jia, Hongyang Li and Junchi Yan (2023). "Llm4drive: A survey of large language models for autonomous driving". In: arXiv e-prints, arXiv-2311.
- Yao, Shunyu, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao and Karthik Narasimhan (2024a). "Tree of thoughts: Deliberate problem solving with large language models". In: Advances in Neural Information Processing Systems 36.
- Yao, Yifan, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun and Yue Zhang (2024b). "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly". In: *High-Confidence Computing*, p. 100211.
- Yin, Ganmin, Zhou Huang, Chen Fu, Shuliang Ren, Yi Bao and Xiaolei Ma (2024a). "Examining active travel behavior through explainable machine learning: Insights from Beijing, China". In: Transportation Research Part D: Transport and Environment 127, p. 104038.
- Yin, Kai, Chengkai Liu, Ali Mostafavi and Xia Hu (2024b). "CrisisSense-LLM: Instruction Fine-Tuned
   Large Language Model for Multi-label Social Media Text Classification in Disaster Informatics".
   In: arXiv preprint arXiv:2406.15477. DOI: 10.48550/arXiv.2406.15477.
- Ying, Shaowei, Zhenlong Li and Manzhu Yu (Sept. 2024). Beyond Words: Evaluating Large Language Models in Transportation Planning. DOI: 10.48550/arXiv.2409.14516. arXiv: 2409.14516 [cs]. (Visited on 01/10/2025).
- Yu, Jiajie, Yuhong Wang and Wei Ma (2024). "Large Language Model-Enhanced Reinforcement Learning for Generic Bus Holding Control Strategies". In: arXiv preprint arXiv:2410.10212.
- Yu, Zhengyao, Guanhao Xu, Vikash V Gayah and Eleni Christofa (2020). "Incorporating phase rotation into a person-based signal timing optimization algorithm". In: *IEEE Transactions on Intelligent Transportation Systems* 23.1, pp. 513–521.
- Yuan, Kang, Yanjun Huang, Shuo Yang, Mingzhi Wu, Dongpu Cao, Qijun Chen and Hong Chen (2024). "Evolutionary Decision-Making and Planning for Autonomous Driving: A Hybrid Augmented Intelligence Framework". In: IEEE Transactions on Intelligent Transportation Systems.
- Zhang, Ge, Yemin Shi, Ruibo Liu, Ruibin Yuan, Yizhi Li, Siwei Dong, Yu Shu, Zhaoqun Li, Zekun Wang, Chenghua Lin, et al. (2023a). "Chinese open instruction generalist: A preliminary release". In: arXiv preprint arXiv:2304.07987.
- Zhang, Kunpeng, Shipu Wang, Ning Jia, Liang Zhao, Chunyang Han and Li Li (Apr. 2024a). "Integrating Visual Large Language Model and Reasoning Chain for Driver Behavior Analysis and Risk Assessment". In: *Accident Analysis & Prevention* 198, p. 107497. ISSN: 00014575. DOI: 10.1016/j.aap.2024.107497. (Visited on 09/10/2024).
- Zhang, Kunpeng, Feng Zhou, Lan Wu, Na Xie and Zhengbing He (2024b). "Semantic understanding and prompt engineering for large-scale traffic data imputation". In: *Information Fusion* 102, p. 102038.

- Zhang, Siyao, Daocheng Fu, Wenzhe Liang, Zhao Zhang, Bin Yu, Pinlong Cai and Baozhen Yao (2024c). "Trafficgpt: Viewing, processing and interacting with traffic foundation models". In: *Transport Policy* 150, pp. 95–105.
- Zhang, Wei, Miaoxin Cai, Tong Zhang, Guoqiang Lei, Yin Zhuang and Xuerui Mao (2024d). "Popeye: A Unified Visual-Language Model for Multi-Source Ship Detection from Remote Sensing Imagery". In: arXiv preprint arXiv:2403.03790.
- Zhang, Xue, Xiangyu Shi, Xinyue Lou, Rui Qi, Yufeng Chen, Jinan Xu and Wenjuan Han (2024e). "TransportationGames: Benchmarking Transportation Knowledge of (Multimodal) Large Language Models". In: arXiv preprint arXiv:2401.04471.
- Zhang, Zhexin, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang and Minlie Huang (2023b). "Safetybench: Evaluating the safety of large language models with multiple choice questions". In: arXiv preprint arXiv:2309.07045.
- Zhang, Zijian, Yujie Sun, Zepu Wang, Yuqi Nie, Xiaobo Ma, Peng Sun and Ruolin Li (2024f). "Large language models for mobility in transportation systems: A survey on forecasting tasks". In: arXiv preprint arXiv:2405.02357.
- Zhao, Ming, Omar Hussain, Yu Zhang and Morteza Saberi (2024a). "Optimizing Supply Chain Risk Management: An Integrated Framework Leveraging Large Language Models". In: 2024 IEEE Conference on Artificial Intelligence (CAI). IEEE, pp. 1057–1062.
- Zhao, Rui, Qirui Yuan, Jinyu Li, Yuze Fan, Yun Li and Fei Gao (Jan. 2024b). "DriveLLaVA: Human-Level Behavior Decisions via Vision Language Model". In: Sensors 24.13, p. 4113. ISSN: 1424-8220. DOI: 10.3390/s24134113. (Visited on 09/12/2024).
- Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. (2023). "A survey of large language models". In: arXiv preprint arXiv:2303.18223.
- Zhen, Hao, Yucheng Shi, Yongcan Huang, Jidong J. Yang and Ninghao Liu (2024). "Leveraging Large Language Models with Chain-of-Thought and Prompt Engineering for Traffic Crash Severity Analysis and Inference". In: arXiv preprint arXiv:2408.04652. DOI: 10.48550/arXiv.2408.04652.
- Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. (2023a). "Judging llm-as-a-judge with mt-bench and chatbot arena". In: *Advances in Neural Information Processing Systems* 36, pp. 46595–46623.
- Zheng, Ou, Mohamed Abdel-Aty, Dongdong Wang, Chenzhu Wang and Shengxuan Ding (2023b). "TrafficSafetyGPT: Tuning a pre-trained large language model to a domain-specific expert in transportation safety". In: arXiv preprint arXiv:2307.15311.
- Zhong, Ziyuan, Davis Rempe, Yuxiao Chen, Boris Ivanovic, Yulong Cao, Danfei Xu, Marco Pavone and Baishakhi Ray (2023). "Language-guided traffic simulation via scene-level diffusion". In: *Conference on Robot Learning*. PMLR, pp. 144–177.

- Zhou, Rongyan, Anjali Awasthi and Julie Stal-Le Cardinal (2021). "The main trends for multi-tier supply chain in Industry 4.0 based on Natural Language Processing". In: *Computers in Industry* 125, p. 103369.
- Zhou, Xingcheng, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao and Alois C Knoll (2024). "Vision language models in autonomous driving: A survey and outlook". In: *IEEE Transactions on Intelligent Vehicles*.
- Zhou, Yongchao, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan and Jimmy Ba (2022). "Large language models are human-level prompt engineers". In: arXiv preprint arXiv:2211.01910.
- Zhu, Kaijie, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, et al. (2023). "Promptbench: Towards evaluating the robustness of large language models on adversarial prompts". In: arXiv preprint arXiv:2306.04528.
- Zhu, Yuxuan, Shiyi Wang, Wenqing Zhong, Nianchen Shen, Yunqi Li, Siqi Wang, Zhiheng Li, Cathy Wu, Zhengbing He and Li Li (2024). A Survey on Large Language Model-empowered Autonomous Driving. arXiv: 2409.14165 [cs.AI]. URL: https://arxiv.org/abs/2409.14165.
- Ziemski, Michał, Marcin Junczys-Dowmunt and Bruno Pouliquen (2016). "The united nations parallel corpus v1. 0". In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 3530–3534.

#### 6. Appendix

#### Supplementary Details on Transformer Architectures 6.1.

#### **Attention Mechanism Details**

The self-attention mechanism assigns an attention score to each token based on its relevance to others. These scores are computed using query (Q), key (K), and value (V) matrices, as shown in Eq. 9:

Attention
$$(Q, K, V) = \operatorname{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$
 (9)

Multi-head attention extends self-attention by using multiple attention heads to focus on diverse aspects of input. The combined outputs are linearly transformed as shown in Eq. 10:

$$MultiHead(Q, K, V) = Concat(head_1, head_2, ..., head_h)W^O, W^O \in \mathbb{R}^{(h \cdot d_{head}) \times d_{model}}$$
(10)

To improve scalability, variants like multi-query attention (Shazeer, 2019) and grouped-query attention (Ainslie et al., 2023) reduce computational costs while maintaining performance.

# Positional Encoding Details

Absolute positional encoding, as introduced by Vaswani et al. (Vaswani et al., 2017), is defined as:

$$PE_{pos,2i} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{11}$$

$$PE_{pos,2i} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

$$PE_{pos,2i+1} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

$$(11)$$

This ensures that nearby positions have similar encodings. Relative positional encoding, on the other hand, encodes distances between tokens, making it effective for tasks requiring contextual understanding.

# **Decoder-Only Models**

Operate unidirectionally, processing text left-to-right, and are trained on next-token prediction loss. Examples include GPT (Radford et al., 2018; Brown, 2020).

### **Encoder-Only Models**

Use bidirectional attention for masked language modeling (MLM). Examples include BERT (Devlin, 2018) and RoBERTa (Liu et al., 2019).

# **Encoder-Decoder Models**

Combine bidirectional encoding and autoregressive decoding for sequence-to-sequence tasks like translation (e.g., T5 (Raffel et al., 2020), BART (Lewis et al., 2019)). For more mathematical derivations and examples, readers are directed to (Hadi et al., 2023; Kalyan et al., 2021; Huang et al., 2023b).

#### 6.2. Training Details

# 6.2.1. Pre-training Phase Details

# Next-token prediction

The objective of next-token prediction is to minimize the autoregressive loss:

$$\mathcal{L}_{AR}(\theta) = -\mathbb{E}_{(x_1, \dots, x_T) \sim \mathcal{D}} \left[ \sum_{t=1}^T \log P_{\theta}(x_t \mid x_{1:t-1}) \right]$$
(13)

# Masked Language Modeling (MLM)

In MLM, the model predicts masked tokens by minimizing the following loss:

$$\mathcal{L}_{\text{MLM}}(\theta) = -\mathbb{E}_{(x_1, \dots, x_T) \sim \mathcal{D}} \left[ \sum_{t \in M} \log P_{\theta}(x_t \mid x_{\setminus t}) \right]$$
(14)

# 6.2.2. Post-training Phase Details

# Alignment with human preferences

Alignment with human preferences involves training a reward model  $R_{\phi}(x)$  using human feedback. The reward model's objective is formulated as:

$$\mathcal{L}_{\text{reward}}(\phi) = -\mathbb{E}_{(x_1, x_2) \sim \mathcal{D}_{\text{human}}} \left[ \log \frac{e^{R_{\phi}(x_1)}}{e^{R_{\phi}(x_1)} + e^{R_{\phi}(x_2)}} \right]$$
(15)

The fine-tuning objective for RLHF is:

$$\mathcal{L}_{\text{RLHF}}(\theta) = -\mathbb{E}_{x \sim \pi_{\theta}(x|z)} \left[ R_{\phi}(x) \right] \tag{16}$$

The Proximal Policy Optimization (PPO) objective is used for reinforcement learning:

$$\mathcal{L}_{PPO}(\theta) = \mathbb{E}_{x \sim \pi_{\theta}} \left[ \min \left( r(\theta), \text{clip}(r(\theta), 1 - \epsilon, 1 + \epsilon) \right) \cdot R_{\phi}(x) \right] \tag{17}$$

where  $r(\theta)$  is the probability ratio:

$$r(\theta) = \frac{\pi_{\theta}(x \mid z)}{\pi_{\theta_{\text{old}}}(x \mid z)} \tag{18}$$

# Fine-tuning for downstream tasks

Fine-tuning loss for code generation tasks is:

$$\mathcal{L}_{\text{code}}(\theta) = -\mathbb{E}_{(c_1, \dots, c_T) \sim \mathcal{D}_{\text{code}}} \left[ \sum_{t=1}^T \log P_{\theta}(c_t \mid c_{1:t-1}) \right]$$
(19)

For domain-specific datasets, the fine-tuning objective is:

$$\mathcal{L}_{\text{domain}}(\theta) = -\mathbb{E}_{(x_1, \dots, x_T) \sim \mathcal{D}_{\text{domain}}} \left[ \sum_{t=1}^{T} \log P_{\theta}(x_t \mid x_{1:t-1}) \right]$$
(20)

# 6.3. Other Prompt Engineering Techniques

CoT Prompting with Self-Consistency Also known as the multiple chain-of-thought prompting technique, (Wang et al., 2022) introduces a "sample-and-marginalize" procedure to identify the most consistent answer for complex tasks requiring deliberate reasoning. LLMs generate diverse reasoning paths using chain-of-thought prompting and then reach a consensus through majority voting. This method is effective for tasks with fixed sets of correct answers, where self-consistency can be leveraged. However, for open-ended tasks, careful design of consistency metrics is necessary to ensure reliability.

Tree-of-Thought Prompting The tree-of-thought (Yao et al., 2024a) technique generalizes chain-of-thought prompting by exploring multiple reasoning paths that branch out like a tree, enabling self-evaluation, backtracking, and looking ahead. Unpromising branches are pruned, and each reasoning step acts as an intermediate stop to help decompose complex tasks. For instance, in solving a multi-step problem like the game of 24, tree-of-thought breaks the task into intermediate steps and expands reasoning paths from each node. Heuristics evaluate branches, and search algorithms like breadth-first or depth-first search guide the process, improving performance in multi-step reasoning tasks.

Graph-of-Thought Prompting The graph-of-thought (Besta et al., 2024) technique extends tree-of-thought by connecting intermediate reasoning steps into a network, forming a large graph of thoughts. The key innovation lies in the aggregation process, where edges from separate reasoning paths converge at shared vertices. Sorting and ranking are used to organize branches selectively. This method excels in tasks like sorting, keyword counting, and sequential planning, where multiple goals must be achieved in a specific order.

#### 6.4. Additional LLMs

In this section, we provide a brief overview of additional LLMs that were not included in the main content but are noteworthy for their unique contributions and applications.

Vicuna <sup>††</sup> is an open-source chatbot fine-tuned from LLaMA on approximately 70K user-shared conversations from ShareGPT. It achieves 90% of ChatGPT's quality in preliminary evaluations using GPT-4 as the judge. Vicuna is widely recognized for its ability to produce detailed, well-structured responses, and it serves as a strong alternative to proprietary models in open-source research.

 $<sup>^{\</sup>dagger\dagger}$ https://lmsys.org/blog/2023-03-30-vicuna/

**BLOOM BLOOM** (Le Scao et al., 2023) is a multilingual, open-access model with 176 billion parameters, trained on 59 languages. It emphasizes inclusivity and accessibility for diverse linguistic communities. BLOOM is particularly notable for its large-scale collaborative development by the BigScience project, making it a pioneering effort in democratizing LLM research.

**FLAN FLAN** (Wei et al., 2021) is Google's instruction-tuned model designed to improve zero-shot and few-shot performance. By fine-tuning on chain-of-thought and instruction data, FLAN achieves significant gains in reasoning and comprehension tasks, making it a strong contender in instruction-following applications.

Other Notable LLMs The following models, while not included in the main content, are significant for specific tasks or domains:

- Alpaca (Taori et al., 2023b): Fine-tuned from LLaMA using self-instruct techniques, Alpaca is designed for cost-effective fine-tuning and task-specific applications.
- Mistral-7B (Jiang et al., 2023): A smaller, high-performing model that outperforms larger opensource models like LLaMA-2-13B in several benchmarks.
- Med-PaLM and Med-PaLM 2 (Singhal et al., 2023a; Singhal et al., 2023b): Domain-specific versions of PaLM optimized for medical question answering.
- CodeGen (Nijkamp et al., 2022): An open-source model tailored for programming tasks, capable of generating code in multiple programming languages.
- T0 (Sanh et al., 2021): A model designed for natural language tasks by mapping them into a prompted form, showcasing strong generalization in zero-shot settings.

# 6.5. Prompt Optimization Relevant Concepts (mentioned in Future Research Directions)

Attack Success Rate Attack Success Rate (ASR) is a metric used to evaluate how vulnerable a machine learning model, including LLMs, is to adversarial attacks. These attacks are deliberate attempts to manipulate the model into producing incorrect, biased, or harmful outputs. ASR quantifies the effectiveness of such attacks.

**Perplexity** Perplexity is a key metric used to assess the performance of language models in NLP. It measures how well a model predicts a sequence of words, essentially capturing the model's uncertainty about the next word in a sequence. A lower perplexity score indicates that the model is better at making accurate predictions.

Perplexity (PPL) is mathematically defined as the exponentiation of the average negative loglikelihood of a word sequence. For a sequence of words  $w_1, w_2, \ldots, w_n$ , it is calculated as:

PPL = 
$$\exp\left(-\frac{1}{n}\sum_{i=1}^{n}\log P(w_i \mid w_{1:i-1})\right)$$

where  $P(w_i | w_{1:i-1})$  is the probability of word  $w_i$  given prior words, and n is the sequence length.