Harnessing Large Language Models for Adaptive and Explainable Traffic Forecasting

Haoyang Yan[1] and Xiaolei Ma*[1,2]

[1]School of Transportation Science and Engineering, Beihang University, Beijing, China.

[2]Key Laboratory of Intelligent Transportation Technology and System of the Ministry of Education, Beihang University, Beijing 102206, China

*Address correspondence to: xiaolei@buaa.edu.cn

Abstract

Accurate traffic state prediction is a cornerstone of Intelligent Transportation Systems (ITS). While deep learning models—specifically Graph Neural Networks (GNNs) and Transformers—have achieved state-of-the-art performance in routine forecasting, they exhibit significant fragility under anomalous conditions (e.g., accidents, extreme weather, public events) due to their reliance on historical stationarity. Furthermore, the "black-box" nature of these models precludes interpretability, limiting their operational utility. To address these deficiencies, this study proposes Chat-ITS, a novel hybrid framework that synergizes robust probabilistic time-series forecasting with the semantic reasoning capabilities of Large Language Models (LLMs). The methodology comprises three mathematically formalized stages: (1) A foundation model based on discrete tokenization and sequence-to-sequence learning generates a probabilistic distribution of future trajectories; (2) A cross-modal reasoning module, driven by an LLM, processes heterogeneous unstructured text data to perform Bayesian-like contextual adjustments on the candidate trajectories; (3) An operational interface generates explainable diagnostics and actionable control strategies. Extensive experiments on large-scale real-world datasets from Beijing demonstrate that Chat-ITS reduces prediction error by up to 15% during non-recurring congestion events compared to baselines, while offering zero-shot generalization to unseen event types.

1 Introduction

Accurate traffic prediction is fundamental to the efficacy of Intelligent Transportation Systems (ITS), enabling critical functions such as dynamic route guidance, adaptive traffic signal control, and proactive incident management essential for mitigating congestion, reducing emissions, and enhancing urban mobility resilience [1]. Congestion alone costs economies billions annually and degrades quality of life in urban centers [2]. Effective ITS, powered by reliable forecasts, promises substantial improvements in transportation efficiency and sustainability. Recent advances, particularly the application of deep learning techniques like graph neural networks (GNNs) for modeling complex spatial dependencies across road networks [3] and sophisticated sequence models (e.g., temporal convolution networks, attention mechanisms) for capturing temporal dynamics [4, 5], have considerable improved short-term forecasting accuracy under typical, recurring traffic conditions [6]. These methods effectively learn patterns from large historical datasets, providing a strong foundation for next-generation ITS applications operating under predictable circumstances.

Despite these successes, existing state-of-the-art traffic forecasting methods face critical limitations that hinder their real-world operational utility, particularly under non-routine circumstances [7–9]. Firstly, their predictive performance often degrades sharply during anomalous events such as road accidents, unexpected road closures, severe weather conditions, or large-scale public gatherings [10, 11]. Models trained primarily on routine historical patterns often exhibit poor generalization capabilities when faced with data distributions shifted by these irregular occurrences [12, 13]. This fragility undermines their reliability precisely when accurate prediction is most needed for effective incident response and management. Secondly, the fixed input encoding mechanisms of many deep learning models limit their ability to flexibly incorporate diverse, unstructured, or dynamic updates on road work schedules often contains crucial context for anticipating traffic impacts. Integrating textual incident reports, event schedules, social media alerts, or unforeseen disruptions often requires complex feature engineering or extensive model retraining, impeding adaptation to unforeseen event types without clear overhead [14]. Thirdly, and perhaps most crucially for translation into practice, the standard output of these models, typically a high-dimensional matrix or tensor representing predicted speeds or flows, lacks direct interpretability. It fails to convey the underlying reasons for the predicted state or provide actionable guidance for traffic operators and decision-makers [15]. Consequently, even statistically accurate forecasts may not readily translate into effective, timely, and context-aware traffic management interventions, limiting the practical impact of these advanced techniques.

Large language models (LLMs) have emerged as powerful tools demonstrating remarkable capabilities in natural language understanding, contextual reasoning, and generalization across diverse tasks [16]. Their potential to process unstructured text, synthesize information from multiple sources, and generate human-like explanations offers promising avenues to address the challenges of context integration and interpretability in ITS [7, 17, 18]. However, applying LLMs directly to the task of numerical time-series forecasting presents inherent difficulties. Their architectures, primarily optimized for sequential token generation, often struggle with the precise numerical regression required for traffic state prediction and can be inefficient in capturing the complex spatio-temporal statistical dependencies inherent in traffic flow [14, 19]. Furthermore, training or even fine-tuning large LLMs for specialized forecasting tasks demands substantial computational resources and large-scale, domain-specific datasets, often proving impractical for widespread deployment in operational ITS settings where data characteristics can vary across locations and time [20].

Here, we introduce Chat-ITS, a novel hybrid forecasting framework designed to bridge the gap between robust probabilistic time-series modeling and the contextual reasoning capabilities of LLMs, thereby overcoming the aforementioned limitations. Chat-ITS employs a synergistic, multi-stage approach that deliberately leverages the distinct strengths of each component. It first utilizes a dedicated spatio-temporal foundation model, pre-trained on extensive historical traffic data, to generate multiple candidate traffic state trajectories along with associated uncertainty estimates. This ensures statistical rigor and captures complex baseline traffic dynamics. Subsequently, an LLM, operating on these candidate trajectories, is conditioned on flexible natural language prompts. These prompts can seamlessly encode both structured data (e.g., quantitative weather forecasts, road closure notices with coordinates and times) and unstructured descriptions rich with linguistic cues (e.g., ”Event update: sold-out show at the downtown arena, scheduled to end at 10 PM” or ”Dispatch log: report of a multi-vehicle collision with emergency services responding on the northbound lane near exit 15”). The LLM evaluates the candidate trajectories within this broader context, reasoning about the likely impacts to select or adjust towards the most plausible outcome given the real-time information. Crucially, the LLM also generates human-readable explanations for its choice and actionable recommendations tailored for traffic management personnel, integrating insights potentially learned from historical operational data. This architecture deliberately avoids tasking the LLM with direct numerical prediction, instead harnessing its strengths in semantic comprehension, causal inference, and context-aware reasoning.

We demonstrate through comprehensive experiments encompassing both routine traffic patterns and a diverse set of simulated and real-world anomalous scenarios (including construction, accidents, and public events) that Chat-ITS noticeable outperforms conventional deep learning baseline models during irregular events, reducing prediction errors by up to 15% under certain conditions, while matching state-of-the-art accuracy under normal conditions. Crucially, case studies highlight the framework's ability to generalize zero-shot to unseen event types described only via text prompts and deliver context-aware, actionable insights (e.g., suggesting specific signal timing adjustments, disseminating targeted traveler advisories, or recommending dynamic routing strategies). By integrating the statistical power of probabilistic forecasting with the semantic understanding and reasoning capabilities of language-based AI, Chat-ITS presents a new paradigm for traffic prediction, one that is not only accurate and adaptive but also explainable and directly aligned with the practical needs of transportation practitioners for effective real-world ITS deployment.

2 Literature Review

2.1 Deep Learning for Traffic Forecasting

Traffic forecasting has evolved from traditional statistical and regression-based approaches to deep learning models capable of capturing complex temporal dynamics. Early studies primarily relied on Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) architectures, to model sequential dependencies in traffic speed and flow data [21, 22]. While effective in temporal modeling, these approaches are limited in their ability to explicitly represent the non-Euclidean spatial topology of road networks.

To address this limitation, Spatio-Temporal Graph Neural Networks (STGNNs) have been widely adopted. Representative models such as DCRNN [23] and Graph WaveNet [24] combine graph convolutions with temporal sequence modeling to capture spatial correlations and temporal dynamics simultaneously. Subsequent works have extended these frameworks by relaxing the assumption of static spatial dependencies. For instance, Traffic Transformer introduces global–local decoders to hierarchically aggregate spatial features [25], while PDFormer incorporates propagation delay-aware attention to explicitly model temporal lags in traffic interactions [26].

Despite their expressive power, the necessity of graph convolution has recently been questioned. Empirical studies indicate that simplified spatial modeling strategies can achieve comparable performance with substantially reduced computational overhead. SimST demonstrates that lightweight spatial aggregation approximates the effectiveness of GCN-based methods [27]. Similarly, MLP-based architectures such as ST-MLP [28] and STID [29] show that concise spatio-temporal identity mappings can outperform complex GNNs by reducing overfitting to noisy spatial correlations.

In parallel, Transformer-based models have gained attention for their ability to capture long-range temporal dependencies. PatchTST segments time series into patches to preserve local temporal semantics [30], whereas iTransformer inverts the attention mechanism to better model multivariate correlations [31]. To further account for uncertainty and stochasticity in traffic systems, probabilistic generative approaches such as SpecSTG [32] and Diffusion-TS [15] have been proposed, enabling uncertainty-aware forecasting and data imputation.

2.2 Foundation Models for Time Series

Inspired by the “pre-train and fine-tune” paradigm in Natural Language Processing, recent research has shifted toward the development of foundation models for time series analysis [33, 34]. These models aim to learn universal temporal representations that generalize across datasets and tasks, enabling zero-shot or few-shot inference.

Chronos adapts T5-style architectures by discretizing continuous values into token sequences, achieving strong zero-shot performance across diverse domains [35]. Lag-Llama adopts a probabilistic modeling framework to capture scaling behaviors over large-scale time-series corpora [36]. To model multi-periodicity, TimesNet reformulates one-dimensional time series into two-dimensional representations, facilitating variation modeling via convolutional kernels [37].

Recent efforts further emphasize unified modeling across heterogeneous tasks. UniTS proposes a prompt-based backbone capable of jointly addressing forecasting, classification, and imputation [38], while Moirai-MoE introduces a Mixture-of-Experts architecture to handle diverse temporal resolutions without manual frequency alignment [39]. Another research direction explores the reuse of frozen pre-trained language or vision models. One Fits All demonstrates that large language models can be adapted to time series with minimal parameter updates [40]. Nevertheless, most existing foundation models operate exclusively on numerical signals, limiting their ability to incorporate unstructured contextual information, such as event descriptions, that is often critical for interpreting anomalies in intelligent transportation systems [41].

2.3 Large Language Models in Transportation

The application of Large Language Models (LLMs) in transportation research introduces a paradigm shift from purely numerical modeling toward semantic reasoning and agent-based decision making. A central challenge lies in aligning continuous time series data with the discrete token-based representations of LLMs. Time-LLM and LLM4TS address this issue through reprogramming and fine-tuning strategies that encode numerical sequences as language tokens [20, 42].

In urban computing scenarios, UrbanGPT integrates spatio-temporal dependency encoders with instruction tuning to improve generalization under data scarcity [43], while ST-LLM reformulates spatio-temporal observations as token sequences to capture global network dependencies [44]. Beyond forecasting, LLMs have been explored for explanation, simulation, and decision support. TF-LLM and ChatTraffic generate natural language interpretations of traffic conditions and congestion causes, enhancing model interpretability [45, 46]. CityGPT further extends this idea by constructing a city-scale world model in which LLM-based agents perform diverse urban tasks [18].

More advanced frameworks adopt “LLM-in-the-loop” architectures. TimeCAP and TimeXL employ multi-agent systems in which LLMs generate contextual summaries or reasoning paths that guide downstream numerical predictors [47, 48]. Additionally, external information sources such as social events and news reports have been incorporated via generative agents to stabilize forecasting under non-stationary conditions [49]. Nevertheless, a fundamental challenge remains in effectively reconciling the numerical accuracy of specialized time-series models with the high-level semantic reasoning capabilities of LLMs, particularly for real-time anomaly detection and intervention.

3 Preliminary

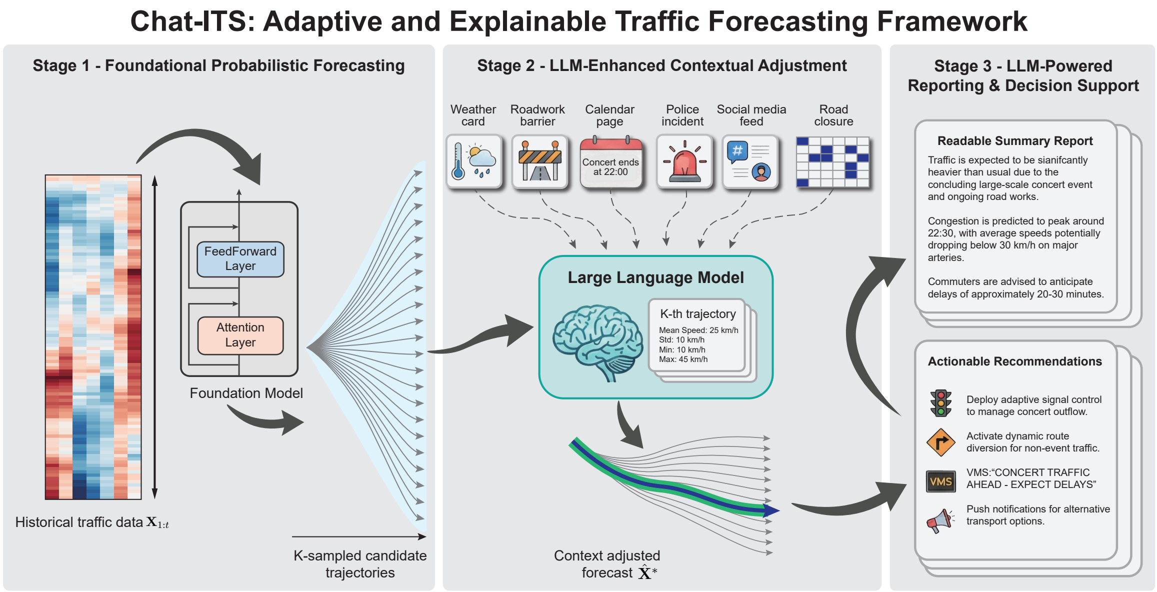
Traffic prediction is typically framed as a short-term time-series forecasting task, where future values $\mathbf{X}_{T+1:T+n}$ are predicted based on historical observations $\mathbf{X}_{1:T}$. This paper tackles a multi-modal version of this problem, recognizing that real-world traffic dynamics are influenced not only by past traffic states but also by a plethora of contextual factors often conveyed through textual or structured non-time-series data. We work with input instances $(\mathbf{X}_{1:T}, \mathbf{s})$, consisting of historical time series data $\mathbf{X}_{1:T} = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$, where each $\mathbf{x}_t \in \mathbb{R}^N$ captures D features of traffic states (e.g., speed, flow, occupancy) for N spatial locations (e.g., road segments, sensors) over T historical time steps, and auxiliary contextual information $\mathbf{s}$. This contextual information $\mathbf{s}$ can be diverse, including structured data (e.g., weather parameters, event schedules, road work logs) and unstructured natural language text (e.g., incident reports, social media alerts, news feeds) that potentially influences the time series and provides valuable context for improving forecast accuracy, especially during non-routine conditions. Our objective is to develop a model $\mathcal{F}$ that these multi-modal inputs to accurate and reliable predictions of future traffic states, potentially including uncertainty quantification. This is formalized as:

$$\mathbf{X}_{T+1:T+n} = \{\mathbf{x}_{T+1}, \mathbf{x}_{T+2}, \dots, \mathbf{x}_{T+n}\} = \mathcal{F}(\mathbf{X}_{1:T}, \mathbf{s}), \tag{1}$$

where $\mathbf{X}_{T+1:T+n}$ is the predicted sequence of n future state vectors or distributions. The ultimate goal is to identify an optimal model $\mathcal{F}$ that delivers accurate and reliable predictions while also being explainable and effectively leveraging the contextual information from $\mathbf{s}$ to adapt to both routine and non-routine conditions.

4 Methodology

4.1 Overall Framework


Chat-ITS: Adaptive and Explainable Traffic Forecasting Framework
Stage 1 - Foundational Probabilistic Forecasting
FeedForward Layer
Attention Layer
Foundation Model
Historical traffic data $\mathbf{X}_{1:t}$
K-sampled candidate trajectories
Stage 2 - LLM-Enhanced Contextual Adjustment
Weather card
Roadwork barrier
Calendar page
Concert ends at 22:00
Police incident
Social media feed
Road closure
Large Language Model
K-th trajectory
Mean Speed: 25 km/h
Std: 10 km/h
Min: 10 km/h
Max: 45 km/h
Context adjusted forecast $\hat{\mathbf{X}}^*$
Stage 3 - LLM-Powered Reporting & Decision Support
Readable Summary Report
Actionable Recommendations
VMS

Traffic is expected to be sianifcantly heavier than usual due to the concluding large-scale concert event and ongoing road works.

Congestion is predicted to peak around 22:30, with average speeds potentially dropping below 30 km/h on major arteries.

Commuters are advised to anticipate delays of approximately 20-30 minutes.

Deploy adaptive signal control to manage concert outflow.

Activate dynamic route diversion for non-event traffic.

VMS:“CONCERT TRAFFIC AHEAD - EXPECT DELAYS”

Push notifications for alternative transport options.

Figure 1: Overall architecture of the Chat-ITS framework. (A) Stage 1: A pre-trained spatio-temporal foundation model processes historical traffic data $\mathbf{X}_{1:T}$ to generate multiple candidate future trajectories $\{\hat{\mathbf{X}}^{(k)}\}$ representing a baseline probabilistic forecast. (B) Stage 2: Real-time contextual information $\mathbf{s}$, including structured and unstructured event data, is processed by an LLM. The LLM reasons about the event’s impact and evaluates the candidate trajectories, selecting or adjusting to the most plausible event-conditioned forecast $\hat{\mathbf{X}}^*$. (C) Stage 3: The adjusted forecast, along with historical dispatch patterns, feeds into the LLM to generate human-readable summary reports and actionable traffic management recommendations.

The Chat-ITS framework, depicted schematically in Fig.1,operates through three synergistic core stages designed to integrate the strengths of advanced time-series modeling and large language models: (1) Foundational Probabilistic Forecasting, (2) LLM-Enhanced Contextual Adjustment, and (3) LLM-Powered Reporting and Decision Support.

- **Stage 1: Foundational Probabilistic Forecasting (Fig.1 A)**: The foundation of Chat-ITS is a robust forecasting model capable of capturing complex dependencies in traffic data and providing probabilistic outputs. We employ a state-of-the-art architecture pre-trained on extensive historical traffic data. To ensure the model learns representative patterns, the pre-training data include curated subsets, such as: (i) time-series from high-volume road segments or grid cells representing typical urban traffic dynamics, and (ii) traffic data aggregated around key venues (stadiums, transport hubs, event centers) known to generate non-standard patterns. Inspired by architectures like Chronos [35] which adapt language transformer-based models for time-series, our foundation model processes the historical input and generates not a single prediction, but multiple trajectory samples $\{\hat{\mathbf{X}}^{(k)}_{T+1:T+n}\}_{k=1}^{K}$. These samples collectively approximate the predictive distribution $P(\mathbf{X}_{T+1:T+n}|\mathbf{X}_{1:T})$, providing a baseline probabilistic forecast and inherent uncertainty quantification, crucial for representing the range of possibilities under routine conditions.

- **Stage 2: LLM-Enhanced Contextual Adjustment (Fig.1 B)**: This stage integrates real-time contextual information $\mathbf{s}$ to refine the baseline forecast, addressing the limitations of models relying solely on historical patterns. The contextual information $\mathbf{s}$, which can include structured data and unstructured text, is processed by an LLM. The LLM executes a chain-of-thought process: first summarizing the event information, then reasoning about its likely causal impact on traffic flow (location, severity, duration), and finally assessing the quantitative effect. Then the LLM selects the most plausible trajectory $\hat{\mathbf{X}}^{*}_{T+1:T+n}$ or potentially generates an adjusted trajectory that better reflects the anticipated impact of the event. This step leverages the LLM's ability to understand and reason about novel or complex situations described in natural language, effectively modulating the initial probabilistic forecast based on real-time context.

- **Stage 3: LLM-Powered Reporting and Decision Support (Fig.1 C)**: The final stage focuses on translating the adjusted forecast $\hat{\mathbf{X}}^{*}_{T+1:T+n}$ into practical outputs for end-users. The LLM receives the context-adjusted forecast and potentially relevant historical traffic guidance data. This historical guidance data allows the LLM to learn implicit operational preferences and common responses implemented by human traffic controllers in similar past situations. Based on the adjusted forecast, the contextual information, and the learned operational patterns, the LLM generates: (i) a concise, human-readable summary report describing the anticipated traffic conditions, highlighting potential issues (e.g., specific bottlenecks, expected delay increases), and explaining the reasoning based on the contextual factors; and (ii) actionable recommendations for traffic management (e.g., ”Consider adjusting signal timing plan B on Corridor X between 8-10 AM,” ”Disseminate advisory regarding lane closure on Highway Y,” ”Prepare diversion route Z”). This stage bridges the gap between raw numerical prediction and practical operational utility, providing explainable insights and decision support.

4.2 Stage 1: Foundational Probabilistic Forecasting

In this stage, we reformulate the time-series forecasting problem as a specialized language modeling task. By mapping continuous traffic dynamics into discrete semantic tokens, we leverage the robust reasoning capabilities of the Transformer architecture to capture complex temporal dependencies and model the aleatoric uncertainty inherent in stochastic traffic flows.

4.2.1 Sequence Serialization and Tokenization Strategy

Unlike traditional point-wise forecasting models, we adopt a patch-based tokenization strategy inspired by the Chronos paradigm [35]. This approach reduces the sequence length complexity from $O(T^2)$ to $O((T/S)^2)$ while preserving local temporal semantics.

Temporal Patching Given a univariate time series $\mathbf{x}^i = \{x_1^i, \ldots, x_T^i\} \in \mathbb{R}^T$ for a spatial node i, we first decompose the sequence into a series of overlapping patches. Let P denote the patch length and S the stride. The sequence is unfolded into a matrix of patches $\mathcal{P}^i = \{\mathbf{p}_1^i, \ldots, \mathbf{p}_N^i\}$, where the j-th patch $\mathbf{p}_j^i \in \mathbb{R}^P$ is defined as:

$$\mathbf{p}_j^i = [x_{(j-1)S+1}^i, \ldots, x_{(j-1)S+P}^i] \tag{2}$$

where $N = \lfloor (T-P)/S \rfloor + 1$ is the number of tokens.

Local Scaling for Stationarity Traffic data exhibits significant non-stationarity (e.g., varying peak hours across days). To ensure the distribution of values within each patch falls into a learnable range for the quantizer, we apply instance-level mean scaling. For each patch $\mathbf{p}_j^i$, we compute a local scale factor $s_j = \frac{1}{P}\sum_{k=1}^{P} |\mathbf{p}_{j,k}^i| + \epsilon$. The scaled patch is obtained via:

$$\tilde{\mathbf{p}}_j^i = \frac{\mathbf{p}_j^i}{s_j} \tag{3}$$

This normalization allows the model to learn scale-invariant temporal patterns, generalizing across different traffic volume levels.

Quantization and Vocabulary Mapping To interface with the categorical nature of language models, we map the continuous scaled domain $\mathbb{R}$ to a discrete codebook $\mathcal{C} = \{c_1, \ldots, c_V\}$ of size V. We employ a quantization function $Q : \mathbb{R} \to \{1, \ldots, V\}$ using uniform binning within a fixed range $[\min, \max]$. The token ID $z_{j,k}$ for the k-th element of the j-th patch is derived as:

$$z_{j,k} = Q(\tilde{p}_{j,k}^i) = \text{clip}\left(\left\lfloor \frac{\tilde{p}_{j,k}^i - \min}{\max - \min} \times (V-1) \right\rfloor, 0, V-1\right) \tag{4}$$

Consequently, the continuous time series $\mathbf{x}^i$ is transformed into a sequence of discrete token IDs $\mathbf{Z}^i = \{z_{1,1}, \ldots, z_{N,P}\}$, which serves as the "sentence" input to the Transformer.

4.2.2 Encoder-Decoder Architecture

We employ a modified T5 (Text-to-Text Transfer Transformer) backbone [50] to process the tokenized traffic sequences. The architecture consists of an encoder that maps the historical token sequence to a latent representation, and a decoder that autoregressively generates future tokens.

Self-Attention with Relative Position Bias The core mechanism is the Multi-Head Self-Attention (MHSA). Unlike standard Transformers that use absolute sinusoidal positional encodings, T5 utilizes relative positional embeddings, which is crucial for time series as it naturally models the "time lag" distance between patches. The attention score between query q and key k is computed as:

$$\mathcal{A}_{q,k} = \text{softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d_{model}}} + \mathbf{B}_{q,k}\right)\mathbf{V} \tag{5}$$

where $\mathbf{B}_{q,k}$ is a learnable scalar bias added to the attention logit, representing the relative temporal distance between position q and k.

Objective Function The model is trained to minimize the standard Cross-Entropy loss over the vocabulary V. Given the historical context $\mathbf{Z}_{<t}$, the model predicts the probability distribution of the next token z_t:

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{L} \log P_\theta(z_t|\mathbf{Z}_{<t}) \tag{6}$$

We incorporate a multi-task learning objective by combining the primary forecasting task with a random masking reconstruction task (BERT-style) to enhance the model's robustness against missing data and noise.

4.2.3 Probabilistic Trajectory Generation

To capture the aleatoric uncertainty inherent in future traffic states, we move beyond point estimation to probabilistic trajectory generation.

Nucleus Sampling (Top-p) During inference, instead of greedy decoding (selecting the token with max probability), we approximate the posterior distribution $P(\mathbf{X}_{\text{future}}|\mathbf{X}_{\text{history}})$ by sampling K independent hypotheses (trajectories). We employ Nucleus Sampling, which truncates the tail of the distribution, considering only the smallest set of top tokens whose cumulative probability exceeds a threshold p:

$$\mathcal{V}^{(p)} = \{z \in \mathcal{C} \mid \sum_{z' \in \mathcal{V}^{(p)}} P_\theta(z'|z_{<t}) \geq p\} \tag{7}$$

At each step t, the next token $z_t^{(k)}$ for the k-th hypothesis is sampled from the re-normalized distribution over $\mathcal{V}^{(p)}$.

De-tokenization and Aggregation The generated token sequences are mapped back to the continuous domain via inverse quantization and inverse scaling using the stored local scale factors s_j. Since patches overlap, the value at a specific time step t is reconstructed by averaging the predictions from all patches covering that step, ensuring smoothness at patch boundaries:

$$\hat{x}_t^{(k)} = \frac{1}{|\Omega_t|} \sum_{j \in \Omega_t} s_j \cdot Q^{-1}(z_{j,\text{idx}(t)}^{(k)}) \tag{8}$$

where Ω_t is the set of patches containing time step t. This process yields a set of candidate trajectories $\mathcal{H} = \{\hat{\mathbf{X}}^{(1)}, \dots, \hat{\mathbf{X}}^{(K)}\}$ representing plausible future scenarios.

4.3 Stage 2: Logic-Enhanced Contextual Trajectory Selection

While Stage 1 provides a robust probabilistic baseline based on historical patterns, it lacks the semantic understanding to account for varying external disruptions (e.g., accidents, extreme weather). Stage 2 bridges this gap by employing a Large Language Model (LLM) as a logic-driven reasoner to evaluate and select the most plausible trajectory $\hat{\mathbf{X}}^*$ from the candidate set $\mathcal{H}$ based on real-time context $\mathbf{s}$.

4.3.1 Semantic Serialization of Probabilistic Forecasts

Since LLMs operate in the semantic space rather than the numerical tensor space, we must project the candidate set $\mathcal{H} = \{\hat{\mathbf{X}}^{(1)}, \dots, \hat{\mathbf{X}}^{(K)}\}$ into a textual representation compatible with the LLM's context window. We define a serialization function $\psi : \mathbb{R}^n \to \mathcal{S}$ that aggregates high-dimensional trajectory data into descriptive statistics. For each candidate k, we compute the statistical summary vector $\mathbf{v}^{(k)}$ over key corridors, including the minimum, median, and maximum speeds across the prediction horizon. This is formatted into a structured prompt segment:

$$\mathcal{T}_{traj}^{(k)} = \text{“Candidate } k : \text{Trend } \in [\min(\mathbf{v}^{(k)}), \max(\mathbf{v}^{(k)})], \text{Dynamics: Description}(\nabla \mathbf{v}^{(k)})\text{”} \tag{9}$$

where Description$(\cdot)$ maps numerical gradients to linguistic descriptors (e.g., “rapidly decaying,” “stable”). This allows the LLM to comprehend the traffic dynamics implied by each hypothesis without processing raw tensors.

4.3.2 Chain-of-Thought Reasoning and Selection

We leverage the reasoning capabilities of the Qwen2.5-14B-Instruct model [51] to model the causal relationship between the context $\mathbf{s}$ and traffic states. The inference process is structured as a conditional probability maximization problem via Chain-of-Thought (CoT) prompting.

Let $\mathcal{P}_{sys}$ be the system instruction and $\mathcal{T}_{ctx}$ be the textualized real-time context (e.g., incident logs, weather reports). The LLM generates a reasoning path $\mathcal{R}$ followed by a selection index k^*:

$$(\mathcal{R}, k^*) \sim P_{LLM}(\cdot \mid \mathcal{P}_{sys}, \mathcal{T}_{ctx}, \{\mathcal{T}_{traj}^{(k)}\}_{k=1}^{K}) \tag{10}$$

The reasoning path $\mathcal{R}$ explicitly decomposes the task into three logical steps:

1. **Event Impact Analysis:** The LLM parses $\mathcal{T}_{ctx}$ to extract event attributes (severity, location) and infers the spatiotemporal scope of the impact (e.g., “Lane closure on Highway A will cause upstream congestion propagation”).

2. **Hypothesis Verification:** The model compares the inferred impact against the statistical properties of each candidate $\mathcal{T}_{traj}^{(k)}$. For instance, if the event implies a significant speed drop, candidates showing “stable high speed” are rejected.

3. **Optimal Selection:** The index k^* corresponding to the candidate that maximizes semantic alignment with the reasoning $\mathcal{R}$ is selected.

The final output is the specific trajectory $\hat{\mathbf{X}}^* = \hat{\mathbf{X}}^{(k^*)}$, which represents the event-conditioned forecast.

4.4 Stage 3: LLM-Powered Reporting and Decision Suppor

The objective of Stage 3 is to transform the selected traffic forecast $\hat{\mathbf{X}}^*$ into structured, actionable outputs by incorporating historical traffic management experience. Rather than relying solely on generative inference, this stage conditions the generation process on retrieved historical cases, thereby grounding the outputs in previously observed traffic contexts and response patterns.

4.4.1 Historical Case Retrieval

We construct a historical case repository $\mathcal{K} = \{(c_m, a_m)\}_{m=1}^{M}$, where each entry consists of a past traffic context c_m (e.g., incident type and observed traffic state) and the corresponding management action a_m (e.g., signal control strategies or information dissemination measures). This repository serves as a source of empirical reference for the current scenario.

Given the current system state $\mathbf{s}$ and the selected forecast summary $\hat{\mathbf{X}}^*$, we form a joint query representation q by concatenation. A dense retrieval model is then employed to identify historical cases that are semantically similar to the current situation. Specifically, cosine similarity is computed between the query embedding $\phi(q)$ and the stored context embeddings $\phi(c_m)$:

$$\mathcal{S}_{rel} = \{(c_m, a_m) \mid \cos(\phi(q), \phi(c_m)) > \tau\}, \tag{11}$$

where $\phi(\cdot)$ denotes a pre-trained sentence encoder and τ is a similarity threshold. The top-N retrieved cases constitute the reference set $\mathcal{S}_{rel}$, which provides contextual constraints for the subsequent generation stage.

4.4.2 Structured Output Generation

The final output is produced by a conditional generator G, which integrates the traffic forecast, the current contextual description, and the retrieved historical cases:

$$\mathbf{Y} = G(\hat{\mathbf{X}}^*, \mathcal{T}_{ctx}, \mathcal{S}_{rel}). \tag{12}$$

The generated output $\mathbf{Y}$ is organized into two complementary components:

- *Situation Summary* ($\mathbf{Y}_{rep}$): a concise description of the anticipated traffic evolution derived from $\hat{\mathbf{X}}^*$, highlighting key temporal and spatial characteristics relevant to traffic management.
- *Action Suggestions* ($\mathbf{Y}_{rec}$): a set of recommended response measures informed by the retrieved cases in $\mathcal{S}_{rel}$ and adapted to the current forecasted conditions.

By conditioning the generation process on historically similar traffic scenarios, this framework encourages consistency with prior management patterns while allowing flexibility to accommodate the specific characteristics of the current forecast.

5 Experiments & Results

To rigorously evaluate the efficacy of the Chat-ITS framework, we present a comprehensive experimental analysis. We first detail the experimental setup—including datasets, implementation details, baselines, and metrics—in Section 5.1. Subsequently, we report the empirical results under routine traffic conditions (Section 5.2), during anomalous events (Section 5.3), and analyze the system's explainability (Section 5.4) and component contributions (Section 5.5).

5.1 Materials and Methods

5.1.1 Datasets and Preprocessing

This study utilizes multiple large-scale, real-world traffic and contextual datasets collected by Amap across China, primarily focusing on Beijing for foundational model training and broader regions for event context and specific analyses. All datasets cover the period from September 2023 to May 2024, unless otherwise specified.

Spatio-Temporal Traffic Data: The core dataset for training the probabilistic foundation model consists of high-resolution traffic state information for the urban core of Beijing. Raw traffic data was aggregated onto a regular grid with a spatial resolution of $500m \times 500m$. This resulted in $N = 5,797$ distinct spatial grid cells covering the main urban road network. For each cell, key traffic state variables, including traffic volume (vehicles per interval) and average speed (km/h), were computed and aggregated into 5-minute intervals. This dataset comprises approximately 1.3 billion data points, providing a comprehensive representation of urban traffic dynamics much larger than many commonly used open-source benchmarks [52].

Venue-Centric Traffic Data: To specifically capture traffic patterns influenced by large-scale public events, supplementary datasets focused on major venues were compiled.

- *Aggregated Venue Flow:* Derived from location-based service data, aggregated traffic volume information was obtained for the precise geographical boundaries and surrounding buffer areas of over 300 major venues (sports stadiums, concert halls, major tourist attractions, transport hubs) across multiple cities in China. This dataset helps model event-specific demand surges and dispersion patterns.

- *Fine-Grained Venue Grid Data:* For a subset of the venues above, traffic volume was aggregated onto a finer $100m \times 100m$ grid covering the venue. Due to the high granularity leading to sparsity, we identified and utilized data primarily from the top-20 grid cells exhibiting the highest average historical traffic volume within each venue's defined area, focusing analysis on the most relevant micro-locations.

Contextual Event Data: Real-time and historical event information, crucial for the LLM reasoning stage (Stage 2) and evaluation during anomalies, was primarily sourced from the Amap open platform APIs and associated historical logs for the relevant periods and geographical areas. This included:

- *Structured Construction Data:* A curated dataset encompassing over 10,000 construction events on highways and major arterials. Each entry typically includes precise location information (coordinates or road segment identifiers), scheduled start and end times, number and type of lanes affected (e.g., closure, partial blockage), and nature of the work.

- *Unstructured Anomaly Reports:* Traffic incident information disseminated via Amap, originating from user reports or official traffic authority alerts. These reports typically contain a natural language description of the incident (e.g., ”Accident involving two cars on Ring Road eastbound near Exit 5, blocking right lane”), an approximate location , and a timestamp. This text serves as direct input for the LLM.

Preprocessing: Standard preprocessing steps were applied to the traffic datasets before model training and evaluation. Missing values in the time-series data less than 5% of points were imputed using linear interpolation while others are dropped. Traffic state features (speed, volume) were normalized using Z-score normalization based on the mean and standard deviation calculated from the training portion of the primary Beijing dataset. For GNN-based baselines, the spatial graph adjacency matrix was constructed based on road network distance threshold, with edge weights typically defined by inverse distance. Unstructured text data from anomaly reports and event information was cleaned to remove irrelevant artifacts before being fed into the LLM prompts.

5.1.2 Baseline Implementations

The chosen baselines represent a broad spectrum of modern time series forecasting methodologies, ensuring a robust and comprehensive comparison against different architectures:

- DLinear[13]: Represents simple yet surprisingly effective linear models, serving as a strong benchmark against more complex architectures by decomposing the time series and applying separate linear layers. It challenges the necessity of intricate designs for certain forecasting tasks.

- FiLM [53]: Represents linear models enhanced with frequency analysis, designed to improve forecasting by better capturing periodicity through specific decomposition techniques applied in the frequency domain.

- Informer [54]: A prominent Transformer-based model optimized for long sequence time-series forecasting (LSTF) efficiency through a ProbSparse self-attention mechanism and distilling operation, representing efficient Transformer variants.

- PatchTST [30]: Represents channel-independent Transformer approaches utilizing patching, where input time series are divided into subseries-level patches that are fed as tokens to the Transformer, capturing local semantic information.

- Chronos [35]: Represents recent large pre-trained foundation models for time series, leveraging language model architectures scaled to time series data for zero-shot or few-shot forecasting, showcasing the potential of large-scale pre-training.
- iTransformer [31]: An innovative Transformer architecture that inverts the standard process by applying attention to embedded variates across the entire time series length, designed to better capture multivariate correlations.

For implementation, we utilized established frameworks such as TSLib [55, 56] or the official public code repositories associated with each baseline model. To guarantee a fair and direct comparison, all baseline models were trained using the identical historical dataset that was employed for training the Chat-ITS foundation model. The sole exception was Chronos, for which we leveraged the publicly available pre-trained weights, applying it directly as a zero-shot forecaster without fine-tuning on our specific dataset.

5.1.3 Evaluation Details

We evaluated the model's performance using standard time-series forecasting metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Weighted Absolute Percentage Error (WAPE). MAE measures the average absolute difference between predictions and actual values. MSE computes the average of the squared errors, which emphasizes larger deviations more heavily. RMSE is the square root of MSE, bringing the error back to the original scale of the data. MAPE and WAPE assess the relative size of errors compared to actual values, providing scale-independent evaluation.

It is worth noting that we preferred WAPE over MAPE for traffic volume forecasting tasks, where the data often contains zero or near-zero values. In such scenarios, MAPE can become undefined or unstable due to division by zero or extremely small actual values, leading to misleading evaluations. Additionally, in cases with large variance in traffic volumes, MAPE tends to overemphasize errors in low-volume periods while underrepresenting high-volume ones. In contrast, WAPE normalizes total absolute error by the sum of actual values across all time steps and locations, offering a more stable and representative metric under these conditions.

The specific metrics are defined as (13), (14), (15), (16), and (17):

$$\mathrm{MAE} = \frac{1}{nN} \sum_{t=T+1}^{T+n} \sum_{i=1}^{N} |x_{t,i} - \hat{x}_{t,i}| \tag{13}$$

$$\mathrm{MSE} = \frac{1}{nN} \sum_{t=T+1}^{T+n} \sum_{i=1}^{N} (x_{t,i} - \hat{x}_{t,i})^2 \tag{14}$$

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{nN} \sum_{t=T+1}^{T+n} \sum_{i=1}^{N} (x_{t,i} - \hat{x}_{t,i})^2} \tag{15}$$

$$\text{MAPE} = \frac{1}{nN} \sum_{t=T+1}^{T+n} \sum_{i=1}^{N} \left| \frac{x_{t,i} - \hat{x}_{t,i}}{x_{t,i}} \right| \tag{16}$$

$$\text{WAPE} = \frac{\sum_{t=T+1}^{T+n} \sum_{i=1}^{N} |x_{t,i} - \hat{x}_{t,i}|}{\sum_{t=T+1}^{T+n} \sum_{i=1}^{N} |x_{t,i}|} \tag{17}$$

Anomaly Identification: For evaluating performance during anomalies, event periods were identified using timestamps from the Amap/Gaode construction and incident logs. Anomalous periods were defined as 1 hour before to 1 hours after the logged event time for relevant locations. For public events, the anomalous period covered 2 hours before the event start to 2 hours after the event end. Zero-shot evaluation used events from categories completely held out during any training/fine-tuning.

Data Splits: Data was split chronologically for each city/dataset. Typically, the first 70% was used for pre-training the foundation model, the next 10% for validation, and the final 20% for testing.

5.2 Baseline Performance under Routine Conditions

To establish the foundational capability of our framework, we evaluated the performance of the pre-trained spatio-temporal model (Stage 1 output) under routine traffic conditions, comparing it against established baselines. The evaluation used Beijing datasets (Jan-May 2024), excluding major anomalies.

Our results, summarized in Table 1, demonstrate that the Chat-ITS foundational model substantially outperforms all evaluated deep learning baselines under these normal conditions. Across all prediction horizons (15, 30, and 60 minutes) and all metrics (MAE, MSE, WAPE), our model consistently achieved the lowest error rates. Notably, while Informer emerged as the most competitive baseline, our model still surpassed its performance considerably. For instance, the average MAE for our model (0.153) was markedly lower than Informer's (0.166) and substantially better than PatchTST (0.259) or FiLM (0.527).

Furthermore, our model displayed remarkable stability across prediction horizons, maintaining consistently low error values even for 60-minute forecasts. This contrasts with baselines like DLinear and PatchTST, which exhibited pronounced accuracy degradation as the horizon increased. This confirms that the foundation model provides a robust starting point for subsequent contextual adjustment.

5.3 Enhanced Prediction Accuracy across Diverse Anomalous Events

We evaluate the effectiveness of incorporating textual context into traffic forecasting under four representative anomalous scenarios, covering both routine disruptions and semantically ambiguous events. The proposed Chat-ITS model is compared with a foundation-model baseline that relies solely on historical traffic observations. Figure 2 shows that, when exposed to anomalous inputs, the baseline frequently produces predictions with large uncertainty, whereas Chat-ITS yields more concentrated and accurate forecasts.

Table 1: Comparison of forecasting metrics (MAE, MSE and WAPE) under routine traffic conditions for different prediction horizons (15, 30, 60 min, and average). Models compared include our pretrained spatio-temporal model and selected deep learning baselines (e.g., DLinear, FiLM, Informer, PatchTST, Chronos, iTransformer) on the Beijing dataset. Best results for each metric and horizon are highlighted.

	15 min			**30 min**			**60 min**			**Average**		
	MAE	**MSE**	**WAPE**	**MAE**	**MSE**	**WAPE**	**MAE**	**MSE**	**WAPE**	**MAE**	**MSE**	**WAPE**
DLinear	0.270	0.139	0.384	0.374	0.257	0.532	0.542	0.508	0.769	0.383	0.285	0.545
FiLM	0.423	0.301	0.602	0.534	0.491	0.759	0.669	0.772	0.949	0.527	0.496	0.748
Informer	0.161	0.060	0.229	0.167	0.066	0.237	0.176	0.074	0.249	0.166	0.066	0.236
PatchTST	0.201	0.090	0.286	0.251	0.147	0.356	0.347	0.285	0.491	0.259	0.164	0.368
chronos-b	0.423	0.427	0.561	0.342	0.323	0.495	0.291	0.325	0.425	0.424	0.489	0.602
chronos-m	0.433	0.426	0.574	0.351	0.328	0.508	0.301	0.343	0.440	0.431	0.495	0.613
chronos-s	0.439	0.443	0.582	0.349	0.325	0.505	0.311	0.374	0.454	0.436	0.504	0.620
iTransformer	0.194	0.088	0.275	0.226	0.130	0.321	0.293	0.234	0.416	0.233	0.143	0.331
Ours	**0.153**	**0.055**	**0.217**	**0.151**	**0.054**	**0.215**	**0.158**	**0.058**	**0.225**	**0.153**	**0.055**	**0.217**

Figure 2 presents illustrative examples across different event types. In common disruption scenarios such as highway construction and traffic incidents (Fig. 2(a) and Fig. 2(b)), Chat-ITS accurately captures the expected reductions in traffic volume or speed and closely follows the ground-truth trajectories. In contrast, the baseline model exhibits substantial predictive uncertainty and struggles to estimate the severity of the impact.

More pronounced differences are observed in semantically nuanced scenarios. In the parking control case (Fig. 2(c)), the baseline extrapolates historical stability and fails to anticipate a complete stop, resulting in a wide predictive interval. By incorporating event descriptions related to destination adjustment, Chat-ITS correctly forecasts a sharp speed reduction approaching zero. Similarly, in the accident with rerouting scenario (Fig. 2(d)), the baseline prediction reflects a strong prior association between accidents and congestion, leading to lower-speed estimates. Chat-ITS, however, accounts for the rerouting information and predicts sustained high-speed flow, consistent with the observed outcome.

Overall, these results indicate that integrating event-level textual context enables the model to disambiguate anomalous conditions and to revise statistically dominant patterns when additional semantic information is available.

Table 2: Forecasting accuracy improvements during anomalous events. Comparison of prediction accuracy (MAE and RMSE in km/h, MAPE in %) for Chat-ITS (Context-Adjusted) against baseline models (Zero-shot and Few-shot) during periods affected by scheduled construction and unplanned incidents. Data evaluated on the anomaly datasets across different prediction horizons. Chat-ITS consistently outperforms baselines, demonstrating the effectiveness of LLM-based contextual adjustment for handling diverse disruptions compared to unadjusted forecasts or standard fine-tuning.

	15 min			**30 min**			**60 min**			**Average**		
	MAE	**RMSE**	**MAPE**	**MAE**	**RMSE**	**MAPE**	**MAE**	**RMSE**	**MAPE**	**MAE**	**RMSE**	**MAPE**
Few-shot	5.11	8.25	18.11%	5.63	9.27	21.18%	6.36	10.57	23.56%	5.59	9.25	20.59%
Zero-shot	5.19	8.33	18.71%	5.69	9.33	21.69%	6.47	10.72	25.90%	5.67	9.34	21.38%
Chat-ITS	**4.86**	**8.03**	**16.51%**	**5.31**	**8.98**	**18.45%**	**5.94**	**10.19**	**20.58%**	**5.27**	**8.97**	**18.18%**

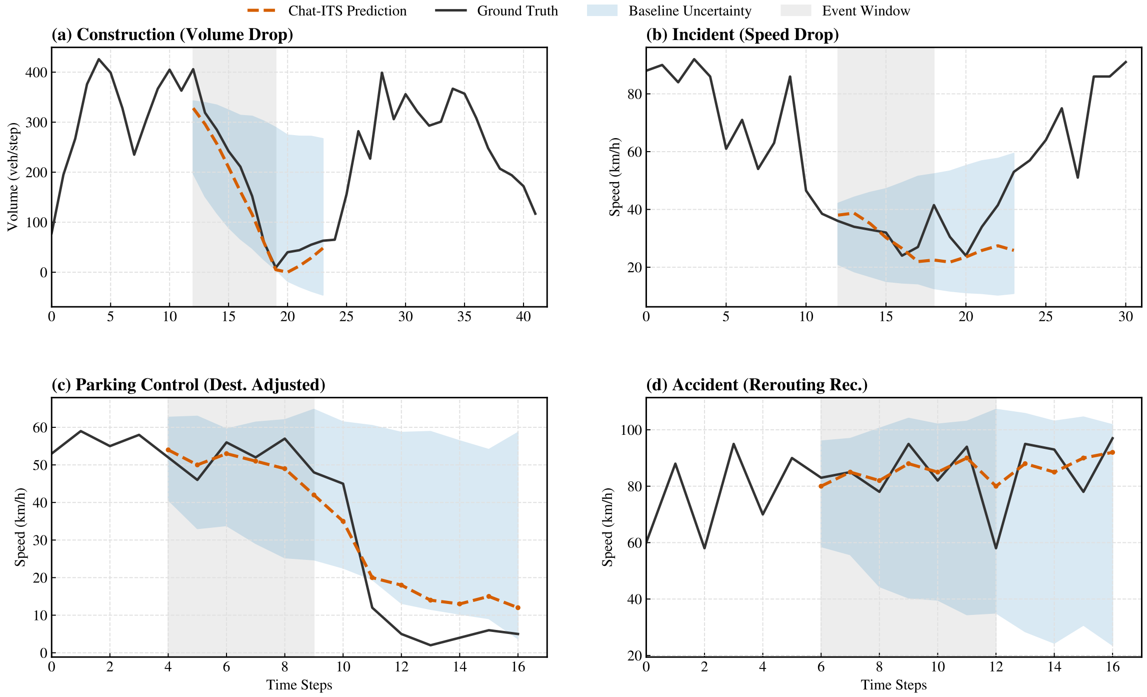


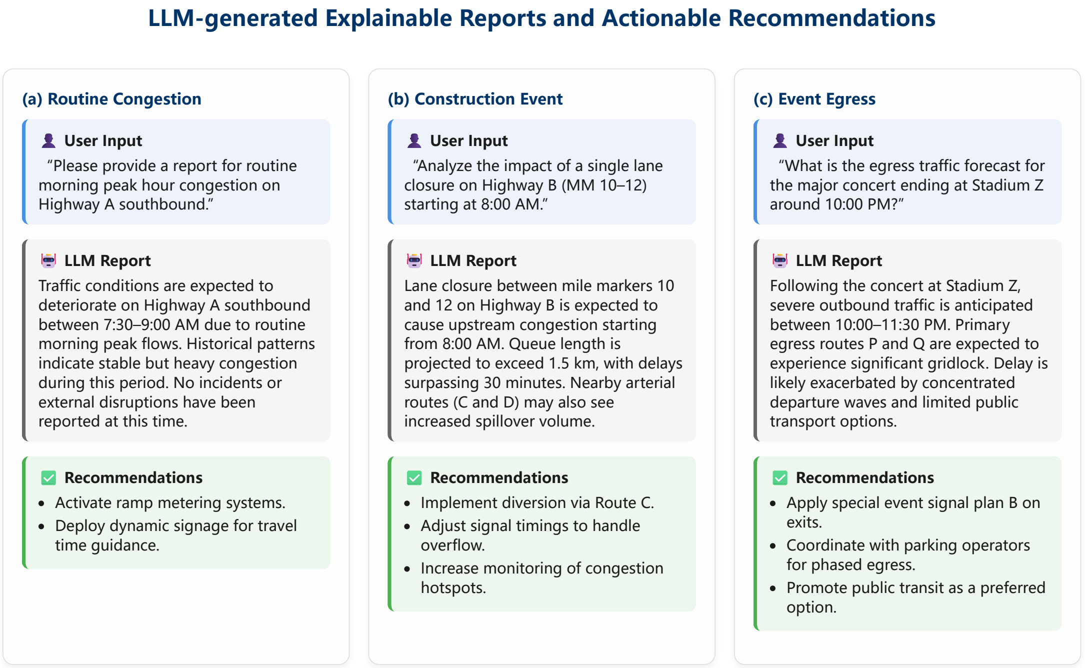
Figure 2: Time-series predictions under different anomalous traffic scenarios. Results are shown for (a) highway construction, (b) traffic incident, (c) parking control, and (d) accident with rerouting. The blue shaded region denotes the predictive uncertainty of the baseline model without textual input. Predictions from Chat-ITS are shown as an orange dashed line, while the ground truth is shown as a black solid line.

5.4 Explainable Reporting and Actionable Recommendations

Stage 3 of the framework utilizes the LLM to translate quantitative forecasts into human-readable reports. Figure 3 showcases this capability:

- **Routine Scenario:** The LLM provides concise confirmation (”Expect typical heavy congestion”) and prompts standard checks (”Ensure ramp metering active”), building trust.
- **Construction Event:** The output is analytical, synthesizing the speed drop with the ”single lane closure” context. Recommendations are proactive, suggesting ”deploying VMS advisory” and ”adjusting signal timings” on alternate routes, demonstrating reasoning about network-wide ripple effects.
- **Public Event:** The report highlights specific egress congestion timing. Recommendations like ”Monitor parking lot clearance rates” are highly tailored to the event type.

These outputs demonstrate Chat-ITS’s capacity to bridge the gap between prediction and action, transforming numerical data into operationally relevant intelligence.



LLM-generated Explainable Reports and Actionable Recommendations

(a) Routine Congestion

User Input

"Please provide a report for routine morning peak hour congestion on Highway A southbound."

LLM Report

Traffic conditions are expected to deteriorate on Highway A southbound between 7:30–9:00 AM due to routine morning peak flows. Historical patterns indicate stable but heavy congestion during this period. No incidents or external disruptions have been reported at this time.

Recommendations

- Activate ramp metering systems.
- Deploy dynamic signage for travel time guidance.

(b) Construction Event

User Input

"Analyze the impact of a single lane closure on Highway B (MM 10–12) starting at 8:00 AM."

LLM Report

Lane closure between mile markers 10 and 12 on Highway B is expected to cause upstream congestion starting from 8:00 AM. Queue length is projected to exceed 1.5 km, with delays surpassing 30 minutes. Nearby arterial routes (C and D) may also see increased spillover volume.

Recommendations

- Implement diversion via Route C.
- Adjust signal timings to handle overflow.
- Increase monitoring of congestion hotspots.

(c) Event Egress

User Input

"What is the egress traffic forecast for the major concert ending at Stadium Z around 10:00 PM?"

LLM Report

Following the concert at Stadium Z, severe outbound traffic is anticipated between 10:00–11:30 PM. Primary egress routes P and Q are expected to experience significant gridlock. Delay is likely exacerbated by concentrated departure waves and limited public transport options.

Recommendations

- Apply special event signal plan B on exits.
- Coordinate with parking operators for phased egress.
- Promote public transit as a preferred option.

Figure 3: **LLM-generated explainable reports and actionable recommendations.** Examples regarding (a) Routine Congestion, (b) Construction Event, and (c) Public Event Egress, showing the translation from context-aware prediction to operational intelligence.

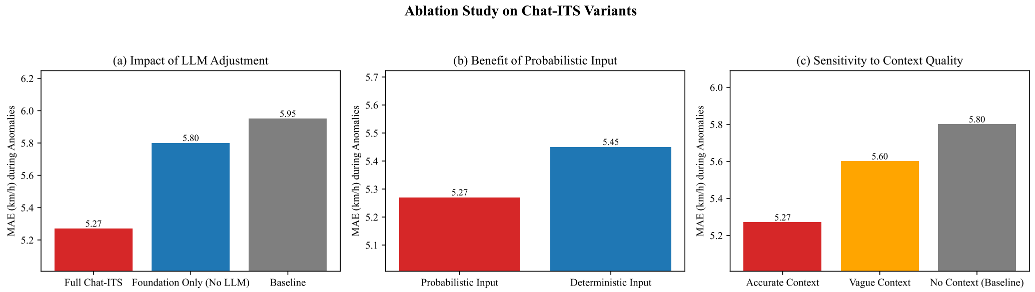


Figure 4: **Ablation study results.** (a) Impact of removing LLM adjustment. (b) Benefit of probabilistic candidates vs. deterministic input. (c) Sensitivity to context quality.

5.5 Ablation Studies

We performed ablation studies on the anomaly dataset to validate component contributions:

1. **Necessity of LLM Adjustment:** Removing Stage 2 and relying solely on the foundation model resulted in noticeable accuracy degradation during anomalies (Figure 4a), confirming the essential role of contextual reasoning.

2. **Probabilistic vs. Deterministic Input:** Providing the LLM with multiple probabilistic candidate trajectories (Stage 1 output) yielded better performance than providing a single deterministic mean forecast (Figure 4b). This suggests the probabilistic distribution offers a richer search space for the LLM to select the most contextually plausible outcome.

3. **Context Quality:** While performance dropped when input context was intentionally degraded (vague/incorrect), Chat-ITS still outperformed context-unaware baselines, indicating a degree of robustness (Figure 4c).

6 Discussion

In this work, we introduced Chat-ITS, a hybrid framework that synergistically combines a pre-trained spatio-temporal foundation model for probabilistic forecasting with the contextual reasoning capabilities of Large Language Models to address key limitations in current traffic prediction systems. Our results demonstrate that Chat-ITS achieves forecasting accuracy comparable to state-of-the-art deep learning methods under routine traffic conditions. More importantly, it marked outperforms these baselines during anomalous events, such as construction, incidents, and public gatherings, by effectively incorporating real-time structured and unstructured contextual information via LLM processing. We showed substantial error reductions (up to 15%) during such events, highlighting the framework's enhanced adaptability and resilience. Furthermore, case studies illustrated Chat-ITS's promising zero-shot generalization capability, allowing it to interpret and respond to novel event types described solely through natural language prompts. Finally, we demonstrated the framework's ability to generate explainable, human-readable summary reports and actionable traffic management recommendations, bridging the critical gap between prediction and operational decision-making.

The effectiveness of Chat-ITS stems from its deliberate hybrid design, which leverages the complementary strengths of deep learning for pattern recognition in high-dimensional spatio-temporal data and LLMs for flexible context understanding, reasoning, and natural language generation. The foundation model provides a statistically robust baseline forecast capturing complex recurring dynamics and uncertainty. The LLM, instead of being burdened with direct numerical prediction, focuses on its core strengths: interpreting diverse inputs (text, structured data), inferring causal impacts of events, evaluating scenarios based on context, and communicating findings effectively. This division of labor allows Chat-ITS to overcome the brittleness of purely data-driven models during anomalies and the limitations of purely LLM-based approaches in precise numerical forecasting. The integration of historical dispatch patterns further enhances the practical relevance of the generated recommendations. This synergistic approach represents a noticeable step towards ITS that are not only predictive but also adaptive, explainable, and operationally relevant.

The uniqueness of Chat-ITS lies in its synergistic integration of two powerful AI paradigms: deep learning for robust forecasting and LLMs for contextual reasoning. While deep learning models have pushed the boundaries of accuracy on benchmark datasets, they often lack robustness to out-of-distribution events and fail to provide actionable insights. Attempts to incorporate auxiliary data often rely on rigid input structures and struggle with unstructured text. LLM applications in transportation have primarily focused on tasks like route planning, dialogue systems, or summarizing traffic reports, but their direct application to end-to-end numerical forecasting remains challenging. Chat-ITS uniquely integrates these two powerful AI paradigms in a way that mitigates their respective weaknesses while harnessing their combined potential for context-aware, explainable, and actionable traffic prediction.

The performance of Chat-ITS during anomalies is inherently dependent on the availability and quality of real-time contextual information ($\mathbf{s}$). Inaccurate or delayed event reports will naturally limit the effectiveness of the LLM adjustment stage. While we demonstrated some robustness, further research is needed on handling noisy or conflicting contextual inputs. The reliance on LLMs also introduces computational costs associated with inference, although using smaller or optimized LLMs could mitigate this, and exploring different LLM architectures, including potentially smaller, domain-adapted models, could optimize this trade-off. Potential issues related to LLM biases or hallucinations, while mitigated by grounding the LLM's task in evaluating pre-generated trajectories, require ongoing vigilance and potentially safety layers in operational deployment. Furthermore, effective prompt engineering is crucial for eliciting the desired reasoning and output from the LLM, which may require domain expertise. Scalability to extremely large, city-wide networks with tens of thousands of sensors also needs further investigation.

Future research will focus on integrating Chat-ITS with real-time traffic control systems (e.g., adaptive signal control, variable speed limits) to enable fully automated, context-aware traffic management. We also aim to incorporate a wider range of contextual data sources, such as real-time social media feeds, advanced weather nowcasting, or connected vehicle data, to further enhance situational awareness. Developing more sophisticated methods for the LLM to not just select but actively modify forecast trajectories based on context could yield further accuracy gains. Finally, conducting user studies with traffic operators to evaluate the usability and effectiveness of the generated reports and recommendations in real-world control room settings is essential for practical validation and refinement.

Acknowledgments

This work was supported by Beijing Natural Science Foundation (No. JQ24051), Beijing Nova Program (No. 20230484432) and Independent Research Project of the State Key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University (No. ZZ-GG-20250406).

7 Data Availability

The datasets central to this study, including the large-scale traffic network data and the anomalous event logs, were provided by Amap. Due to the proprietary nature of this information, which encompasses commercial sensitivities and privacy considerations, these datasets are not publicly available. We acknowledge the importance of reproducibility and regret that these necessary restrictions prevent the public dissemination of these materials. Enquiries regarding the methodology or potential collaborations may be directed to the corresponding author.

8 Supplementary Materials

- **Text S1 and Table S1.** Detailed description and illustrative examples of the Structured Construction Data used in this study.
- **Text S2 and Table S2.** Detailed description and examples of the Unstructured Anomaly Reports.

References

1. Wu K, Ding J, Lin J, et al. Big-data empowered traffic signal control could reduce urban carbon emission. Nature Communications 2025;16. Publisher: Nature Publishing Group:2013.
2. Avila AM and Mezić I. Data-driven analysis and forecasting of highway traffic dynamics. Nature Communications 2020;11. Publisher: Nature Publishing Group TLDR: Here it is demonstrated how the Koopman mode decomposition can offer a model-free, data-driven approach for analyzing and forecasting traffic dynamics.:2090.
3. Shao Z, Zhang Z, Wang F, and Xu Y. Pre-training Enhanced Spatial-temporal Graph Neural Network for Multivariate Time Series Forecasting. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 2022:1567–77. DOI: 10.1145/3534678.3539396. arXiv: 2206.09113[cs]. URL: http://arxiv.org/abs/2206.09113 (visited on 10/24/2023).

4. Li Z, Cai R, Fu TZJ, Hao Z, and Zhang K. Transferable Time-Series Forecasting Under Causal Conditional Shift. IEEE Transactions on Pattern Analysis and Machine Intelligence 2024;46. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence:1932–49.

5. Runge J, Bathiany S, Bollt E, et al. Inferring causation from time series in Earth system sciences. Nature Communications 2019;10. Publisher: Nature Publishing Group:2553.

6. Yuan Y, Ding J, Feng J, Jin D, and Li Y. UniST: A Prompt-Empowered Universal Model for Urban Spatio-Temporal Prediction. 2024. DOI: 10.1145/3637528.3671662. arXiv: 2402.11838[cs]. URL: http://arxiv.org/abs/2402.11838 (visited on 07/15/2024).

7. Guo X, Zhang Q, Jiang J, Peng M, Zhu M, and Yang HF. Towards explainable traffic flow prediction with large language models. Communications in Transportation Research 2024;4. TLDR: This paper proposes a Traffic flow Prediction model based on Large Language Models (LLMs) to generate explainable traffic predictions, named xTP-LLM, and is the first study to use LLM for explainable prediction of traffic flows.:100150.

8. Williams AR, Ashok A, Marcotte É, et al. Context is Key: A Benchmark for Forecasting with Essential Textual Information. TLDR: A time series forecasting benchmark that pairs numerical data with diverse types of carefully crafted textual context, requiring models to integrate both modalities, and a simple yet effective LLM prompting method that outperforms all other tested methods on this benchmark. 2024. DOI: 10.48550/arXiv.2410.18959. arXiv: 2410.18959[cs]. URL: http://arxiv.org/abs/2410.18959 (visited on 01/08/2025).

9. Liu H, Xu S, Zhao Z, et al. Time-MMD: A New Multi-Domain Multimodal Dataset for Time Series Analysis. TLDR: Time-MMD is introduced, the first multi-domain, multimodal time series dataset covering 9 primary data domains, and MM-TSFlib, the first multimodal time-series forecasting (TSF) library, seamlessly pipelining multimodal TSF evaluations based on Time-MMD for in-depth analyses. 2024. DOI: 10.48550/arXiv.2406.08627. arXiv: 2406.08627[cs]. URL: http://arxiv.org/abs/2406.08627 (visited on 07/26/2024).

10. Li Z, Lin X, Liu Z, et al. Language in the Flow of Time: Time-Series-Paired Texts Weaved into a Unified Temporal Narrative. 2025. DOI: 10.48550/arXiv.2502.08942. arXiv: 2502.08942[cs]. URL: http://arxiv.org/abs/2502.08942 (visited on 02/17/2025).

11. Xue H and Salim FD. PromptCast: A New Prompt-based Learning Paradigm for Time Series Forecasting. 2023. DOI: 10.48550/arXiv.2210.08964. arXiv: 2210.08964[cs,math,stat]. URL: http://arxiv.org/abs/2210.08964 (visited on 01/17/2024).

12. Arango SP, Mercado P, Kapoor S, et al. ChronosX: Adapting Pretrained Time Series Models with Exogenous Variables. TLDR: This paper introduces a new method to incorporate covariates into pretrained time series forecasting models through modular blocks that inject past and future covariate information, without necessarily modifying the pretrained model in consideration. 2025. DOI: 10.48550/arXiv.2503.12107. arXiv: 2503.12107[cs]. URL: http://arxiv.org/abs/2503.12107 (visited on 03/27/2025).

13. Zeng A, Chen M, Zhang L, and Xu Q. Are Transformers Effective for Time Series Forecasting? 2022. arXiv: 2205.13504[cs]. URL: http://arxiv.org/abs/2205.13504 (visited on 07/25/2023).

14. Tan M, Merrill MA, Gupta V, Althoff T, and Hartvigsen T. Are Language Models Actually Useful for Time Series Forecasting? 2024. DOI: 10.48550/arXiv.2406.16964. arXiv: 2406.16964[cs]. URL: http://arxiv.org/abs/2406.16964 (visited on 08/06/2024).

15. Yuan X and Qiao Y. Diffusion-TS: Interpretable Diffusion for General Time Series Generation. In: The Twelfth International Conference on Learning Representations. 2023. URL: https://openreview.net/forum?id=4h1apFjO99 (visited on 12/13/2024).

16. Su J, Jiang C, Jin X, et al. Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review. 2024. DOI: 10.48550/arXiv.2402.10350. arXiv: 2402.10350[cs]. URL: http://arxiv.org/abs/2402.10350 (visited on 04/02/2024).

17. Wang X, Fang M, Zeng Z, and Cheng T. Where Would I Go Next? Large Language Models as Human Mobility Predictors. 2023. DOI: 10.48550/arXiv.2308.15197. arXiv: 2308.15197[physics]. URL: http://arxiv.org/abs/2308.15197 (visited on 10/30/2023).

18. Feng J, Du Y, Liu T, Guo S, Lin Y, and Li Y. CityGPT: Empowering Urban Spatial Cognition of Large Language Models. 2024. DOI: 10.48550/arXiv.2406.13948. arXiv: 2406.13948[cs]. URL: http://arxiv.org/abs/2406.13948 (visited on 08/01/2024).

19. Liu P, Guo H, Dai T, et al. Taming Pre-trained LLMs for Generalised Time Series Forecasting via Cross-modal Knowledge Distillation. 2024. DOI: 10.48550/arXiv.2403.07300. arXiv: 2403.07300[cs]. URL: http://arxiv.org/abs/2403.07300 (visited on 03/19/2024).

20. Jin M, Wang S, Ma L, et al. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. 2023. DOI: 10.48550/arXiv.2310.01728. arXiv: 2310.01728[cs]. URL: http://arxiv.org/abs/2310.01728 (visited on 10/19/2023).

21. Ma X, Tao Z, Wang Y, Yu H, and Wang Y. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C: Emerging Technologies 2015;54. Publisher: Elsevier:187–97.

22. Ma X, Zhong H, Li Y, Ma J, Cui Z, and Wang Y. Forecasting Transportation Network Speed Using Deep Capsule Networks With Nested LSTM Models. IEEE Transactions on Intelligent Transportation Systems 2021;22. Conference Name: IEEE Transactions on Intelligent Transportation Systems:4813–24.

23. Li Y, Yu R, Shahabi C, and Liu Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. arXiv preprint arXiv:1707.01926 2017.

24. Wu Z, Pan S, Long G, Jiang J, and Zhang C. Graph wavenet for deep spatial-temporal graph modeling. arXiv preprint arXiv:1906.00121 2019.

25. Yan H, Ma X, and Pu Z. Learning Dynamic and Hierarchical Traffic Spatiotemporal Features With Transformer. IEEE Transactions on Intelligent Transportation Systems 2022;23. Conference Name: IEEE Transactions on Intelligent Transportation Systems:22386–99.

26. Jiang J, Han C, Zhao WX, and Wang J. PDFormer: Propagation Delay-Aware Dynamic Long-Range Transformer for Traffic Flow Prediction. 2023. DOI: 10.48550/arXiv.2301.07945. arXiv: 2301.07945[cs]. URL: http://arxiv.org/abs/2301.07945 (visited on 02/06/2024).

27. Liu X, Liang Y, Huang C, et al. Do We Really Need Graph Neural Networks for Traffic Forecasting? TLDR: Empirical results show that SimST improves the prediction throughput by up to 39 times compared to more sophisticated STGNNs while attaining comparable performance, which indicates that GNNs are not the only option for spatial modeling in traffic forecasting. 2023. DOI: 10.48550/arXiv.2301.12603. arXiv: 2301.12603[cs]. URL: http://arxiv.org/abs/2301.12603 (visited on 01/05/2024).

28. Wang Z, Nie Y, Sun P, Nguyen NH, Mulvey J, and Poor HV. ST-MLP: A Cascaded Spatio-Temporal Linear Framework with Channel-Independence Strategy for Traffic Forecasting. 2023. DOI: 10.48550/arXiv.2308.07496. arXiv: 2308.07496[cs]. URL: http://arxiv.org/abs/2308.07496 (visited on 01/05/2024).

29. Shao Z, Zhang Z, Wang F, Wei W, and Xu Y. Spatial-Temporal Identity: A Simple yet Effective Baseline for Multivariate Time Series Forecasting. 2022. DOI: 10.48550/arXiv.2208.05233. arXiv: 2208.05233[cs]. URL: http://arxiv.org/abs/2208.05233 (visited on 01/08/2024).

30. Nie Y, Nguyen NH, Sinthong P, and Kalagnanam J. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. 2023. arXiv: 2211.14730[cs]. URL: http://arxiv.org/abs/2211.14730 (visited on 12/15/2023).

31. Liu Y, Hu T, Zhang H, et al. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. 2023. DOI: 10.48550/arXiv.2310.06625. arXiv: 2310.06625[cs]. URL: http://arxiv.org/abs/2310.06625 (visited on 10/24/2023).

32. Lin L, Shi D, Han A, and Gao J. SpecSTG: A Fast Spectral Diffusion Framework for Probabilistic Spatio-Temporal Traffic Forecasting. 2024. DOI: 10.48550/arXiv.2401.08119. arXiv: 2401.08119[cs]. URL: http://arxiv.org/abs/2401.08119 (visited on 05/09/2024).

33. Liang Y, Wen H, Nie Y, et al. Foundation Models for Time Series Analysis: A Tutorial and Survey. 2024. DOI: 10.48550/arXiv.2403.14735. arXiv: 2403.14735[cs]. URL: http://arxiv.org/abs/2403.14735 (visited on 05/07/2024).

34. Jin M, Wen Q, Liang Y, et al. Large Models for Time Series and Spatio-Temporal Data: A Survey and Outlook. 2023. DOI: 10.48550/arXiv.2310.10196. arXiv: 2310.10196[cs]. URL: http://arxiv.org/abs/2310.10196 (visited on 10/25/2023).

35. Ansari AF, Stella L, Turkmen C, et al. Chronos: Learning the Language of Time Series. 2024. arXiv: 2403.07815[cs]. URL: http://arxiv.org/abs/2403.07815 (visited on 03/18/2024).

36. Rasul K, Ashok A, Williams AR, et al. Lag-Llama: Towards Foundation Models for Time Series Forecasting. 2023. DOI: 10.48550/arXiv.2310.08278. arXiv: 2310.08278[cs]. URL: http://arxiv.org/abs/2310.08278 (visited on 10/18/2023).

37. Wu H, Hu T, Liu Y, Zhou H, Wang J, and Long M. TIMESNET: TEMPORAL 2D-VARIATION MODELING FOR GENERAL TIME SERIES ANALYSIS. 2023.

38. Gao S, Koker T, Queen O, Hartvigsen T, Tsiligkaridis T, and Zitnik M. UniTS: Building a Unified Time Series Model. 2024. arXiv: 2403.00131[cs]. URL: http://arxiv.org/abs/2403.00131 (visited on 03/18/2024).

39. Liu X, Liu J, Woo G, et al. Moirai-MoE: Empowering Time Series Foundation Models with Sparse Mixture of Experts. 2024. DOI: 10.48550/arXiv.2410.10469. arXiv: 2410.10469. URL: http://arxiv.org/abs/2410.10469 (visited on 11/12/2024).

40. Zhou T, Niu P, Wang X, Sun L, and Jin R. One Fits All:Power General Time Series Analysis by Pretrained LM. 2023. arXiv: 2302.11939[cs]. URL: http://arxiv.org/abs/2302.11939 (visited on 10/18/2023).

41. Jin M, Tang H, Zhang C, et al. Time Series Forecasting with LLMs: Understanding and Enhancing Model Capabilities. arXiv.org. 2024. URL: https://arxiv.org/abs/2402.10835v2 (visited on 05/07/2024).

42. Chang C, Peng WC, and Chen TF. LLM4TS: Two-Stage Fine-Tuning for Time-Series Forecasting with Pre-Trained LLMs. 2023. DOI: 10.48550/arXiv.2308.08469. arXiv: 2308.08469[cs]. URL: http://arxiv.org/abs/2308.08469 (visited on 10/19/2023).

43. Li Z, Xia L, Tang J, et al. UrbanGPT: Spatio-Temporal Large Language Models. 2024. DOI: 10.48550/arXiv.2403.00813. arXiv: 2403.00813[cs]. URL: http://arxiv.org/abs/2403.00813 (visited on 03/21/2024).

44. Liu C, Yang S, Xu Q, et al. Spatial-Temporal Large Language Model for Traffic Prediction. 2024. DOI: 10.48550/arXiv.2401.10134. arXiv: 2401.10134[cs]. URL: http://arxiv.org/abs/2401.10134 (visited on 01/22/2024).

45. Guo X, Zhang Q, Peng M, Zhu M, Hao, and Yang. Explainable Traffic Flow Prediction with Large Language Models. version: 2. 2024. arXiv: 2404.02937[cs]. URL: http://arxiv.org/abs/2404.02937 (visited on 04/09/2024).

46. Zhang C, Zhang Y, Shao Q, et al. ChatTraffic: Text-to-Traffic Generation via Diffusion Model. 2024. DOI: 10.48550/arXiv.2311.16203. arXiv: 2311.16203[cs]. URL: http://arxiv.org/abs/2311.16203 (visited on 12/19/2024).

47. Lee G, Yu W, Shin K, Cheng W, and Chen H. TimeCAP: Learning to Contextualize, Augment, and Predict Time Series Events with Large Language Model Agents. TLDR: Experimental results on real-world datasets demonstrate that TimeCAP outperforms state-of-the-art methods for time series event prediction, including those utilizing LLMs as predictors, achieving an average improvement of 28.75% in F1 score. 2025. DOI: 10.48550/arXiv.2502.11418. arXiv: 2502.11418[cs]. URL: http://arxiv.org/abs/2502.11418 (visited on 03/18/2025).

48. Jiang Y, Yu W, Lee G, et al. Explainable Multi-modal Time Series Prediction with LLM-in-the-Loop. TLDR: TimeXL, a multi-modal prediction framework that integrates a prototype-based time series encoder with three collaborating Large Language Models to deliver more accurate predictions and interpretable explanations, is introduced. 2025. DOI: 10.48550/arXiv.2503.01013. arXiv: 2503.01013[cs]. URL: http://arxiv.org/abs/2503.01013 (visited on 03/12/2025).

49. Wang X, Feng M, Qiu J, Gu J, and Zhao J. From News to Forecast: Integrating Event Analysis in LLM-Based Time Series Forecasting with Reflection. TLDR: This paper utilizes LLM-based agents to iteratively filter out irrelevant news and employ human-like reasoning to evaluate predictions, which enables the model to analyze complex events, such as unexpected incidents and shifts in social behavior. 2024. DOI: 10.48550/arXiv.2409.17515. arXiv: 2409.17515[cs]. URL: http://arxiv.org/abs/2409.17515 (visited on 02/28/2025).

50. Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research 2020;21:1–67.

51. Bai J, Bai S, Chu Y, et al. Qwen Technical Report. arXiv preprint arXiv:2309.16609 2023.

52. Liu X, Xia Y, Liang Y, et al. LargeST: A Benchmark Dataset for Large-Scale Traffic Forecasting. 2023. arXiv: 2306.08259[cs]. URL: http://arxiv.org/abs/2306.08259 (visited on 10/31/2023).

53. Zhou T, Ma Z, Wen Q, et al. Film: Frequency improved legendre memory model for long-term time series forecasting. Advances in neural information processing systems 2022;35:12677–90.

54. Zhou H, Zhang S, Peng J, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 12. 2021:11106–15.

55. Wu H, Hu T, Liu Y, Zhou H, Wang J, and Long M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In: *International Conference on Learning Representations*. 2023.

56. Wang Y, Wu H, Dong J, Liu Y, Long M, and Wang J. Deep Time Series Models: A Comprehensive Survey and Benchmark. 2024.