Rapid #: -25307380

CROSS REF ID:	**33103362430003412**
LENDER:	**BUB (University of Bristol) :: Main Library**
BORROWER:	**HNK (Hong Kong Univ. of Science & Technology) :: Main Library**
TYPE:	Article CC:CCG
JOURNAL TITLE:	SAE technical paper series
USER JOURNAL TITLE:	SAE Technical Paper
ARTICLE TITLE:	Construction and Application of Traffic Accident Knowledge Graph Based on LLM 2025-01-7139
ARTICLE AUTHOR:	Yingqi Hou
VOLUME:	2025
ISSUE:	
MONTH:	
YEAR:	2025
PAGES:	
ISSN:	0148-7191
OCLC #:	

Processed by RapidX:	9/26/2025 2:52:46 AM

2025-01-7139 Published 21 Feb 2025

Construction and Application of Traffic Accident Knowledge Graph Based on LLM

Yingqi Hou and Yichang Shao Southeast University

Zhongyi Han Shandong Provincial Communications Planning and Design Institute Group Co. Ltd.

Zhirui Ye Southeast University

Citation: Hou, Y., Shao, Y., Han, Z., and Ye, Z., "Construction and Application of Traffic Accident Knowledge Graph Based on LLM," SAE Technical Paper 2025-01-7139, 2025, doi:10.4271/2025-01-7139.

Received: 14 Nov 2024 Revised: 01 Dec 2024 Accepted: 08 Jan 2025

Abstract

Records of traffic accidents contain a wealth of information regarding accident causes and consequences. It provides a valuable data foundation for accident analysis. The diversity and complexity of textual data pose significant challenges in knowledge extracting. Previous research primarily relies on Natural Language Processing (NLP) to extract knowledge from texts and uses knowledge graphs (KGs) to store information in a structured way. However, the process based on NLP typically necessitates extensive annotated datasets for model training, which is complex and time-consuming. Moreover, the application of traffic accident knowledge graphs by direct information querying within the graph requiring complex commands, which leads to poor interaction capabilities. In this study, we adapt an innovative approach integrates Large Language Models (LLMs) for the construction and application of a traffic accident knowledge graph. Based on the defined schema layer of the traffic accident knowledge graph, we employ LLMs to extract knowledge from accident records and refine the extraction process by using prompts and few-shot learning mechanism. To ensure the accuracy of the extracted result, we employ a dual verification method combines self-verification of LLMs with manual inspection. Then we visualize the knowledge by using Neo4j. Finally, we explore the application of KGs within the framework of Retrieval-Augmented Generation (RAG) and construct an intelligent question-answering system. The combination of LLMs and KGs facilitates a framework of semi-automated knowledge extraction and analysis. The Knowledge Graph-Based Retrieval-Augmented Generation Question Answering System for Traffic Accidents enables complex query and answering tasks such as causation analysis and scenario generation for autonomous driving tests. The integration of KGs and LLMs not only expands the application scenarios of KGs but also reduces the risk of hallucination in responses generated by LLMs. This method efficiently Extracting information from unstructured textual data, advances the digitalization and intelligence of traffic accident management.

Keywords

Traffic Accident Knowledge Graph, Large Language Model, Question-Answer System

1. Introduction

With the development of urbanization and the continuous increase in car ownership, traffic accidents have become an extremely serious social problem. Road accidents pose a huge challenge to global public health, causing a large number of deaths and injuries every year and seriously threatening people's safety [1,2]. In urbanization process of China, while traffic safety is continually improving, the number of annual traffic accident fatalities remains significant. The number of deaths due to traffic accidents in China in 2023 has exceeded 60,000 [3]. Traffic accidents not only constitute a major cause of death but also place a heavy economic burden on societal development. Therefore, in-depth study of occurrence mechanism, preventive measures and emergency handling methods of traffic accidents are of crucial practical significance. These surveys can reduce the occurrence of traffic accidents and alleviate the severity of collisions around the world and in China [4].

The occurrence and severity of traffic accidents are influenced by various factors, including driver behavior [5,6], vehicle characteristics [7] and environmental conditions [8]. Data-driven approaches play a critical role in analyzing and predicting accident severity, employing a range of statistical models and machine learning methods. Commonly used statistical models include logistic regression [9] and multivariate regression [10], as well as machine learning methods like dynamic Bayesian networks [11] and probabilistic neural networks [12]. However, most research on traffic accident analysis and prediction based on these methods relies on manually summarized structured data. There is a lack of limited direct exploration and analysis of unstructured textual data related to traffic accidents, such as accident reports and news articles. These unstructured texts often contain rich information, including details on the time, location, participants, weather conditions and causes of accidents. Due to the diversity and complexity of textual data, the aforementioned models struggle to effectively and directly extract useful knowledge from them. Thus, utilizing unstructured data to reconstruct accident scenarios more accurately for analyzing and predicting accident causes has become a pressing issue [13].

To address the limitations in current research regarding the extraction of insights from unstructured data like traffic accident reports, Knowledge Graphs (KGs) provide an effective solution as a tool for data management and semantic analysis. The concept of the KGs was first introduced by Google in 2012 [14]. It is a directed labeled graph composed of entities, attributes and relation [15], which typically be constructed as basic units shaped like ternary groups, such as “entity-relation-entity” or “entity-attribute-attribute entity”. These ternary groups use nodes and edges to represent entities and their interrelations. The vast amount of knowledge extracted from unstructured text data can be structured and visualized by KGs, which can subsequently be employed for tasks such as question-answering, reasoning and analysis by using graph theory or machine learning techniques [16]. This structured data format has proven to be effective and is currently applied to case studies and reasoning across multiple domains, including healthcare [17], finance [18] and law [19].

In the transportation field, road traffic accident reports are often semi-structured or unstructured, which is lack of standardized descriptions and clear semantic relationships. The application of KGs for this kind of data can serve as a foundation for more comprehensive knowledge extraction and reasoning. KGs have been widely applied in various research areas within transportation, including traffic flow prediction [20, 21], travel trajectory tracking and forecasting [22, 23], vehicle emissions survey [24] and speed prediction [25]. KGs have also been found to be applied in traffic accident analysis recently. Yu et al. have developed a road traffic accident knowledge graph providing new methods and conceptual frameworks for the digitalization and refinement of traffic safety management [26]. Zhang et al. set up a traffic accident knowledge graph by following steps of knowledge requirements, modeling, extraction, storage and subsequently performed visual analysis from multiple perspectives, including accident profiling, classification, statistics and association pathways [27]. Furthermore, based on the analysis of the correlations among various factors influencing traffic accidents, Zhu et al. propose a KG-CWT-RGCNN-BiLSTM model for traffic accident prediction [28]. The aforementioned studies demonstrate the effectiveness of KGs in the field of traffic accident analysis; however, there remains several issues need to be optimized. First, the construction process of KGs in most studies relies on manually defined methods or knowledge extraction techniques based on Natural Language Processing (NLP). These methods typically require large annotated datasets for model training and involve multiple steps such as entity extraction and knowledge fusion. It makes the construction process of KGs time-consuming and poor in scalability. Second, the interactive capabilities of KGs require further optimization, as querying and reasoning with graphs often necessitate complex commands and human interpretation.

Since 2023, with the continuous development of Large Language Model (LLM) technology, LLMs have demonstrated powerful language processing capabilities in tasks such as zero-shot learning question answering. We have observed that the methods for automating the creation and processing of KGs by using LLMs, as well as enhancing the quality of LLMs responses through KGs are gradually taking shape. Currently, some studies have indicated that it is feasible to automate the creation of KGs from text by using LLMs, particularly in tasks involving entity and relationship extraction and link prediction [29, 30, 31, 32]. As black-box models, LLMs inevitably face issues of knowledge insufficiency and hallucination [32]. Providing LLMs with accurate factual knowledge from KGs as input can effectively reduce hallucination errors. In summary, the integration of LLMs and KGs for applications mainly follow three primary approaches [33]: 1) inputting knowledge stored in KGs as reasoning references to enhance the interpretability of LLMs; 2) using LLMs to optimize the construction and reasoning processes of KGs; 3) integrating LLMs and KGs in knowledge representation and reasoning tasks. Integrating LLMs with KGs enables efficient structuring of unstructured textual and leverages LLMs’ powerful conversational capabilities to utilize the knowledge stored within KGs. Retrieval-Augmented Generation (RAG) is an architecture that assists LLMs in generating responses by combining retrieving relevant information from a data source and then generating prompts to enhance LLMs' accuracy in question-answering tasks [34]. This search mechanism can effectively reduce hallucination of LLMs. When given a query, RAG retrieves information related to the query from a vectorized knowledge base and combines this retrieved information with the query to form a prompt for instructing LLMs to generate responses. However, when applying RAG to long text data, challenges arise due to the low interpretability of embeddings generated from lengthy text. RAG cannot rely solely on semantic similarity to capture structured relational knowledge, which leads to limited overall understanding and a lack of global information in application. Integrating KGs with RAG can address these limitations by using the entity and relational knowledge stored in KGs to enhance RAG’s context-sensitive retrieval capabilities. This combination enables more efficient and reliable retrieval of complex information, which can improve the accuracy of LLM-generated responses in complex question-answering applications [35,36].

In summary, although existing research has recognized the value of the rich information contained in traffic accident record text data for traffic accident analysis, and methods such as NLP have been used to extract useful knowledge from these texts [13], some studies have also constructed traffic accident knowledge graphs based on NLP processing [37]. However, the construction process of such KGs typically requires a large amount of annotated data sets for model training, which makes the construction process of KGs inefficient and poorly scalable. Furthermore, the application methods of traffic accident knowledge graphs also need to be further optimized, as these applications are mostly based on direct querying of knowledge within the knowledge graph, usually requiring complex commands and manual interpretation analysis to draw final conclusions, with poor interaction capabilities.

Thus, this study harnesses Chinese road traffic accident data, employing LLMs to facilitate the construction of KGs and to expand the application of the knowledge. The research endeavors to answer the following questions: 1) What strategies can be employed to extract vital information from accident text records for enhanced application? 2) How can the labor-intensive process of constructing KGs based on domain-specific knowledge be optimized for greater efficiency? 3) Amidst the rapid growth of LLMs, how can LLMs and KGs be synergistically integrated to enhance the performance of LLMs in traffic accident analysis? Our research aims to bridge these gaps, striving to develop a comprehensive and efficient framework that combines LLMs with KGs for the extraction and utilization knowledge in traffic accident text. This work lays the groundwork for future studies in intelligent traffic accident management and predictive analytics for traffic accidents.

In this paper, we propose a semi-automated framework based on LLMs for constructing and applying a road traffic accident knowledge graph. Leveraging prompt engineering and few-shot learning, we enhance LLMs' performance in entity and relation extraction tasks through self-verification mechanisms and human validation. Additionally, we enable more efficient application of the knowledge graph for accident analysis tasks by using RAG to tap into LLMs' powerful question-answering abilities. This end-to-end workflow, from knowledge extraction and graph construction to analytical processing, achieves an organic integration of LLMs and KGs in tasks involving knowledge representation and reasoning. To meet practical needs for accident warning and decision-making, our approach utilizes real traffic accident data from local transportation authorities. We enable the semi-automatic storage of unstructured accident data within a structured knowledge graph in the field of traffic accident analysis. Leveraging RAG technology, this integrated framework for LLM-based knowledge graph construction and application supports multiple tasks, including causality visualization, dynamic interactive data querying and intelligent question-answering system development. These applications offer deeper insights into accident scenarios and have potential for future applications, such as accident warning systems, collision scene reconstruction and scenario generation for autonomous vehicle testing.

The remainder of this paper is structured as follows. Section 2 describes the data used in this study. Section 3 presents the methods and implementation processes, including the entities and relationship extraction by using LLMs, and the construction and application of road traffic accident knowledge graph based on RAG. Section 4 provides the process of case study and application. Section 5 is the discussion and Section 6 outlines the primary work and applications of the paper, along with the future research directions.

2. Data Description

The data used in this study comes from official traffic accident reports recorded by local traffic management authorities. It includes records of 1,254 road traffic accidents between January 1, 2020, and June 12, 2022 that occurred in 13 cities in Jiangsu Province, China. In the event of a severe accident, the police are required to arrive at the scene in person to collect detailed information about the accident. They are required to follow up on the conditions of the involved individuals until their recovery or unfortunate death, summarizing into a comprehensive textual record from the scene to the consequences of the accident.

The dataset consists of semi-structured and unstructured data. The key parameters contained in the dataset are shown in Table 1, including accident ID, occurrence time, location, collision type, main cause of accidents, specific crash cause and a textual description of accidents. The data is highly authentic and accurate with substantial research value. The proportional distribution of the number of accidents for three types of accidents in the dataset and the specific quantities of accidents corresponding to the five main causes are shown in Figure 1.

3. Methodology

In this study, we employ an innovative approach that integrates LLMs for the construction of a traffic accident knowledge graph. Initially, we establish the schema layer of the knowledge graph, defining its structural framework. Subsequently, we utilize LLMs to extract knowledge from traffic accident text records, refining the extraction process through the implementation of prompts and few-shot learning techniques. To ensure the accuracy of the extracted knowledge, we employ a dual verification method, combining LLM self-verification with manual

TABLE 1 Data description.

Attribute	Type	Values
Accident ID	Continuous	Integer.
Time	Timestamp	Data and time.
Location	String	Location information.
Crash Description	String	The detailed description of road traffic accidents, including time, place, person, vehicle type, behavior and consequence.
Collision Type	Categorical	0 = Head-on collision 1 = Side collision 2 = Rear-end collision
Main Cause Category	Categorical	0 = Driver status issues Categorical 1 = Improper driving behavior 2 = Traffic violation 3 = Vehicle mechanical faults 4 = Effect of environment and traffic condition
Specific Crash Cause	String	122 kinds of detailed causation.

inspection. Following this, we integrate the validated knowledge into the knowledge graph and employ Neo4j for visualization purposes. Finally, we explore the application potential of the knowledge graph within the framework of RAG.

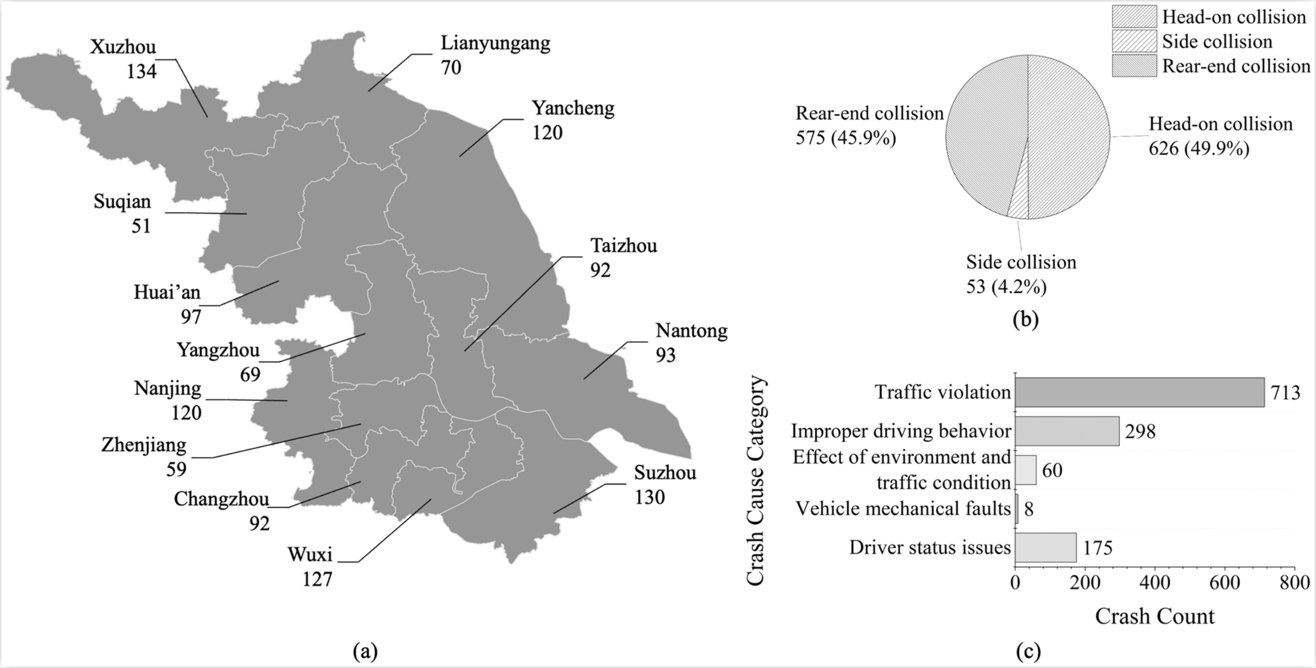
3.1. The Design of Conceptual Layer of KG

The ontology of a knowledge graph is a structured representation of concepts and their relationships within a specific domain, providing the theoretical framework for the decomposition of elements and attributes in the data layer [38]. The conceptual layer defines the hierarchical structure and relationships of key knowledge components, including entities, relationships, and attributes. These are typically expressed in the form of triples, such as "entity-relation-entity" or "entity-relation-entity attribute".

Starting from the practical needs for accident warning and decision-making, this study constructs the schema layer of traffic accident knowledge graph based on the essential elements embedded within accident text descriptions. In the process of defining entities, relationships, and entity attributes, we apply Stanford University's Seven-step Method to develop a top-down conceptual framework. This approach ensures a clear and scalable structure for the knowledge graph. The specific steps include: identifying the domain and scope of ontology; considering the reuse of existing ontologies; listing important terms within the ontology; defining classes and their hierarchical structure; specifying properties for each class slot; defining facets for these slots and creating instances [39].

We develop a framework for traffic accident knowledge graph based on domain expertise in traffic accident analysis and Stanford's Seven-step Method. The schema layer consists of two hierarchical levels. The first level organizes accident causes into five major categories. Derived from a synthesis of previous traffic accident studies, we further subdividing them into 122 specific causes. Each specific cause is linked to accident cases through "cause" relationship. This can present commonly causes of accidents in a more practical format, thus enhancing the applicability of the knowledge graph in traffic accident analysis tasks. The first layer includes three entity types: main cause category, specific cause, and accident case; two types of relationship, include "cause"

FIGURE 1 (a) crash distribution in Jiangsu province; (b) collision type distribution and (c) crash cause distribution.



and "include". The knowledge stored in this layer comes from the causes of accidents defined by traffic management staff when registering traffic accident records in the original dataset.

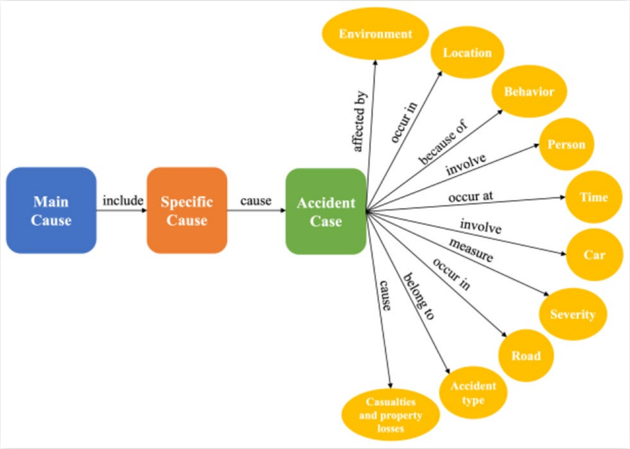
The second level of the graph structure elaborates on the specific information associated with individual accident cases. Drawing on textual records within the accident case dataset, we apply Stanford's Seven-step Method to identify 10 entity types, 8 relationship types and 15 attribute entity types. This level captures detailed information on each accident, including accident type, severity, casualties and property losses; and contextual details including time, location, road, environment, individuals and cars involved and so on. These entities in this layer are used to store the knowledge extracted by LLMs from accident records.

The structure of the schema layer of traffic accident knowledge graph is illustrated in Figure 2, comprising a total of 13 entity types, 10 relationship types and 15 entity attributes. Detailed descriptions of each ontology type are provided below.

3.1.1. Entity Explanation Specific Cause: When recording accident details, staffs of traffic management authorities document the causes based on the specialized accident dictionary developed by Chinese Ministry of Transport. The accident records contained in dataset involves 122 specific causes from accident dictionary. Analyzing these specific causes aids in designing targeted accident prevention measures.

Main Cause: Given the numerous specific causes, we define five main categories to facilitate more comprehensive analysis. Drawing on extensive experience in traffic accident analysis, primary accident causes can be grouped into three domains: human, vehicle and environmental factors. Vehicle and environmental factors are further specified as "vehicle mechanical faults" and "effect of environment and traffic condition". For in-depth analysis of human factor, we categorized human behavior into three subcategories: 1) driver status issues categorical, such as fatigue driving; 2) improper driving behaviors, such as failing to reduce speed at intersections; and 3) traffic violations, such as driving against the flow of traffic. As illustrated in Figure 3, 122 specific causes are classified into five main cause categories.

FIGURE 2 The structure of traffic accident knowledge graph.



Accident Case: Represents the primary node that records text descriptions pertinent to each accident case.

Casualties and Property Losses: Represents the outcomes of traffic accidents, indicated by the number of fatalities, number of injuries and direct economic losses.

Severity: Denotes the severity level of a traffic accident. The severity is categorized according to China's current grading standards for accident severity. There are four levels, including minor, general, major and catastrophic. The severity is primarily determined based on casualties and property losses with classification standards shown in Table 2.

Location: Records information about the location where the accident occurred.

FIGURE 3 The classification relationship between specific cause and main cause.

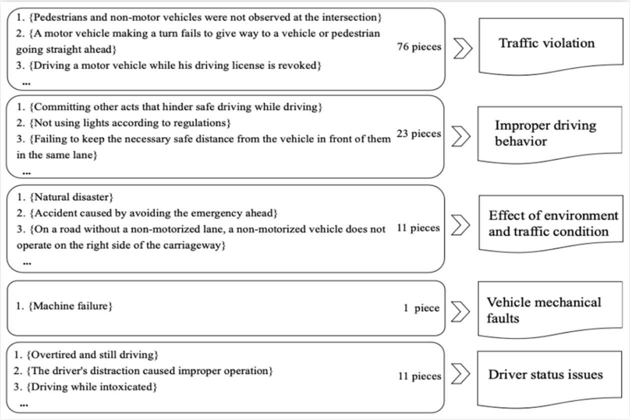

1. {Pedestrians and non-motor vehicles were not observed at the intersection}
2. {A motor vehicle making a turn fails to give way to a vehicle or pedestrian going straight ahead}
3. {Driving a motor vehicle while his driving license is revoked}
...
76 pieces
Traffic violation
1. {Committing other acts that hinder safe driving while driving}
2. {Not using lights according to regulations}
3. {Failing to keep the necessary safe distance from the vehicle in front of them in the same lane}
...
23 pieces
Improper driving behavior
1. {Natural disaster}
2. {Accident caused by avoiding the emergency ahead}
3. {On a road without a non-motorized lane, a non-motorized vehicle does not operate on the right side of the carriageway}
...
11 pieces
Effect of environment and traffic condition
1. {Machine failure}
1 piece
Vehicle mechanical faults
1. {Overtired and still driving}
2. {The driver's distraction caused improper operation}
3. {Driving while intoxicated}
...
11 pieces
Driver status issues

TABLE 2 The classification standards of severity level.

Level	Classification Standard
Minor	Refers to an accident causes minor injuries to 1 or 2 people or resulting in property damage of less than 1,000 RMB for motor vehicle accidents and less than 200 RMB for non-motor vehicle accidents.
General	Refers to an accident causes serious injuries to 1 or 2 people, minor injuries to more than 3 people, or resulting in property damage of less than 30,000 RMB.
Major	Refers to an accident causes the death of 1 or 2 people, serious injuries to more than 3 but fewer than 10 people, or property damage exceeding 30,000 RMB but less than 60,000 RMB.
Catastrophic	Refers to an accident causes the death of more than 3 people, serious injuries to more than 11 people, or the death of 1 person along with serious injuries to more than 8 people, or the death of 2 people along with serious injuries to more than 5 people, or property damage exceeding 60,000 RMB.

Time: Records information about the time of the accident, including specific date and time. Additionally, the time is segmented into six periods within a day. The attribute entities included and definitions are shown in the Table 3.

Environment: Records information about the surrounding environment of the accident.

Car: Represents the vehicles involved in the accident, documenting specific attribute entities contain ID, type and direction of vehicles.

Road: Details about the road where the accident occurred, with road types categorized into seven specific classes, including expressway, national highway, provincial highway, county highway, urban expressway, urban road and other road.

Accident Type: Classifies the type of traffic accident, with categories including head-on collision, rear-end collision and side collision.

Person: Represents individuals involved in the accident, containing specific attribute entities including name, gender, age, condition and mode of transportation. The attribute entities included and classification standard are shown in Table 4.

Behavior: Records behavioral information extracted from textual data regarding the actions of individuals involved, primarily focusing on pre-accident behavior.

TABLE 3 Attribute entities and definitions of time.

Attribute Entity	Definition
Date	Specific date of traffic accident.
Time	Specific time of traffic accident.
Period	Time period during which the accident occurred. The classification standard is as follows: -early morning: 0:00 to 5:00; -morning: 5:00 to 8:00; -late morning: 8:00 to 11:00; -noon: 11:00 to 13:00; -afternoon: 13:00 to 17:00; -evening: 17:00 to 19:00; -night: 19:00 to 24:00.

TABLE 4 Attribute entities and definitions of person.

Attribute Entity	Definition
Name	Specific information of the person's name.
Gender	Specific information of the person's gender.
Age	Specific information of the person's age.
Transportation mode	The mode of transportation chosen by the person involved, categorized as driving motor vehicle, riding in motor vehicle, driving non-motor vehicle, riding in non-motor vehicle, and walking.
Condition	The condition of the person involved after the accident, categorized as deceased, injured, uninjured and unknown.

3.1.2. Relationship Explanation

The relationships and corresponding definitions in traffic accident knowledge graph are shown in Table 5.

This study utilizes three different data types to record various entities, contains text, numerical and categorical data. Numerical data is primarily used to represent casualty statistics and economic losses; categorical data is used to classify accident causes, accident types, accident severity levels, road types and so on; the remaining entity information is recorded as text data.

3.2. Automatic Knowledge Extraction of KG Based on LLM

Knowledge extraction at the data layer involves extracting data corresponding to entities and entity attributes defined in the conceptual layer. This information is recorded in the form of triples to enable the storage of knowledge in the knowledge graph database. This process relies on techniques such as knowledge extraction and fusion. Specifically, knowledge extraction is divided into named entity recognition and relationship extraction.

TABLE 5 Relationships and definitions in traffic accident knowledge graph.

Relationship	Definition
Include	The relationship between main cause and specific cause, which indicates that an entity encompasses specific elements as part of its structure or details.
Cause	The relationship between specific cause and accident case, as well as accident case and casualties and property losses, denotes a relationship where one factor leads to or results in another event or condition.
Measure	The relationship between accident case and severity, refers to the quantification or assessment of a specific attribute or parameter.
Occur in	The relationship between accident case and location or road, specifies the broader context or environment where an accident takes place.
Occur at	The relationship between accident case and time, denotes the precise time where an accident happens.
Affected by	The relationship between accident case and environment, indicates an influence or impact on an entity by environment factor or condition.
Involve	The relationship between accident case and car or person. Represents entities or individuals participating in or associated with an accident.
Belong to	The relationship between accident case and accident type, means being part of or owned by something.
Because of	The relationship between accident case and behavior, represents that something happens because of a particular behavior.
Attribute	The relationship between entities and their attributes, regarding as a characteristic or inherent part of someone or something.

The data layer of the knowledge graph stores specific information in the form of triples, typically represented as:

$$G = \left\{ E,R,T \right\} \tag{1}$$

In this context, G represents the knowledge graph, and E denotes the set of entities, which is represented as a node within the graph. R represents the set of relationships r, depicted as edges between nodes in the knowledge graph, symbolizing connections between different entities. T denotes the set of facts, with each fact defined as a triple $\{h,\ r,\ t\} \in f$, where h is the head entity, r is the relationship, and t is the tail entity.

The key steps in knowledge graph construction involve extracting knowledge from unstructured text data and representing it as triples suitable for storage within the knowledge graph. Currently, the most common approach to automate this process is using statistical machine learning and deep learning models. These techniques are specifically relying on manually annotated data based on predefined rules for model training. Common models include BiLSTM-CRF, BERT-CRF, and BERT-BiLSTM-CRF. Deep learning-based knowledge extraction typically consists of four main steps: tokenization, named entity recognition, relation extraction and triple generation. Among these, entity recognition and relation extraction require a substantial amount of manually labeled data to train the models effectively. The annotation process demands considerable time investments to guarantee the quality and reliability of the data, and to create models with superior performance.

This paper utilizes the powerful comprehension abilities of LLMs to reduce the manual workload in the knowledge extraction process. LLMs derive their language understanding from pre-training on extensive domain-specific data, capturing ontological structures that include various entities, attributes, and relationships. Through this training, LLMs learn semantic patterns and relationships between entities via an implicit mapping of text entities and relations into internal representational spaces. It enables a deep semantic understanding of natural language. Currently, automatic annotation of unstructured text data using LLMs is in an exploratory stage. Some researches have shown the feasibility of this application, demonstrating that effective results can be achieved without the need for extensive manually labeled data for model training. Besides, LLMs can identify entities and relationships simultaneously. It removes the need to separate these tasks and thus avoids the complex interactions and error propagation commonly seen between named entity recognition (NER) and relation extraction (RE) in traditional methods. This approach simplifies model architecture and increases processing speed.

To mitigate hallucinations in LLMs and enhance interpretability, this paper employs a dual strategy of prompt-based fine-tuning and few-shot training of LLMs to identify accident details and extract relationships from traffic accident records. By inputting background knowledge and iteratively training the model through prompt-based dialogue, the LLM performs self-supervised learning on a limited set of labeled data, continuously enhancing its performance in traffic accident analysis task. Additionally, the process is monitored and accuracy reinforced by using self-checking mechanisms within LLM and a human-in-the-loop (HITL) to oversee extraction and validate results.

3.3. The Design of Prompts for LLM

Prompt can be used to stimulate the reasoning abilities of LLMs, allowing users to guide LLMs in handling complex tasks and extracting key information [40,41]. This approach serves as an end-to-end processing mechanism, enabling LLMs to directly extract the required triples from text, significantly reducing the need for manual annotation. In this paper, we design appropriate prompts to input domain-specific knowledge and task requirements related to traffic accidents into pre-trained LLMs, thereby improving the ability of handling unstructured accident-related texts.

A prompt is a natural language input sequence for LLMs, which need to be designed according to the specific task requirements. Prompts may contain multiple components, such as background information, instructions, and examples. Background information provides relevant task knowledge to help LLMs understand the task requirements. Instructions are typically short phrases that describe key information of tasks, such as goals and objectives. Examples help clarify the desired output format and content for LLMs. By designing and optimizing prompts and engaging in iterative interactions with LLMs, the quality of output results can be improved. In terms of prompt design, this paper adopts a step-by-step prompt framework. The framework consists of four components: role setting, task instruction, domain knowledge guidance and few-shot learning.

1. Role setting: This component involves using prompts such as “You are an expert in construction of a knowledge graph for traffic accidents” to specify the role LLMs should adopt. This helps LLMs better align with the specific task requirements.
2. Task instructions: It is divided into three parts: 1) clarifying the task, 2)listing the requirements, and 3)setting the standards. The goal is to define the objectives and expectations of tasks, outline the necessary steps and processes for completion and specify the rules to follow. Additionally, by setting output format standards and evaluation criteria, we can ensure the consistency of the results better.
3. Domain knowledge guidance: Building upon the extensive general knowledge acquired by LLMs during pre-training, domain-specific information relevant to traffic accident analysis is incorporated by prompts. It includes background knowledge from domain-specific sources, such as official

dictionaries of entries in traffic accident reports, classifications of accident types, and criteria for measuring accident severity. Annotated corpora are used to demonstrate how entities are identified and extracted from data, enabling LLMs to better understand and apply related concepts. Additionally, a structured schema of predefined entities and relationships is supplied as standardized input for LLMs.

4. Few-shot learning: Specific examples are provided to guide LLMs learn from fewer annotated samples than typically required for training NLP models. These examples help the model grasp domain characteristics and data features associated with the task, enabling it to generate outputs in alignment with the predefined entity and relationship schema.

To enhance extraction performance, we employ iterative sampling with LLMs based on small-sample datasets. After analyzing each output, the prompt content is refined and re-entered into LLMs. Through iterative instructions, small-sample experiments, comparative analyses and incremental optimization, the model progressively acquires a more nuanced understanding of each operational step until achieves the desired performance level in knowledge extraction. This approach significantly enhances the LLMs' analytical and operational capabilities for the given task.

3.4. Correction and Evaluation of LLM Extraction Results

LLMs encounter difficulties in maintaining a balance between text comprehension and generation during entity extraction tasks, which predisposes them to committing errors. To ensure accuracy and reliability, it is essential to conduct both LLMs' self-validation and human review of the extracted knowledge. During self-validation, LLMs concentrate on assessing the correctness of entities by relying solely on text comprehension, thereby minimizing the need for text generation. This approach allows the model to concentrate more effectively on potential errors in entity extraction. To implement self-validation, the LLM re-evaluates its extracted results by reprocessing segments of the output. Prompts like "Is the above relationship assessment correct or incorrect?" guide LLMs to verify its own output. Given the complex and opaque mechanisms underlying LLMs' operations, human semantic analysis and rationality checks are subsequently applied to the final output to further enhance validity of results. This combined approach not only improves the quality of the extracted knowledge but also aids in evaluating whether logical inconsistencies have occurred during the LLMs' task execution.

For evaluating the output results of knowledge extraction, commonly used metrics include precision, recall and the F1-score.

Precision measures the proportion of correctly identified entities out of all entities recognized by the model. The calculation formula is:

$$Precision = \frac{Number\ of\ correctly\ identified\ entities}{Total\ number\ of\ entitie\ identified\ by\ the\ model} \tag{2}$$

Recall assesses the proportion of correctly identified entities out of all entities present in the text. The calculation formula is:

$$Recall = \frac{Number\ of\ correctly\ identified\ entities}{Total\ number\ of\ entitie} \tag{3}$$

The F1-score, as the harmonic mean of precision and recall, provides a balanced assessment of the model's overall performance. The calculation formula is:

$$F1 = \frac{2 * precision * Recall}{Precision + Recall} \tag{4}$$

3.5. Knowledge Storage and Visualization

In this study, Neo4j is employed to facilitate knowledge storage and graph visualization. Neo4j is a high-performance NoSQL graph database optimized for storing structured data on networks based on Java. Its scalable and flexible architecture supports dynamic additions, deletions and queries on nodes and relationships, making it well-suited for visualization of KGs.

Neo4j's built-in graph theory algorithms, including shortest path, breadth-first search, and depth-first search, facilitate efficient graph search and pattern matching, enabling rapid retrieval and response to knowledge graph queries. By utilizing the Neo4j driver in Python, we can establish a seamless connection between Python and the Neo4j database, enabling automated data extraction and storage within the graph database. This setup facilitates both the storage and visualization of knowledge within the graph. In this knowledge graph, nodes represent entities, while edges represent the relationships and associations between entities.

3.6. Application of KG

3.6.1. Information Query with Cypher

When the knowledge graph is applied, the multidimensional data stored in it facilitates the visualization of accident profiles across various dimensions, including temporal, spatial, causal, and outcome aspects. By retrieving the causes of accidents, the system could intuitively present real historical accident data, facilitating timely access to specific case details such as the time and location of the accident, weather conditions, and the individuals and vehicles involved. The querying process primarily utilizes the Cypher query language, for example, using "MATCH (a:Accident_Case)-[:SURROUND_BY]->(e:Environment) RETURN a,e. LIMIT 20" can retrieve the extracted environmental information and display 20 data entries. This approach requires users to learn how to use Cypher.

Sometimes it is challenging for users to handle complex retrieval and analysis tasks directly.

3.6.2. A Knowledge Graph Question-Answering System Based on RAG We introduce Retrieval-Augmented Generation (RAG) at the application layer of KGs. RAG, developed by the FAIR team in 2020, combines information retrieval and language generation to enhance the performance of LLMs on question-answering tasks and reduce potential hallucinations in responses [34]. Generally, implementing RAG requires a vector database constructed from domain-specific data. Building vector database involves data preparation, text segmentation and vectorization. Text segmentation can be achieved by using models like LangChain. Vectorization converts the process of translating text data into a matrix of vectors. Open-source models such as m3e-base and ChatGPT-Embedding are often be utilized for this task. In the application phase, the system responds to user queries by efficiently retrieving relevant background knowledge from the vector database to generate prompts for instructing LLMs to generate answers. In this step, common retrieval methods include similarity search and full-text search. Prompts generated from the retrieved knowledge typically contain a task description, relevant background information and the user query in the form of a prompt. The general framework of RAG is shown in Figure 4.

In KG-based dialogue tasks, structured knowledge stored within the graph enables graph-based retrieval generation to produce superior results compared to context retrieval relying solely on semantic similarity [42]. The structured format of KGs allows LLMs to better understand and leverage relationships between different pieces of information. By applying the RAG framework, keywords extracted from the query guide knowledge retrieval within the knowledge graph. Simultaneously, while performing retrieval from a knowledge-based vector database, the system queries in the knowledge graph database and locates relevant subgraph information [43]. This subgraph information is then integrated as background knowledge with the query to generate a more specific and accurate prompt for LLMs. The prompt thus incorporates information embedded within the knowledge graph subgraph relevant to the query. Then LLMs generate an inference result based on the knowledge within the knowledge graph when receiving the Q-A task. This approach highlights how KGs enhance the accuracy of LLMs' retrieval tasks. The implementation process is primarily divided into three stages: 1) graph-based indexing; 2) graph-guided retrieval and 3) graph-enhanced generation, as shown in Figure 5.

FIGURE 4 Framework of Retrieval-Augmented Generation.

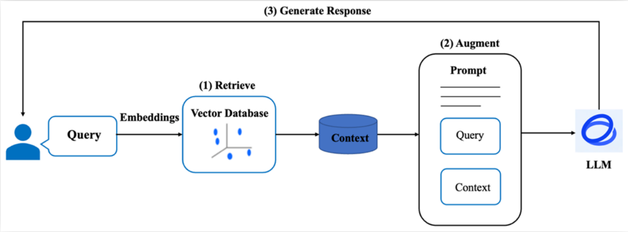
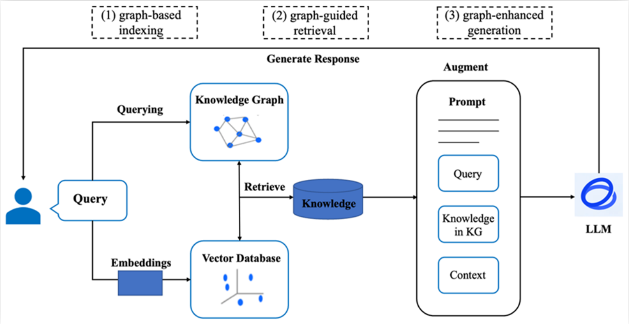


FIGURE 5 Framework of Retrieval-Augmented Generation based on Knowledge Graph.



4. Case Study & Results

4.1. Data Preprocessing

The initial step involves examining the raw dataset to eliminate accident records with missing information, such as lacking details on accident consequences. This refinement yields a final dataset of 1,186 official traffic accident records. The main cause and specific cause of each accident with structured data form official annotations is directly integrated into the knowledge graph. Other details such as accident time, location, consequences, and the vehicles and individuals involved are embedded within the unstructured textual content of each accident record. Knowledge extraction techniques are applied to convert these unstructured text elements into structured data in the form of triplets for knowledge graph construction.

4.2. Knowledge Extraction

This study utilizes ChatGLM4 to extract knowledge from accident reports. It is an open-source LLM developed by Tang Jie's team from Tsinghua University [44]. ChatGLM4 is a bilingual (Chinese-English) conversational LLM that performances as well as GPT-4-turbo in Chinese-dominated applications. It incorporating Chinese culture and values during the training process, which enabling it to better understand and express content with Chinese characteristics. Compared to commercial models like ChatGPT, ChatGLM4 excels across various Chinese and English datasets while maintaining a relatively smaller parameter size and lower training costs, which is more suitable for academic research.

To process unstructured accident report texts, prompts are generated based on our defined knowledge graph schema, sample extraction instances and output formats. Through iterative instruction fine-tuning, ChatGLM4's capability for automated knowledge extraction is continuously refined. Additionally, in order to be adapt for event-relation extraction tasks, specific trigger rules for certain relationship types are introduced into ChatGLM4. These rules define criteria such as temporal segmentation based on time information and severity assessment based on casualty numbers. The format designed for prompt is shown in the Figure 6.

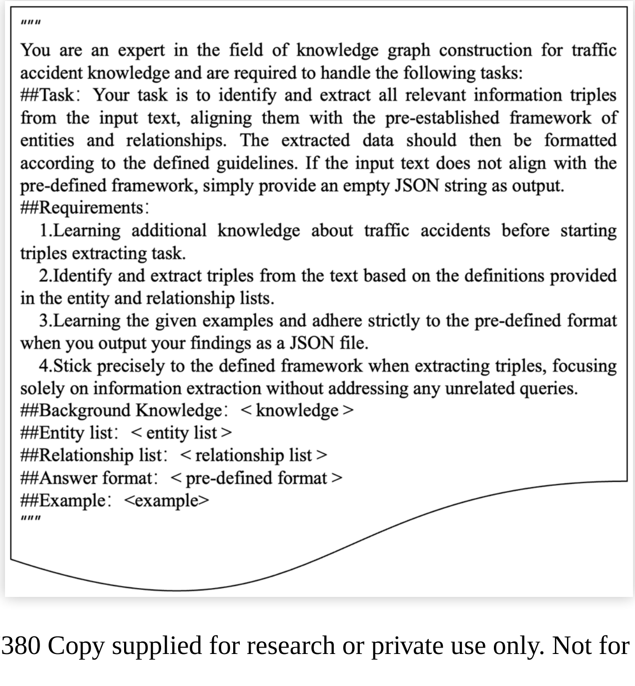
4.3. Knowledge Correction and Fusion

To enhance the construction of traffic accident knowledge graph, knowledge is organized and standardized through self-validation by ChatGLM4 and manual verification. This process involved three key steps: 1) entity alignment, 2) relationship alignment and 3) conflict resolution. Entity alignment ensures the uniqueness of entities within the knowledge graph to prevent redundancy. Relationship alignment standardizes different expressions of the same relationship type to maintain relational accuracy and completeness. Conflict resolution addresses potential contradictions and inconsistencies in the data through appropriate strategies.

Initially, disambiguation is performed on extracted entities to ensure consistency in entity representation. For example, in cases where “Feitian Avenue” and “Feitian Avenue in Jiangning District” refer to the same road. After training, ChatGLM4 could recognize these references as the same entity based on contextual understanding and consolidate them under the unified name “Feitian Avenue”. This step significantly improved the consistency of the results.

Secondly, ChatGLM4 is optimized to deduplicate synonymous relationship expressions in the extracted results. Under the instruction for recognizing synonymous relationships, ChatGLM4 is guided to understand the equivalence of relationship like “caused” and “resulted in” and categorize them under a unified relational label. This also enhanced the ability to extract information about the consequences of accidents from report texts of ChatGLM4.

FIGURE 6 The format designed for prompt.



"""

You are an expert in the field of knowledge graph construction for traffic accident knowledge and are required to handle the following tasks:

##Task：Your task is to identify and extract all relevant information triples from the input text, aligning them with the pre-established framework of entities and relationships. The extracted data should then be formatted according to the defined guidelines. If the input text does not align with the pre-defined framework, simply provide an empty JSON string as output.

##Requirements：

1.Learning additional knowledge about traffic accidents before starting triples extracting task.

2.Identify and extract triples from the text based on the definitions provided in the entity and relationship lists.

3.Learning the given examples and adhere strictly to the pre-defined format when you output your findings as a JSON file.

4.Stick precisely to the defined framework when extracting triples, focusing solely on information extraction without addressing any unrelated queries.

##Background Knowledge：< knowledge >

##Entity list：< entity list >

##Relationship list：< relationship list >

##Answer format：< pre-defined format >

##Example：<example>

"""

During the experimental process, we observed a recurring issue in severe accident records where the same individual is often first recorded as injured and later as death due to unsuccessful medical treatment. Typically, these records of injurie and subsequent death are described in separate sentences. In the initial training stages, ChatGLM4 frequently made errors by counting both injury and death occurrences for the same individual, resulting in an inaccurate assessment of the accident's severity. To address this issue, we refined the prompts by instructing ChatGLM4 to “Read the entire text and extract the final state of each individual involved as the result of condition, thereby correcting the injury and death counts.”. We also used prompts such as “Determine whether the statement ‘The < final condition > of < name of a person > is injured’ is correct.” to guide the model re-evaluate the Q-A pairs. These prompts adjustment helped guide ChatGLM4 to avoid logical errors and conflict in counting. After multiple rounds of interaction, the likelihood of ChatGLM4 repeating this error significantly decreased.

After multiple rounds of instruction fine-tuning, the performance of ChatGLM4 improved significantly. As shown in Table 6, The performance of the fine-tuned ChatGLM4 in entity extraction is comparable to BERT-BiLSTM-CRF, one of the most commonly used supervised knowledge extraction models known for its effectiveness in text-based extraction tasks.

This demonstrates the overall robust performance of ChatGLM4 in traffic accident knowledge extraction tasks. It is partly due to the structured presentation of information in accident reports, where details are typically provided in a fixed sequence: time and location of accidents, followed by the basic information of the individuals involved and their respective vehicles, possible causes of the accident, and finally, the consequences. In future work, more diverse datasets could be introduced to further evaluate the performance of ChatGLM4.

4.4. Construction and Visualization of Traffic Accident Knowledge Graph

Based on the designed framework of knowledge graph schema, the construction of the traffic accident knowledge graph data layer is carried out in two steps. First, for structured data such as main cause, specific cause, and accident case, data and entities are aligned and then

TABLE 6 Performance comparison between ChatGLM4 and BERT-BiLSTM-CRF.

Model	Precise	Recall	F1
ChatGLM4	0.848	0.892	0.869
BERT-BiLSTM-CRF	0.851	0.904	0.877

batch-imported into Neo4j storage by using Python. Second, ChatGLM4 is employed to extract and verify knowledge from the detailed information of each accident case. This process produces a JSON string that contains entity and relationship extraction results formatted as required. The output includes structured data for all accident cases, aligning with the defined entities and relationships in the knowledge graph conceptual layer. The data is then stored in the graph and linked to corresponding accident case nodes, completing the construction of the traffic accident knowledge graph. This graph comprises 16,132 entities and 28,091 relationships.

Using the visualization tool Neo4j, a partial view of traffic accident knowledge graph is displayed in Figure 7. Nodes of the same entity type share a unified style, and lines between nodes represent relationships. Compared to traditional relational databases, the traffic accident knowledge graph offers greater flexibility, allowing for dynamic additions of new nodes, relationships, and attributes.

4.5. Application of Traffic Accident Knowledge Graph

4.5.1. Information Query in Knowledge Graph

Targeted data queries can be conducted within the traffic accident knowledge graph using Cypher. Depending on the requirements, diverse queries can retrieve various stored data from the graph. For instance, Figure 8 displays the visualization of certain environmental factors influencing accidents. The graph returns all environmental factors stored within the graph, such as road ice, heavy fog, and intersections without signal control lights. This information can be used for analyzing the impact of different environmental factors on traffic accidents. This also indicates that the original accident records currently involve limited descriptions of environmental impact factors, and the dataset is needed for supplementation in the future. In addition to direct information display, the graph can be used for clustering queries. For example, it can generate statistics on the number of minor, general, major, and catastrophic accidents, as shown in Table 7.

FIGURE 7 Part of traffic accident knowledge graph.

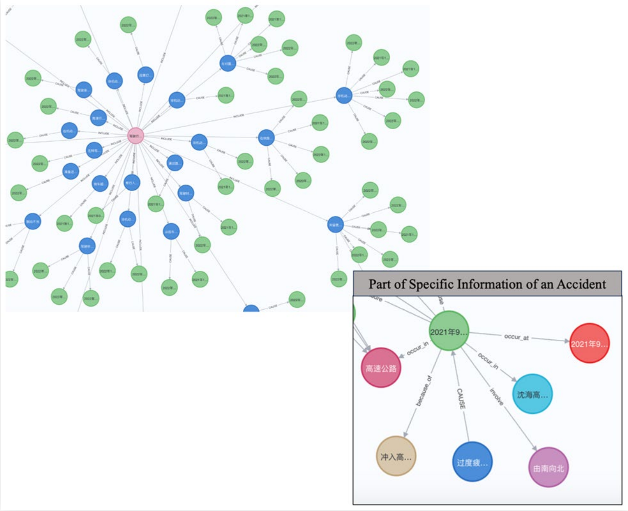
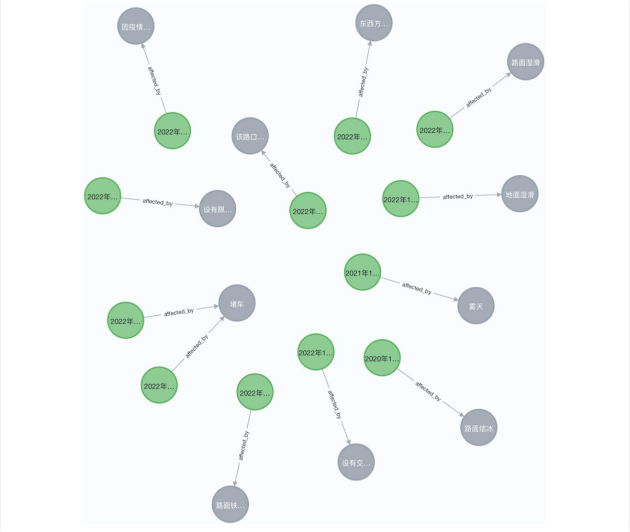


FIGURE 8 Example of visualization of the result of information query.



Overall, most accidents are classified as minor collisions and major collisions, together accounting for nearly 95% of the total, with major accidents alone making up 42%. This indicates that accidents with significant consequences are frequent. General accidents and particularly severe accidents have much lower proportions, together accounting for only 5% of the total. This distribution highlights the need to focus on major and catastrophic accidents when developing preventive measures and policies. In the following sections, we will further analyze major and catastrophic accident in detail.

4.5.2. Knowledge Graph-Based Retrieval-Augmented Generation Question-Answering System for Traffic Accidents

By integrating ChatGLM4 with traffic accident knowledge graph, we develop a Retrieval-Augmented generation question-answering system. This system leverages traffic accident data stored in the knowledge graph to handle more complex analytical tasks directly.

We combine the traditional retrieval approach of RAG with Cypher query-based retrieval for knowledge graphs. First, we use m3e-base model to generate a vector database based on the constructed knowledge graph. When a user inputs a traffic accident-related query through the Gradio interface, ChatGLM4 extracts key

TABLE 7 Statistics number of minor, moderate, major, and catastrophic accidents,

Traffic Accidents of Different Severity Levels	Number of Accidents
Minor	628
General	49
Major	500
Catastrophic	9

entities from the query that correspond to entities stored in the graph. By using SentenceTransformer model for embedding task, the extracted entities are vectorized and entity vectors are computed. Then, we use FAISS model for vector retrieval to match these vectors with entity nodes in Neo4j to identify the most similar entities. Subsequently, construct the corresponding Cypher statement to query the subgraph structure that includes these similar entities in our knowledge graph.

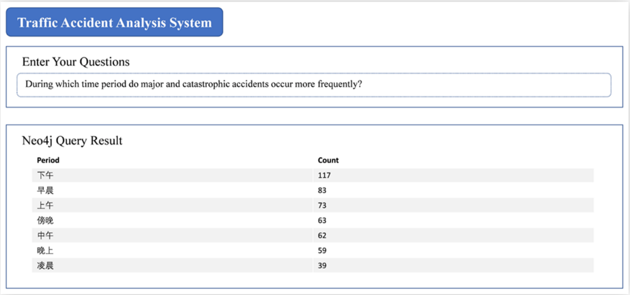
Additionally, we also try to use ChatGLM4 directly generates dynamic Cypher queries based on the graph structure and question intent. These queries are executed in Neo4j to retrieve knowledge. The results from two retrieval methods are then integrated into a unified context. If the subgraph query yields no results, only the results from the Cypher query generated by ChatGLM4 are used. If both retrieval methods return results, they are merged and used as background knowledge for ChatGLM4. Finally, ChatGLM4 generates answers to the user's question based on retrieval background knowledge. The Q-A system presenting both the knowledge graph query results and the natural language answer on the Gradio interface for users. Below are examples of the application in 1) Analyzing causes for major and catastrophic accidents; and 2) Generating scenarios for autonomous driving testing based on the traffic accident knowledge graph.

Causal Analysis of Major and Catastrophic Accidents: Major and catastrophic accidents result in fatalities, making the analysis of them critically important. By examining these incidents, we can identify behavioral, environmental, and other factors that exacerbate accident severity. This enables the development of more effective countermeasures. Common factors affecting the severity of accidents include the type of collision, time of occurrence, and the types of vehicles involved.

This study focuses on analyzing how collision type and time of occurrence impact the severity of major and catastrophic accidents. For instance, questions such as "Which type of accident is more likely to lead to major and catastrophic accidents?" and "During which time periods do major and catastrophic accidents occur more frequently?" are investigated. As illustrated in Figure 9, a retrieval-augmented question-answering system based on a traffic accident knowledge graph uses retrieved data as a reference to generate outputs for these specific queries.

The statistical results retrieved from traffic accident knowledge graph regarding the relationship between time periods and accident severity are presented in Table 8. Based on these data, ChatGLM4 generated a response with analysis, according to the query results, major and catastrophic accidents are more likely to occur in the afternoon, with a total of 117 incidents, the highest among all time periods. It is followed by the morning period, with 83 incidents. Overall, accident frequency is higher in the afternoon and morning periods, while the fewest accidents occur in the early morning, with only 39 incidents. This suggests that during certain times of the day, factors such as fatigue, lack of attention, or other influences may increase the likelihood of accidents.

FIGURE 9 The retrieval-augmented generation question-answering system page display (due to the origin dataset being in Chinese, the Q&A system performs better in Chinese, this original conversation is in Chinese).



Traffic Accident Analysis System

Enter Your Questions

During which time period do major and catastrophic accidents occur more frequently?

Neo4j Query Result

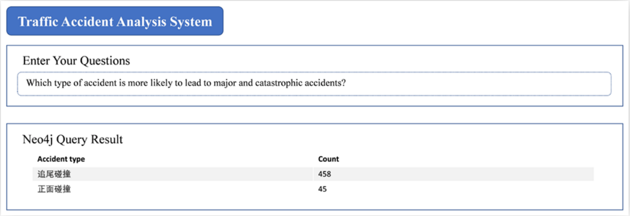
Period	Count
下午	117
早晨	83
上午	73
傍晚	63
中午	62
晚上	59
凌晨	39

TABLE 8 Time periods and the corresponding number of Major and Catastrophic Accidents.

Time Period	Number of Accidents
afternoon	117
morning	83
late morning	73
evening	63
Noon	62
night	39
early morning	39

As depicted in Figure 10, during the analysis of traffic accident types and their severity, ChatGLM4 provided a response incorporated knowledge regarding the potential associations between accident severity and types as represented in the knowledge graph. The query results from the knowledge graph indicated that there were 458 rear-end collisions and 45 head-on collisions. Based on the provided search results, ChatGLM4 generated a response appears that rear-end collisions are more likely to lead to major and catastrophic accidents than front-end collisions. The count for rear-end collisions is significantly higher at 458, compared to only 45 for front-end collisions. This statistical trend may suggest that rear-end collisions are more prevalent and may have a higher risk of causing severe outcomes. This response aligns with the observations of Sun et al. [45] and Shao et al. [46, 47].

FIGURE 10 The potential associations between accident severity and types as represented in the knowledge graph (the original conversation is in Chinese).



Traffic Accident Analysis System

Enter Your Questions

Which type of accident is more likely to lead to major and catastrophic accidents?

Neo4j Query Result

Accident type	Count
追尾碰撞	458
正面碰撞	45

The example above demonstrates that when integrated into a retrieval-augmented question-answering system based on a traffic accident knowledge graph, ChatGLM4 can generate accurate and insightful responses that meet the analytical needs of traffic accident analysis by leveraging the knowledge stored within the graph.

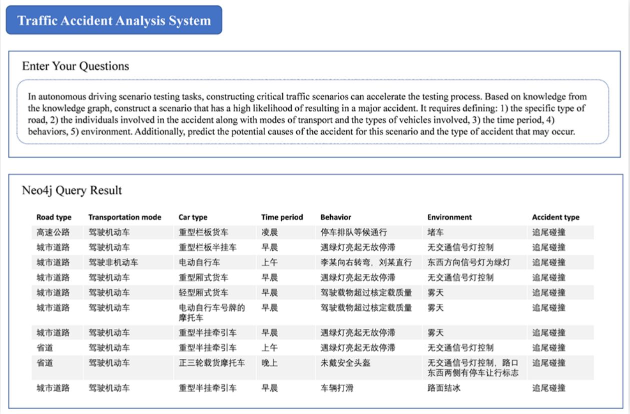
Autonomous Driving Test Scenario Generation: Autonomous driving test scenarios refer to virtual or physical testing environments that simulate real driving conditions to validate and optimize the performance of autonomous driving systems. These test scenarios can encompass a wide range of situations, from simple straight-line driving to complex urban streets, in order to evaluate the system's ability to handle various road conditions. The generation of key scenarios in autonomous driving tests helps quickly identify and focus on those situations that have a significant impact on the safety and stability of autonomous driving. These test scenarios, by simulating extreme conditions that may be encountered during real-world driving, enable an assessment of the system's response capabilities in critical environment. It contributes to improved testing efficiency, targeted evaluation, and enhances the robustness and safety of the system.

The autonomous driving test scenarios generated from real traffic accident data possess authenticity and representativeness. Test scenarios created using real-world data closely reflect actual conditions, ensuring autonomous driving systems to achieve higher reliability when responding to real road environments. Additionally, generating test scenarios based on accident data allows a focus on high-risk situations, accelerating the autonomous driving testing process and helping these systems to enhance safety, robustness, and their ability to handle complex situations more quickly and effectively.

By inputting a question in the form like "In autonomous driving scenario testing tasks, constructing critical traffic scenarios can accelerate the testing process. Based on knowledge from the knowledge graph, construct a scenario that has a high likelihood of resulting in a severe accident. It requires defining: 1) the specific type of road, 2) the individuals involved in the accident along with their modes of transport and the types of vehicles involved, 3) the time period, 4) behaviors, 5) environment. Additionally, predict the potential causes of the accident for this scenario and the type of accident that may occur". A detailed specification of the scenario content to be generated need to be defined.

As depicted in Figure 11, prior to answering the question, ChatGLM4 first learned lots of real traffic accident information within the traffic accident knowledge graph. Utilizing its robust transfer learning capabilities, it generated a novel testing scenario definition, outlining a scenario with a high probability of severe accidents in the autonomous driving scenario testing task: 1) type of road: urban road; 2) individuals involved and their modes of transport and vehicle types: driver operating a motor vehicle; 3) vehicle type: heavy box truck; 4) time period: morning; 5) behavior: the vehicle stops unexpectedly at a green light; 6) environment: controlled by traffic signals; 7) predicted accident type: rear-end collision; 8) cause of accident: behavior interfering with safe driving, such as stopping without reason when the green light is on. Moreover, ChatGLM4 leverages its analytical skills refined from the knowledge within the graph to offer an analysis of the probable time period for the scenario and to suggest appropriate countermeasures based on historical dialogue content. The response highlighted that, according to the query results, this scenario is common on urban roads in the morning and often involves heavy box trucks. The primary accident type is rear-end collision, triggered by drivers' traffic rule infractions through sudden stops at green lights. This trend suggests that during testing, autonomous driving systems should be especially vigilant of heavy truck behavior on urban roads, particularly under traffic signal regulation.

FIGURE 11 The relevant knowledge retrieved from KG learned by ChatGLM4 before generating answers.



Enter Your Questions

In autonomous driving scenario testing tasks, constructing critical traffic scenarios can accelerate the testing process. Based on knowledge from the knowledge graph, construct a scenario that has a high likelihood of resulting in a major accident. It requires defining: 1) the specific type of road, 2) the individuals involved in the accident along with modes of transport and the types of vehicles involved, 3) the time period, 4) behaviors, 5) environment. Additionally, predict the potential causes of the accident for this scenario and the type of accident that may occur.

Neo4j Query Result

Road type	Transportation mode	Car type	Time period	Behavior	Environment	Accident type
高速公路	驾驶机动车	重型栏板货车	凌晨	停车排队等候通行	堵车	追尾碰撞
城市道路	驾驶机动车	重型栏板半挂车	早晨	遇绿灯亮起无故停滞	无交通信号灯控制	追尾碰撞
城市道路	驾驶非机动车	电动自行车	上午	李某向右转弯，刘某直行	东西方向信号灯为绿灯	追尾碰撞
城市道路	驾驶机动车	重型厢式货车	早晨	遇绿灯亮起无故停滞	无交通信号灯控制	追尾碰撞
城市道路	驾驶机动车	轻型厢式货车	早晨	驾驶载物超过核定载质量	雾天	追尾碰撞
城市道路	驾驶机动车	电动自行车号牌的摩托车	早晨	驾驶载物超过核定载质量	雾天	追尾碰撞
城市道路	驾驶机动车	重型半挂牵引车	早晨	遇绿灯亮起无故停滞	雾天	追尾碰撞
省道	驾驶机动车	重型半挂牵引车	上午	遇绿灯亮起无故停滞	无交通信号灯控制	追尾碰撞
省道	驾驶机动车	正三轮载货摩托车	晚上	未戴安全头盔	无交通信号灯控制，路口东西两侧有停车让行标志	追尾碰撞
城市道路	驾驶机动车	重型半挂牵引车	早晨	车辆打滑	路面结冰	追尾碰撞

5. Discussion

KGs store, query, and reason over complex semi-structured and unstructured data, leading to a growing number of applications across diverse fields. While using traditional NLP methods for construction of KGs often rely heavily on manually annotated datasets and rule-based approaches, LLMs offer a more scalable and adaptable solution. Once deployed, traditional NLP systems tend to remain static unless updated with new data and retrained. This rigidity limits their adaptability to emerging trends or newly acquired knowledge. Unlike NLP models that require extensive fine-tuning and domain-specific annotations, LLMs like ChatGLM4 can generalize across different domains with minimal adaptation, making them possess an inherent capability to continuously learn to adapt and become highly versatile tools for construction KGs for different domain. Through techniques like prompt engineering and few-shot learning, LLMs can quickly incorporate new information into their existing knowledge base and become highly responsive to changing environments. With the continuous advancements in the language processing capabilities, LLMs have shown accuracy in knowledge extraction tasks as well as manually trained NLP models. In specific application contexts, the retrieval and text generation capabilities of LLMs further enhance the effectiveness of utilizing the knowledge and complex relational information stored in KGs. This paper explores the integration of LLMs with KGs, leveraging the robust language processing capabilities of LLMs to streamline KGs construction. Focusing on the domain of traffic accident analysis, we demonstrate a feasible and efficient approach for processing and applying large-scale unstructured traffic accident text records within a short timeframe. By using Retrieval-Augmented Generation (RAG) techniques, we combine the strengths of LLMs' language understanding with the structured querying power of KGs, enabling more sophisticated and nuanced analyses of unstructured data. By using the framework of RAG, KGs serve as a dynamic database accessible to LLMs and provide background knowledge to enhance the factual accuracy of LLM-generated answers in traffic accident analysis. It leverages relationships and context within the graph to provide deeper analysis and insights. By employing LLMs to interpret textual queries and perform reasoning on KGs, a connection between textual and structured information is established. This framework of construction and application a traffic accident knowledge graph showcases strong scalability, which is applicable to various fields and scenarios. We present some application examples in some complex query tasks, including accident data clustering analysis, causation analysis of accident severity and scenario generation for accidents.

This study has certain limitations. Although the data included in the knowledge graph is relatively extensive, it is region-specific and primarily reflecting conditions in Jiangsu Province, China. This limited geographical scope may introduce potential biases, as the dataset may not be fully representative of traffic accident scenarios in other regions or countries. The current data collection also has limitations in terms of quantity and informational diversity, which could affect the model's generalizability to a broader range of traffic accident scenarios. Other than that, due to the length constraints of the article, this paper mainly focuses on the application patterns of using LLMs to simplify the construction of KGs and to better use the knowledge stored in KGs, thereby providing more constructive answers in professional domain Q&A. We mainly pay attention to the scalability of this framework and choose ChatGLM4 for this paper's work considering its strong capabilities demonstrated in tasks in Chinese. In the process of experimentation, we have found that after multiple rounds of instruction-based dialogue adjustments, ChatGLM4 has shown stable excellent performance in knowledge extraction tasks. Due to the length constraints of the article, we have noticed that this paper has some limitations in model selection and further work for improving the model performance, and we plan to optimize these limitations in our subsequent work.

In the future, more real accident text records are planned to be collected from various regions and with diverse accident types to update and expand the traffic accident knowledge graph regularly. Additionally, in the future, LLMs could be further tested and adjusted for different types of traffic accident scenarios to improve the accuracy in knowledge extraction and Q&A tasks. Moreover, since the process of answer generation by LLMs remains a black box, it may result in unpredictability in the outcomes. Therefore, multiple rounds of self-validation and other advanced training techniques are still essential for research. It is also worth considering comparing with other advanced models and conducting ablation studies to optimize the effectiveness of model applications. Furthermore, in-depth mechanism analysis could be conducted to explore the impact of different system components on the performance of LLMs. Besides, based on this foundation, attempts can be made to realize more application scenarios of the traffic accident knowledge graph, such as providing users with recommendations for routes with low accident rates and travel safety tips to achieve traffic accident early warning; or using it to identify high-incidence areas of traffic accidents, providing traffic planning and management suggestions for government departments.

6. Conclusion

This study proposes an efficient framework for processing and analyzing unstructured traffic accident data by combining LLMs and KGs. By integrating the powerful language processing capabilities of LLMs with the structured data storage advantages of KGs, we achieve semi-automated knowledge extraction and detailed analysis. The main contributions of this study are as follows: 1) use LLMs to replace traditional NLP methods for knowledge extraction, enhancing processing speed and accuracy; 2) a prompt-based human-machine collaboration is implemented for self-verification and manual checks, ensuring the reliability of data extraction; 3)RAG technology is utilized to improve the practicality of KGs in traffic accident analysis, resulting in the development of a Knowledge Graph-Based Retrieval-Enhanced Question Answering System for Traffic Accidents.

This system supports complex query tasks, such as causation analysis and scenario generation, broadening the application scope of KG. In the field of traffic accident analysis, this system can be applied to: 1) intelligent querying and analysis of traffic accident data, providing support for accident investigation and risk assessment; 2) accident prediction and causation analysis based on historical data and KG-stored knowledge, 3) use realistic test cases generated to evaluate the performance of autonomous vehicles in complex scenarios; 4) perform deeper analysis based on graph relationships and offer more comprehensive answers.This work of this paper provides a solid foundation and innovative approach for the intelligent development of traffic accident management and decision-making.

Acknowledgements

This study is sponsored by the National Key R&D Program of China (2022YFB4300300), the Fundamental Research

Funds for the Central Universities (2242024RCB0047, 2242023R40057), the National Natural Science Foundation of China (52202410), and the Nanjing Science and Technology Program (202309010).

Contact Information

Yingqi Hou
Southeast University, Nanjing, China
houyingqi@seu.edu.cn

Yichang Shao
Southeast University, Nanjing, China
yshao@seu.edu.cn

Zhongyi Han
Shandong Provincial Communications Planning and Design Institute Group Co. Ltd., Jinan, China
hanzhongyi@aa.seu.edu.cn

Zhirui Ye (corresponding author)
Southeast University, Nanjing, China
yezhirui@seu.edu.cn

References

1. European Commission, ‘20,400 Lives Lost in EU Road Crashes Last Year,” 2024-11-11, 2024, https://transport.ec.europa.eu/news-events/news/20400-lives-lost-eu-road-crashes-last-year-2024-10-10_en.
2. Administration, N.H.T.S, “NHTSA Early Estimates: 2022 Traffic Crash Deaths | NHTSA,” 2024-11-10, 2024, https://www.nhtsa.gov/press-releases/traffic-crash-death-estimates-2022.
3. National Bureau of Statistics of China., "China statistical yearbook 2024," 2024-11-10, 2024 https://www.stats.gov.cn/sj/ndsj/2024/indexeh.htm.
4. Shao, Y., Shi, X., Zhang, Y., Xu, Y. et al., “Adaptive Forward Collision Warning System for Hazmat Truck Drivers: Considering Differential Driving Behavior and Risk Levels,” *Accid. Anal. Prev.* 191 (2023): 107221, doi:10.1016/j.aap.2023.107221.
5. Zhang, Y., Jing, L., Sun, C., Fang, J. et al., “Human Factors Related to Major Road Traffic Accidents in China,” *Traffic Inj. Prev.* 20, no. 8 (2019): 796-800, doi:10.1080/15389588.2019.1670817.
6. Bucsuházy, K., Matuchová, E., Zuuvala, R., Moravcová, P. et al., “Human Factors Contributing to the Road Traffic Accident Occurrence,” *Transp. Res. Procedia* 45 (2020): 555-561, doi:https://doi.org/10.1016/j.trpro.2020.03.057.
7. Chand, A., Jayesh, S., and Bhasi, A.B., “Road Traffic Accidents: An Overview of Data Sources, Analysis Techniques and Contributing Factors,” *Mater. Today Proc.* 47 (2021): 5135-5141, doi:https://doi.org/10.1016/j.matpr.2021.05.415.
8. Wang, K., Zhang, W., Jin, L., Feng, Z. et al., “Diagnostic Analysis of Environmental Factors Affecting the Severity of Traffic Crashes: From the Perspective of Pedestrian–Vehicle and Vehicle–Vehicle Collisions,” *Traffic Inj. Prev.* 23, no. 1 (2022): 17-22, doi:10.1080/15389588.2021.1995602.
9. Nasri, M., Aghabayk, K., Esmaili, A., and Shiwakoti, N., “Using Ordered and Unordered Logistic Regressions to Investigate Risk Factors Associated with Pedestrian Crash Injury Severity in Victoria, Australia,” *J. Safety Res.* 81 (2022): 78-90, doi:https://doi.org/10.1016/j.jsr.2022.01.008.
10. Shaon, M.R.R., Qin, X., Afghari, A.P., Washington, S. et al., “Incorporating Behavioral Variables into Crash Count Prediction by Severity: A Multivariate Multiple Risk Source Approach,” *Accid. Anal. Prev.* 129 (2019): 277-288, doi:https://doi.org/10.1016/j.aap.2019.05.010.
11. Sun, J. and Sun, J., “A Dynamic Bayesian Network Model for Real-Time Crash Prediction Using Traffic Speed Conditions Data,” *Transp. Res. Part C Emerg. Technol.* 54 (2015): 176-186, doi:https://doi.org/10.1016/j.trc.2015.03.006.
12. Parsa, A.B., Taghipour, H., Derrible, S., and Mohammadian, A.K., “Real-Time Accident Detection: Coping with Imbalanced Data,” *Accid. Anal. Prev.* 129 (2019): 202-210, doi:https://doi.org/10.1016/j.aap.2019.05.014.
13. Shao, Y., Shi, X., Zhang, Y., Shiwakoti, N. et al., “Injury Severity Prediction and Exploration of Behavior-Cause Relationships in Automotive Crashes Using Natural Language Processing and Extreme Gradient Boosting,” *Eng. Appl. Artif. Intell.* 133 (2024): 108542, doi:10.1016/j.engappai.2024.108542.
14. Amit, S., “Introducing the Knowledge Graph: Things, Not Strings,” Off. Google Blog, 2012.
15. Xu, J., Kim, S., Song, M., Jeong, M. et al., “Building a PubMed Knowledge Graph,” *Sci. Data* 7, no. 1 (2020): 205, doi:https://doi.org/10.1038/s41597-020-0543-2.
16. Chen, X., Jia, S., and Xiang, Y., “A Review: Knowledge Reasoning over Knowledge Graph,” *Expert Syst. Appl.* 141 (2020): 112948, doi:https://doi.org/10.1016/j.eswa.2019.112948.
17. Yang, P., Wang, H., Huang, Y., Yang, S. et al., “LMKG: A Large-Scale and Multi-Source Medical Knowledge Graph for Intelligent Medicine Applications,” *Knowl.-Based Syst.* 284 (2024): 111323, doi:10.1016/j.knosys.2023.111323.
18. Elhammadi, S., Lakshmanan, L.V., Ng, R., Simpson, M., Huai, B., Wang, Z., and Wang, L., “A High Precision Pipeline for Financial Knowledge Graph Construction,” in *Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics*, 2020, 967-977, https://aclanthology.org/2020.coling-main.84. DOI:10.18653/v1/2020.coling-main.84.
19. Sovrano, F., Palmirani, M., and Vitali, F., “Legal Knowledge Extraction for Knowledge Graph Based Question-Answering,” *Leg. Knowl. Inf. Syst.* (2020): 143-153, doi:10.3233/FAIA200858.
20. Yang C., Qi G., “An Urban Traffic Knowledge Graph-Driven Spatial-Temporal Graph Convolutional Network for Traffic Flow Prediction,” in *Proceedings of the 11th International Joint Conference on Knowledge Graphs*. New York: Association for Computing Machinery, 2023: 110-

114, https://dl.acm.org/doi/10.1145/3579051.3579058. DOI:10.1145/3579051.3579058.

21. Zeng, J. and Tang, J., "Combining Knowledge Graph into Metro Passenger Flow Prediction: A Split-Attention Relational Graph Convolutional Network," *Expert Syst. Appl.* 213 (2023): 118790, doi:10.1016/j.eswa.2022.118790.
22. Chen, T., Zhang, Y., Qian, X., and Li, J., "A Knowledge Graph-Based Method for Epidemic Contact Tracing in Public Transportation," *Transp. Res. Part C Emerg. Technol.* 137 (2022): 103587, doi:10.1016/j.trc.2022.103587.
23. Hu, S., Weng, J., Liang, Q., Zhou, W. et al., "Individual Travel Knowledge Graph-Based Public Transport Commuter Identification: A Mixed Data Learning Approach," *J. Adv. Transp.* 2022, no. 1 (2022): 2012579, doi:10.1155/2022/2012579.
24. Zhao, Z., "Quantification of Carbon Emission Technologies Based on Knowledge Graph Bert-BiLSTM-Attention-CRF Model," in *2023 International Conference on Electronics and Devices, Computational Science (ICEDCS)*. 2024, https://ieeexplore.ieee.org/abstract/document/10361660. DOI:10.1109/ICEDCS60513.2023.00014.
25. Zhang Y., Wang Y., Gao S., Raubal M., "Context-Aware Knowledge Graph Framework for Traffic Speed Forecasting Using Graph Neural Network," arXiv, 2024(2024-07-25), 2024, http://arxiv.org/abs/2407.17703. DOI:10.48550/arXiv.2407.17703.
26. Yu D., Peng W., Chen Y., Yang Y. et al., "Road Traffic Accident Data Management and Application Analysis Based on Knowledge Graph Technology," in *Proceedings of the 2024 International Conference on Digital Society and Artificial Intelligence*. New York: Association for Computing Machinery, 2024, 297-302. https://doi.org/10.1145/3677892.3677940.
27. Zhang, L., Zhang, M., Tang, J., Ma, J. et al., "Analysis of Traffic Accident Based on Knowledge Graph," *J. Adv. Transp.* 2022, no. 1 (2022): 3915467, doi:10.1155/2022/3915467.
28. Zhu, C., Wang, Y., Wang, Q., Fang, J. et al., "Research on Traffic Accident Prediction Basedon KG-CWT-RGCNN-BiLSTM," *Eng. Lett.* 31, no. 4 (2023). https://www.engineeringletters.com/issues_v31/issue_4/EL_31_4_09.pdf.
29. Zhu, Y., Wang, X., Chen, J., Qiao, S. et al., "LLMs for Knowledge Graph Construction and Reasoning: Recent Capabilities and Future Opportunities," *World Wide Web* 27, no. 5 (2024): 58, doi:10.1007/s11280-024-01297-w.
30. Trajanoska M., Stojanov R., Trajanov D., "Enhancing Knowledge Graph Construction Using Large Language Models," arXiv, 2023(2023-05-08)[2024-11-11]. http://arxiv.org/abs/2305.04676. DOI:10.48550/arXiv.2305.04676.
31. Zhang B., Reklos I., Jain N., Peñuela A.M. et al., "Using Large Language Models for Knowledge Engineering (LLMKE): A Case Study on Wikidata," arXiv, 2023(2023-09-15)[2024-11-11]. http://arxiv.org/abs/2309.08491. DOI:10.48550/arXiv.2309.08491.
32. Hofer M., Frey J., Rahm E., "Towards Self-Configuring Knowledge Graph Construction Pipelines Using LLMS-A Case Study with RML," in *Fifth International Workshop on Knowledge Graph Construction@ ESWC2024*, 2024, https://kg-construct.github.io/workshop/2024/resources/paper8.pdf.
33. Pan, S., Luo, L., Wang, Y., Chen, C. et al., "Unifying Large Language Models and Knowledge Graphs: A Roadmap," *IEEE Trans. Knowl. Data Eng.* 36, no. 7 (2024): 3580-3599, doi:10.1109/TKDE.2024.3352100.
34. Lewis P., Perez E., Piktus A., Petroni F. et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks," Advances in Neural Information Processing Systems. Curran Associates, Inc., 2020: 9459-9474[2024-11-13]. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.
35. Soman, K., Rose, P.W., Morris, J.H., Akbas, R.E. et al., "Biomedical Knowledge Graph-Optimized Prompt Generation for Large Language Models," *Bioinformatics* 40, no. 9 (2024): btae560, doi:10.1093/bioinformatics/btae560.
36. Matsumoto, N., Moran, J., Choi, H., Hernandez, M.E. et al., "KRAGEN: A Knowledge Graph-Enhanced RAG Framework for Biomedical Problem Solving Using Large Language Models," *Bioinformatics* 40, no. 6 (2024): btae353, doi:10.1093/bioinformatics/btae353.
37. Yuan, D., Zhou, K., and Yang, C., "Architecture and Application of Traffic Safety Management Knowledge Graph Based on Neo4j," *Sustainability* 15, no. 12 (2023): 9786, doi:https://doi.org/10.3390/su15129786.
38. Nguyen, H.L., Vu, D.T., and Jung, J.J., "Knowledge Graph Fusion for Smart Systems: A Survey," *Inf. Fusion* 61 (2020): 56-70, doi:10.1016/j.inffus.2020.03.014.
39. Noy N.F., Mcxuinness D.L., "Ontology Development 101: A Guide to Creating Your First Ontology," Stanford Knowledge Systems Laboratory Technical Report KSL-01-05, 2001.
40. Strobelt, H., Webson, A., Sanh, V., Hoover, B. et al., "Interactive and Visual Prompt Engineering for Ad-Hoc Task Adaptation with Large Language Models," *IEEE Trans. Vis. Comput. Graph.* 29, no. 1 (2023): 1146-1156, doi:10.1109/TVCG.2022.3209479.
41. Wang Y., Shen S., Lim B.Y., "RePrompt: Automatic Prompt Editing to Refine AI-Generative Art Towards Precise Expressions," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Hamburg Germany: ACM, 2023: 1-29[2024-11-11]. https://dl.acm.org/doi/10.1145/3544548.3581402. DOI:10.1145/3544548.3581402.
42. Matsumoto, N., Moran, J., Choi, H. et al., "KRAGEN: A Knowledge Graph-Enhanced RAG Framework for Biomedical Problem Solving Using Large Language Models," *Bioinformatics* 40, no. 6 (2024): btae353, doi:10.1093/bioinformatics/btae353.
43. Li M., Miao S., Li P.. "Simple Is Effective: The Roles of Graphs and Large Language Models in Knowledge-Graph-Based Retrieval-Augmented Generation," arXiv, 2024(2024-10-28)[2024-11-09]. http://arxiv.org/abs/2410.20724.
44. GLM T, Zeng A., Xu B., Wang B. et al., "ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools," arXiv, 2024(2024-07-30)[2024-11-14].

http://arxiv.org/abs/2406.12793. DOI:10.48550/arXiv.2406.12793.

45. Sun, Z., Xing, Y., Wang, J., Gu, X. et al., “Exploring Injury Severity of Vulnerable Road User Involved Crashes Across Seasons: A Hybrid Method Integrating Random Parameter Logit Model and Bayesian Network,” *Saf. Sci.* 150 (2022): 105682, doi:10.1016/j.ssci.2022.105682.

46. Shao, Y., Zhang, Y., Zhang, Y., Shi, X. et al., “A Virtual Vehicle–Based Car-Following Model to Reproduce Hazmat Truck Drivers’ Differential Behaviors,” *Journal of Advanced Transportation* 2024, no. 1 (2024): 5041012, doi:10.1155/2024/5041012.

47. Shao, Y., Han, Z., Shi, X., Zhang, Y. et al., “Risk-Informed Longitudinal Control in Autonomous Vehicles: A Safety Potential Field Modeling Approach,” *Physica A: Statistical Mechanics and its Applications* 633 (2024): 129419, doi:10.1016/j.physa.2023.129419.

Positions and opinions advanced in this work are those of the author(s) and not necessarily those of SAE International. Responsibility for the content of the work lies solely with the author(s).

This paper is based upon a presentation at the *2024 International Conference on Smart Transportation Interdisciplinary Studies*.

ISSN 0148-7191