CRASHSAGE: A LARGE LANGUAGE MODEL-CENTERED FRAMEWORK FOR CONTEXTUAL AND INTERPRETABLE TRAFFIC CRASH ANALYSIS

Hao Zhen and Jidong J. Yang

Smart Mobility and Infrastructure Lab College of Engineering University of Georgia, Athens, GA, USA Hao.Zhen, Jidong.Yang@uga.edu

ABSTRACT

Road crashes claim over 1.3 million lives annually worldwide and incur global economic losses exceeding \$1.8 trillion. Such profound societal and financial impacts underscore the urgent need for road safety research that uncovers crash mechanisms and delivers actionable insights. Conventional statistical models and tree ensemble approaches typically rely on structured crash data, overlooking contextual nuances and struggling to capture complex relationships and underlying semantics. Moreover, these approaches tend to incur significant information loss, particularly in narrative elements related to multi-vehicle interactions, crash progression, and rare event characteristics. This study presents CrashSage, a novel Large Language Model (LLM)-centered framework designed to advance crash analysis and modeling through four key innovations. First, we introduce a tabular-to-text transformation strategy paired with relational data integration schema, enabling the conversion of raw, heterogeneous crash data into enriched, structured textual narratives that retain essential structural and relational context. Second, we apply context-aware data augmentation using a base LLM model to improve narrative coherence while preserving factual integrity. Third, we fine-tune the LLaMA3-8B model for crash severity inference, demonstrating superior performance over baseline approaches, including zero-shot, zero-shot with chain-of-thought prompting, and few-shot learning, with multiple models (GPT-40, GPT-40-mini, LLaMA3-70B). Finally, we employ a gradient-based explainability technique to elucidate model decisions at both the individual crash level and across broader risk factor dimensions. This interpretability mechanism enhances transparency and enables targeted road safety interventions by providing deeper insights into the most influential factors.

Keywords Road safety; traffic crashes, crash severity prediction; Large Language Models (LLMs); Explainable AI (XAI), tabular-to-text transformation, data augmentation, supervised fine-tuning, gradient-based explainability

1 Introduction

Traffic crashes remain a persistent global public health crisis, resulting in over 1.3 million fatalities annually and imposing economic costs estimated at more than \$1.8 trillion. In the United States alone, approximately 42,000 lives are lost each year, despite continuous advancements in vehicle safety features, roadway design improvements, and safety policy implementation [1]. This enduring challenge highlights the inherently complex and multifaceted nature of traffic crashes, which result from dynamic interactions among human factors, vehicle characteristics, environmental conditions, and infrastructure elements.

The complexity of traffic safety analysis lies in capturing the intricate relationships between these factors. Human factors (e.g., variations in physical and physiological status, attentiveness levels, risk-taking behaviors, and social responsibility) interact with dynamic vehicle movements, changing weather conditions, and diverse roadway conditions and characteristics. This complexity necessitates sophisticated analytical approaches that can effectively model these high-dimensional interactions to develop context-aware, targeted interventions and policies.

Over the decades, traffic safety research has evolved through several methodological paradigms, each offering distinct advantages and facing specific limitations. Traditional statistical and econometric models have long served as the cornerstone of crash severity analysis [2, 3, 4, 5, 6, 7]. These methods, including random parameters multinomial logit models [8, 9], ordered probability models [10, 11], latent class models [12, 13], and markov-switching models [14], offer explicit interpretability through parameter estimates that quantify relationships between crash outcomes and related factors.

However, these traditional models impose rigid functional forms and distributional assumptions, which constrain their ability to capture the nonlinearities, complex high-order interactions, and inherent heterogeneity present in crash data [15]. These limitations often lead to inconsistent findings across datasets and geographic regions, an issue compounded by the relatively small sample sizes that are common in transportation safety research. Furthermore, specifying appropriate model structures typically requires considerable domain expertise, introducing potential subjectivity in the selection of variables and the specification of interactions.

In response to these limitations, researchers have increasingly resorted to machine learning (ML) techniques, such as random forests and deep neural networks, which have demonstrated improved predictive performance in crash modeling [16]. For interpretability, tree-based ensemble models typically rely on post-hoc methods like SHapley Additive exPlanations (SHAP) [17], which attribute importance scores to input features. Lares et al. [18] introduces a Feature Group Tabular Transformer (FGTT) model that enhances interpretability by grouping semantically related features into tokens and leveraging transformer attention heatmaps to reveal key interactions between feature groups, thereby uncovering relationships behind different traffic crash types. While these approaches offer greater flexibility in modeling complex, nonlinear relationships, they still suffer from certain limitations. Firstly, they focus on feature-level attribution without revealing the underlying reasoning process. Secondly, they fall short in capturing contextual and causal relationships beyond statistical correlations. Lastly, they lack a native mechanism for processing unstructured data such as textual crash reports, requiring feature engineering to encode relevant information. This encoding process can lead to substantial information loss when qualitative factors (e.g., the sequential nature of crash events, contextual factors, and narrative elements) are reduced to categorical variables or omitted entirely.

Large Language Models (LLM) have emerged as a transformative paradigm in traffic safety analysis, addressing many of aforementioned limitations inherent in both traditional statistical methods and conventional machine learning techniques. The evolution of language models, from encoder-only architectures like BERT [19] and RoBERTa [20] to decoder-only variants including the GPT series [21] and LLaMA [22], has demonstrated remarkable generalization capabilities across diverse domains [23, 24, 25], suggesting their potential for transportation safety applications. The unique advantages of LLMs in crash analysis stem from two key strengths: (1) their ability to process and derive insights from unstructured textual narratives, which often contain rich contextual information that is lost in structured tabular formats; and (2) the extensive world knowledge embedded within their pretrained parameters, which may enable nuanced reasoning about complex circumstances described in textual narratives.

Table 1 provides a comparative overview of the different methodological approaches to crash severity analysis, highlighting their respective input data requirements, interpretability characteristics, and key limitations. By leveraging unstructured crash narratives, LLMs can preserve the the richness and completeness of crash events by utilizing extensive contextual information, maintaining sequential flow of actions as the event unfolds, and capturing multivehicle interactions that are often fragmented or lost in traditional tabular formats.

Furthermore, LLMs are capable of generating natural language explanations that mirror human reasoning, offering intuitive and comprehensive insights into the dynamics of crash events. Unlike abstract feature importance scores, these explanations are easily interpretable by transportation officials and safety practitioners, even without specialized technical expertise.

Previous research [26] suggests that LLM-based crash severity analysis aligns well with domain knowledge while providing competitive accuracy. Various prompting techniques, including zero-shot, few-shot, and chain-of-thought approaches, have shown promise in improving modeling accuracy, while expert validation confirms that LLM-generated explanations correspond with established traffic safety knowledge. Fan et al. [27] fine-tuned LLaMA2 [28] for traffic crashes to predict accident outcomes and demonstrated the overall better performance than machine learning baselines: Random forest, Decision Trees, Adaptive boosting (AdaBoost), Bayesian Network(BN), LogisticRegression (LR), and Categorical boosting (CatBoost). It shows the promising application for traffic safety area. However, the limitation exists in the interpretability with intuitive "what-if" analysis. Their explanation method involves making controlled modifications to input data and observing the changes in the model's predictions, which helps to identify the influence of different features on the output. Despite this intuitive method, their approach has two significant limitations that restrict its practical utility in transportation safety applications. First, while their what-if analysis reveals how output distributions shift under hypothetical conditions, it does not expose the internal reasoning process of the model, which is the advantage of LLMs; the transformation of raw data into narrative form and subsequent classification remains

largely a black box. Second, the method offers only post-hoc explanations based on output perturbations, rather than providing a detailed, explanation that could enhance transparency and trust.

This study introduces CrashSage, a novel framework specifically designed to address critical gaps in crash analysis through four principal contributions. First, we develop a comprehensive *tabular-to-text transformation* method along with relational data integration schema that converts raw, heterogeneous crash data from Washington State datasets into richly detailed textual narratives, thereby preserving crucial structural and relational information commonly lost in conventional tabular formats. Second, we implement *context-aware data augmentation* via LLaMA-8B, enhancing the coherence of these crash narratives while rigorously maintaining factual accuracy. Third, we perform *supervised fine-tuning* of a LLaMA3-8B model [29] tailored for crash severity inference, demonstrably surpassing baseline approaches such as zero-shot, zero-shot with chain-of-thought, and few-shot across multiple model configurations (GPT-4o [30], GPT-4o-mini [31], and LLaMA3-70B [29]). Finally, we integrate a *gradient-based explainability* strategy to illuminate model decisions at both the individual crash level and in broader risk factor co-occurrence analyses. This interpretability mechanism not only enhances transparency and trustworthiness in model outputs but also provides actionable insights for targeted interventions in road safety management through a deeper understanding of how diverse factors interact to influence crash outcomes.

The CrashSage framework represents a paradigm shift in traffic safety analysis by transforming sparse crash records into actionable narratives. This approach enables transportation agencies to move beyond retrospective statistical analysis toward proactive risk identification and mitigation. By leveraging the semantic understanding capabilities of LLMs while maintaining explainability, our framework bridges the gap between advancement in AI and practical deployment in safety-critical transportation applications.

Table 1: Comparison of methods used in crash severity analysis.

Method	Input Data	Interpretability	Limitations
Econometric and Statistical Methods	Structured data	Explicit, model-based	Assumes fixed functional forms, limited complexity, lacks context
Tree Ensemble Models	Structured data	SHAP-based feature importance	Lacks context, needs feature engineering
Large Language Models (LLMs)	Unstructured data with context	Natural language explanations	Computationally expensive, potential biases

2 Related Work

This section explores two pivotal domains that underpin our CrashSage framework. First, we examine the evolution of Large Language Models (LLMs) and their emerging applications in transportation safety, emphasizing both technical advancements and domain-specific implementations. Second, we investigate interpretability techniques for language models, with a particular focus on gradient-based attribution methods that offer meaningful explanations in safety-critical contexts. Together, these complementary areas of research establish the technical foundations of our framework and highlight the key gaps it addresses in advancing explainable LLM capabilities for traffic safety analysis.

2.1 LLMs and Their Applications in Transportation Safety

Since their inception, LLMs have rapidly progressed from niche research tools to versatile, general-purpose systems with wide-ranging applications across diverse domains. The development of LLMs is driven by transformer-based architectures. BERT [19] employed encoder-only design well-suited for tasks like classification and information extraction. This was followed by a shift toward decoder-only architectures, exemplified by models such as GPT [21], LLaMA-3 [22], Claude [32], and GPT-4 [33]. This evolution has been characterized by notable improvements in contextual reasoning, factual accuracy, and adaptability, positioning LLMs as versatile tools for complex technical applications. The decoder-only models are particularly effective at generating coherent narratives, modeling sequential events, and capturing long-range dependencies. These capabilities are especially valuable for analyzing traffic crash sequences, which could unfold as temporally linked chains of events. Additionally, decoder-based LLMs exhibit strong few-shot and zero-shot learning capabilities, enabling effective generalization to new crash scenarios even in the absence of domain-specific training examples [26]. Their ability to handle increasingly longer contexts enables the integration of comprehensive information about road conditions, weather, vehicle states, and driver behaviors within a unified analysis framework.

The transportation sector has begun to explore the potential of LLMs across a range of applications, though their adoption in safety-critical contexts remains limited. Recent studies have demonstrated promising use cases, such as traffic forecasting [34]. Expanding on these advancements, Jonnala et al. [35] investigated the role of LLMs in optimizing transit operations, focusing on enhanced route planning, reduced wait times, and personalized travel assistance. By leveraging GTFS data alongside advanced natural language processing techniques, the research demonstrates that, with

careful engineering and fine-tuning, LLMs can significantly improve resource allocation and passenger satisfaction. These findings highlight the potential of LLMs to support data-driven decision-making in urban transit systems.

In the specific domain of transportation safety, early applications have focused on analyzing accident reports and extracting structured information from unstructured narratives. Notably, Zhen et al. [26] demonstrated that LLMs can effectively classify crash severity from textual descriptions while providing explanations that align with domain knowledge. This approach marks a stark departure from traditional methods that rely exclusively on structured tabular data.

Despite these promising developments, existing applications of LLMs in transportation safety face several limitations. First, most approaches rely on general-purpose models without domain-specific fine-tuning, limiting their ability to accurately interpret specialized terminology and contextual nuances unique to transportation domain. Second, LLMs are often employed as isolated analytical tools rather than being embedded within integrating frameworks that span the entire decision-making pipeline, from data preparation and feature engineering to actionable insights and policy support. Most notably, current efforts rarely address the explainability requirements that are essential for safety-critical contexts. In such settings, understanding the rationale behind a model's output is important, particularly when decisions may directly impact public safety.

These limitations underscore the need for transportation-specific adaptations of LLMs that incorporate domain knowledge, systematic data transformation, and robust explainability mechanisms. Addressing these challenges is essential for advancing LLMs from experimental use cases to practical tools that support evidence-based, actionable decision-making in safety-critical transportation contexts.

2.2 Explanation in Language Models

Unlike traditional statistical models with explicit parameters that directly link input features to predictions, LLMs rely on distributed representations that pose significant challenges for interpretation [36]. This section reviews the evolution of explainability approaches for LLMs, with a particular focus on gradient-based methods that inform the design of our CrashSage framework.

Early explanation methods for language models primarily relied on attention visualization [37], which offered intuitive but often misleading insights into model behavior. While attention maps can highlight tokens a model emphasizes during processing, subsequent research has shown that attention weights alone do not provide a reliable causal account of model predictions [38]. This limitation is especially problematic in transportation safety contexts, where accurately identifying causal factors is essential for designing effective interventions and informing policy decisions.

To overcome these shortcomings, more recent approaches have employed influence functions [39] and integrated gradients [40] to attribute model predictions to specific input tokens. These methods provide more reliable explanations by quantifying the impact of input perturbations on model outputs. In safety-critical domains like crash analysis, these techniques can potentially identify which aspects of an incident description most significantly influence severity predictions, thereby aligning with traditional safety analysis objectives focused on uncovering key risk factors.

Our CrashSage framework builds upon the comprehensive explanation techniques introduced by Wu et al. [41], incorporating gradient-based attribution to assess the influence of input tokens on specific model outputs. This approach uses a Taylor approximation to quantify how the inclusion or exclusion of individual input tokens affect output probabilities. By normalizing these attribution scores, the method yields a robust measure of each input token's contribution to the model predictions. When applied to crash analysis, this technique can highlight key terms within crash descriptions (e.g., "high speed," "distracted," "intersection") that strongly influence severity inference. These insights provide transportation safety professionals with a clearer understanding of the factors driving model decisions, supporting more informed and interpretable use of LLMs in the transportation safety domain.

Adapting explanation methods to transportation safety requires domain-specific considerations that go beyond general text analysis. Crash narratives often contain specialized terminology, structured inter-dependencies, and causal sequences that must be interpreted through the lens of traffic safety principles. To address these nuances, our framework extends the methods proposed by Wu et al.'s [41] by adjusting key hyperparameters and fine-tuning the model to align with established safety analysis practices. These adaptations ensure that the resulting explanations accurately reflect model behavior while offering actionable insights for transportation safety professionals.

The interpretability approach in our framework operates across multiple levels of granularity. At the individual crash level, token-level attributions identify critical factors influencing severity predictions, such as vehicle types, environmental conditions, and driver characteristics. At a broader scale, analyzing attribution patterns across multiple incidents enables to reveal systemic safety issues, particularly through co-occurrence analysis of high-attribution factors. This multi-level strategy addresses key interpretability gaps in current LLM applications within transportation

safety. As demonstrated in Section 6, the visualization of co-occurring factors reveals complex inter-dependencies among environmental conditions, driver behaviors, vehicle attributes, and infrastructure features that contribute to crash outcomes. This approach moves beyond isolated factor analysis to uncover compound effects and interaction dynamics, providing transportation agencies a more holistic understanding of crash causation mechanisms, which is critical for developing targeted and effective safety interventions.

3 Data Sources and Processing

This study leverages Washington traffic crash datasets, sourced from the Washington State Department of Transportation (WSDOT), encompassing crash records from 2020 to 2022. It has four different table, including crash table, road segment table, vehicle/unit table, and person table. Transportation crash data is typically stored in structured relational databases with complex schemas that separate information across multiple tables. While this structure was designed for storage and querying, it presents challenges for natural language processing applications. To address this, we utilized relational database schema combined with tabular-to-text transformation technique, introduced in the subsequent section, to convert structured crash data into coherent narrative descriptions.

3.1 Relational Schema for Crash Data Integration

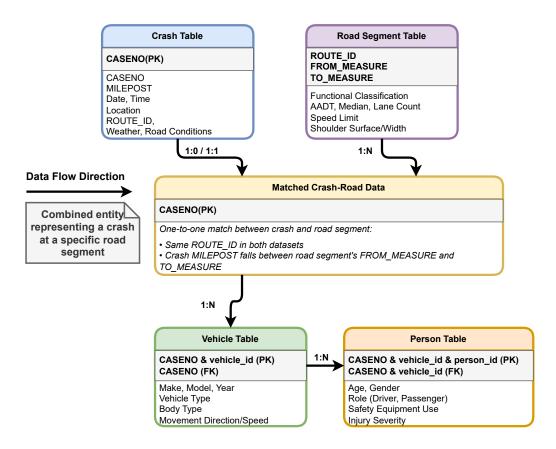


Figure 1: Data integration through relational schema

Our framework employs a relational schema, detailed in Figure 1) to integrate heterogeneous crash data sources through four normalized tables: Crash, Road Segment, Vehicle, and Person. The crash table forms the core entity with primary key CASENO, containing spatiotemporal attributes (location, timestamp) and environmental conditions. Through foreign key relationships, each crash record links to its corresponding Road Segment via spatial matching between MILEPOST and segment boundaries (FROM MEASURE, TO MEASURE), ensuring accurate geolocation mapping.

The schema maintains data integrity through hierarchical one-to-many relationships: each crash record connects to multiple vehicle entries through CASENO, and each vehicle/unit links to multiple person records. This structure preserves the natural hierarchy of crash events while enabling efficient querying of participant-level details. We

implemented this schema through a nested dictionary structure, with JSONL serialization facilitating language model integration through instruction-based learning templates.

Initial analysis revealed a severe disparity between 'No Apparent or Minor Injury' cases (n=49,648) and 'Serious injury or fatal' incidents (n=1,779). To address the class imbalance, we employed a stratified down-sampling approach, resulting in a more balanced dataset, comprising 2,654 no apparent or minor injury cases and 1,779 serious injury or fatal cases.

3.2 Tabular-to-Text Transformation

The tabular-to-text conversion pipeline transforms structured crash records into natural language narratives through a two-phase process: semantic normalization and template-based generation. In the normalization phase, numerical codes are mapped to natural language descriptors using domain-specific lexicons, translating categorical encoding into human-interpretable terms while removing duplicates and non-informative null values. The subsequent template-based generation uses fill-in-the-blank templates, illustrated below, to construct coherent, chronological event sequences that highlight key crash dynamics, enabling the generation of context-rich crash narratives suitable for analysis by LLMs.

On [date], a [day of week] at [time], an accident involving [number] vehicles occurred [lighting conditions], with [weather conditions]. The road condition at the time was [surface condition]. The accident took place on [road name] ([road type])...

Particularly, we structurally separate descriptive narratives (pre-crash conditions, collision mechanics) from outcome related narratives (injury severity, vehicle damage). This division support supervised learning objectives and enables models to learn potential causal relationships between antecedent conditions and resulting consequences.

Consequently, the transformation yields LLM-readable narratives that preserve relational semantics through natural language encoding. This methodology effectively bridges structured crash analytics with unstructured text processing capabilities, facilitating direct application of LLMs to transportation safety analysis while maintaining computational tractability.

4 CrashSage Framework

4.1 System Architecture Overview

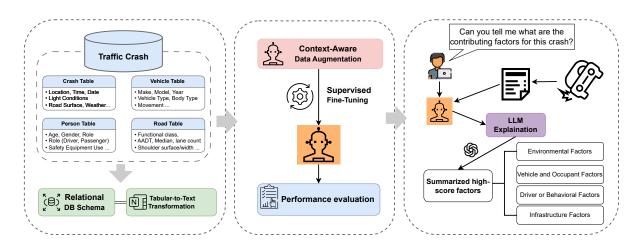


Figure 2: CrashSage framework.

Our CrashSage framework presented in Figure 2 offers a comprehensive approach to traffic safety analysis through the integration of LLMs with structured crash data. At its foundation, the framework utilizes entity-oriented analysis, leveraging the *Crash, Vehicle, Person, and Road/Unit* tables, each containing specific attributes that characterize different aspects of crash events. These structured data sources undergo a sequential transformation process beginning with a relational database schema that organizes the hierarchical relationships between entities (Section 3.1), followed by a Tabular-to-Text Transformation that converts structured records into coherent narratives (described previously in Section 3.2). The resulting template-based narratives are further refined through context-aware data augmentation

using a pretrained LLM agent (referred to as the base LLM in this paper), which improves textual coherence while preserving factual accuracy. The augmented narratives serve as the training data for the following supervised fine-tuning of our CrashSage agent, another LLM specialized in traffic safety domain. Additionally, the analytical capabilities of our framework are enhanced through gradient-based explainability methods, which offer interpretable insights via attribution analysis. This enables the generation of entity- or aspect-aware explanations, focusing on contributing factors grouped in to key categories: environmental conditions, vehicle and occupant characteristics, driver behavior, and infrastructure related elements, as shown in the right panel of Figure 2.

The core components of our CrashSage framework, including *Context-Aware Data Augmentation*, *Supervised Fine-Tuning*, *and Gradient-Based Explanation*, are detailed in the following sections.

4.2 Context-Aware Data Augmentation

Our framework implements a semantic preservation augmentation technique leveraging the LLaMA3-8B model [29] to enhance narrative coherence while maintaining factual fidelity in crash reports. The augmentation pipeline operates through a linguistic transformation process formalized as:

$$R = LLM(x, \{c_i\}_{i=1}^n), p_t), \tag{1}$$

where R represents the enhanced narrative, and $\{c_i\}$ defines a set of preservation constraints. The model receives specialized instructions p_t through system prompts that position it as a domain expert in transportation safety reporting. We employ a conservative temperature setting $(\tau = 0.1)$ to balance minimal creative variation with strong factual consistency.

In our implementation, we utilize the chat template to standardize input-output formats. The system prompt designates the model as a "professional editor specializing in rewriting traffic accident reports," with explicit preservation requirements communicated through detailed guidelines. These include maintaining all factual information (times, dates, locations, vehicle details), removing uninformative placeholders (e.g., "nan," "unknown"), preserving chronological order, and employing consistent professional language. The augmentation process handles batch processing of instances for computational efficiency.

This methodology delivers three significant improvements over conventional preprocessing approaches: standardization of linguistic variability across temporal and jurisdictional dimensions; enhancement of narrative fluidity without introducing factual distortion; and preservation of complex crash dynamics involving human, vehicular, and environmental factors. By maintaining semantic integrity while improving textual coherence, our technique creates crash narratives that are more consistent and amenable to downstream natural language processing tasks.

4.3 Supervised Fine-tuning of LLM

During the fine-tuning phase, the traffic crash severity inference task is framed as a next-token generation task. This process can be described as [42]:

$$p_{\theta}(T_i) = \prod_{j=1}^{|T_i|} p_{\theta}(t_j^{(i)}|t_1^{(i)}, \cdots, t_{j-1}^{(i)}),$$
(2)

where T_i is the *i*-th example in the training data, p_{θ} is the LLM model, $t_i^{(i)}$ denotes the *j*-th token in T_i .

The LLM's parameters are fine-tuned by maximizing the likelihood $p_{\theta}(T) = \prod_{i=1}^{N} p_{\theta}(T_i)$. Both the system prompt and the user prompt are masked for loss computation during training [43]. In our setting, the model is fine-tuned to predict crash severity levels, a task that demands a deep understanding of crash dynamics.

4.4 Gradient-based Explanation

4.4.1 Gradient-based Attribution

To ensure trustworthiness, we seek to understand the complex inner workings of our supervised fine-tuned LLM in the traffic safety domain by applying a gradient-based explanation method originally developed for analyzing the impact of instruction tuning on language models [41].

Gradient-based attribution techniques, which have been widely used to explain deep learning models [40], can help identify which words in an input text have the greatest influence on the model's output. In the context of traffic safety, this could enable pinpointing specific terms or phrases in incident reports that the model deems most indicative of safety issues. Following [41], we will use a first-order Taylor approximation of the difference in output probabilities when including or excluding each input token. Normalizing and thresholding these attribution scores yield a robust measure of each input token's importance. The importance $I_{n,m}$ of input token x_n to output token y_m is defined as:

$$I_{n,m} = p(y_m|Z_m) - p(y_m|Z_{m,/n})$$
(3)

 $I_{n,m} = p(y_m|Z_m) - p(y_m|Z_{m,/n})$ where Z_m is the context for generating y_m consisting of the concatenation of prompt X and the first m-1 tokens of response Y, and $Z_{m,n}$ omits token x_n from Z_m . This is approximated using a first-order Taylor expansion:

$$I_{n,m} \approx \left\langle \frac{\partial f(y_m | Z_m)}{\partial E_i[x_n]}, E_i[x_n] \right\rangle$$
 (4)

where $E_i[x_n]$ is the input word embedding of token x_n extracted from the fine-tuned LLM. The normalized pairwise importance score $\hat{S}_{n,m}$ is then defined as:

$$\hat{S}_{n,m} = \begin{cases} \left\lceil \frac{L \times I_{n,m}}{\max_{n'}^{N} I_{n',m}} \right\rceil & \text{if } \left\lceil \frac{L \times I_{n,m}}{\max_{n'}^{N} I_{n',m}} \right\rceil > b \\ 0 & \text{otherwise} \end{cases}$$
 (5)

where a scaling factor L and a binary threshold b are hyperparameters. In this study, we set L=100 and b=1.

4.4.2 Crash Severity Analysis with Word-level Attribution

The foundation of our analytical framework relies on word-level attribution, which assigns importance scores to individual words or phrases within crash narratives based on the gradient attribution, described previously in Section 4.4.1. The attribution scores quantify each token's contribution to the final prediction. Higher attribution scores suggest stronger associations with crash severity outcomes.

To enable aspect-aware explanation, our methodology decomposes crash narratives with word-level attribution into five key categories of contributing factors: 1) environmental conditions (weather, lighting, road surface quality, etc.); 2) vehicle and occupant characteristics (vehicle types, protection systems, etc.); 3) driver behavioral elements (speed, intoxication, maneuvers, etc.); 4) infrastructure features (road design, traffic control devices); and 5) unusual aspects with unexpectedly high attribution scores that may represent unique contributing factors. This multi-aspect approach enables comprehensive assessment of crash severity determinants.

The implementation consists of a semi-automated pipeline utilizing GPT-40 [30] to process and interpret word-level attribution data. Raw crash narratives, augmented with attribution scores, undergo systematic analysis through a carefully engineered prompt structure. This prompt directs the language model to process text with embedded attribution values, identify high-scoring words relevant to each factor category, and provide concise analytical summaries while maintaining output consistency.

To facilitate systematic analysis, we standardized the output using a JSON structure that preserves organizational consistency across all analyzed reports. For each factor category, the output includes a narrative summary capturing key insights and an array of high-scoring words with their associated attribution values. This structured approach enables qualitative assessment through the summaries as well as quantitative analysis via the attribution scores.

The core advantage of this approach lies in the combination of domain-specific prompting with robust natural language processing capabilities of LLMs, resulting in precisely formatted analyses that highlight each crash's salient risk factors in a transparent, interpretable manner. It advances crash severity analysis by leveraging the linguistic pattern recognition capabilities of LLMs while maintaining analytical rigor through attribution-based evidence.

CrashSage: Crash Severity Inference

In this study, we focus on fine-tuning a LLM for the task of crash severity inference, aligning a general-purpose LLM with domain-specific knowledge in road safety.

Supervised Fine-tuning and Hyperparameters

We performed supervised fine-tuning on the Llama3-8B model [29] using parameter-efficient fine-tuning via LoRA [44]. The model was using AdamW [45] as optimizer and trained for 30 epochs using the DeepSpeed framework

Table 2: Experimental Settings and Abbreviations

Approaches	
Zero-shot	Models evaluated without any examples
Zero-shot with Chain-of-Thought	Models prompted to explain reasoning step-by-step
Few-shot	Models provided with a small number of examples
Supervised Fine-Tuning	Model fine-tuned on task-specific data
Models Evaluated	
LLaMA3-8B [29]	8 billion parameter open-source model
LLaMA3-70B [29]	70 billion parameter open-source model
GPT-4o-mini [31]	gpt-4o-mini-2024-07-18, multimodal model from OpenAI
GPT-4o [30]	gpt-4o-2024-11-20, advanced multimodal model from OpenAI
Sampling Strategy	
Sampling Strategy	Greedy decoding for all models

[46]. We employed a LoRA configuration with rank (r) of 128, alpha scaling factor of 256, and dropout rate of 0.1, targeting all linear layers in the model. Training was conducted with a learning rate of 3e-5 using a cosine scheduler with 5% warmup, weight decay of 1e-4, and maximum gradient norm of 1.0. For optimization efficiency, we utilized gradient checkpointing and accumulated gradients over 16 steps with a per-device batch size of 1. The model processed sequences with a maximum length of 2,048 tokens and was trained in bfloat16 precision. This configuration balances computational efficiency with effective knowledge transfer while maintaining reasonable memory requirements. The experiments are conducted on a server with four Nvidia A6000 48GB GPUs.

5.2 Baseline Methods

This study establishes several baseline configurations to evaluate the performance of our supervised fine-tuned LLaMA3-8B [29] model against state-of-the-art proprietary models for the traffic crash severity classification task. We implemented three distinct prompting strategies with GPT-4o [30] as our primary baselines, while also extending our evaluation to include LLaMA3-8B, LLaMA3-70B, and GPT-4o mini [31] for broader comparative analysis.

The baselines are structured to assess how different prompting techniques influence model performance in classifying traffic crash severity into two severity levels: "No apparent or minor injury" or "Serious injury or fatal." These baselines provide crucial reference points for evaluating the efficacy of our supervised fine-tuning approach against powerful foundation models using various prompting strategies.

5.2.1 Zero-Shot Prompting

Our first baseline employs a zero-shot prompting strategy, wherein models receive a concise instruction without examples. The model is prompted with domain-specific context identifying it as a professional road safety engineer and tasked with classifying crash severity based solely on the provided crash description. The prompt used is shown below:

You are a professional road safety engineer.

You are given a detailed description for a traffic crash.

Please classify the severity of the crash into one of two categories: 'No apparent or minor injury',

'Serious injury or fatal accident'.

You can only output one of the classification result in your answer.

This approach evaluates the model's inherent ability to perform the classification task without prior examples, relying exclusively on its pre-trained knowledge on traffic safety.

5.2.2 Zero-Shot Chain-of-Thought

The second baseline implements a zero-shot chain-of-thought (CoT) approach, which extends the zero-shot prompting by explicitly instructing the model to analyze the traffic crash before outputting the classification result. The prompt used for this approach is:

You are a professional road safety engineer.

You are given a detailed description for a traffic crash.

Please analyze this traffic crash with careful reasoning first, and then classify the severity of the crash into one of the two categories: 'No apparent or minor injury', 'Serious injury or fatal accident'.

You can only output one of the classification result at the end of your answer.

This modification encourages the model to engage in a more deliberate reasoning process, potentially leading to improved decision-making. The CoT approach is designed to assess whether explicit instructions for analytical reasoning enhance classification accuracy compared to direct zero-shot prompting.

5.2.3 Few-Shot Learning

The third baseline utilizes a few-shot learning paradigm, presenting the model with two exemplar traffic crashes, one in each severity category, and their corresponding severity outcomes prior to requesting classification of the target case. The prompt includes carefully selected examples that serve as implicit demonstrations of the reasoning process and decision criteria:

You are a professional road safety engineer.

Here are two examples of traffic crashes and their severity classification:

[Example 1 with label 'No apparent or minor injury']

No apparent or minor injury

[Example 2 with label 'Serious injury or fatal']

Serious injury or fatal

You are given a detailed description for a traffic crash.

Please classify the severity of the crash into one of two categories: 'No apparent or minor injury', 'Serious injury or fatal'.

You can only output one of the classification result in your answer.

The few-shot approach leverages LLMs' in-context learning capabilities to learn from a minimal set of examples and apply that knowledge to new cases, a strategy demonstrated to improve performance across a range of natural language processing tasks.

5.3 Results

Table 3: Performance comparison of different models across evaluation metrics.

Setting	Model	Macro-F1	Accuracy	Macro-Recall	Macro-Precision
	LLaMA3-8B	0.6726	0.6883	0.67	0.68
ZS	LLaMA3-70B	0.6345	0.6355	0.67	0.67
ZS	GPT-4o-mini	0.6711	0.7078	0.67	0.72
	GPT-4o	0.7067	0.7229	0.70	0.72
	LLaMA3-8B	0.5071	0.5346	0.59	0.66
ZS_CoT	LLaMA3-70B	0.6059	0.6099	0.65	0.67
	GPT-40 mini	0.3717	0.4413	0.52	0.55
	GPT-40	0.3693	0.4443	0.52	0.58
	LLaMA3-8B	0.6851	0.6867	0.70	0.69
FS	LLaMA3-70B	0.7051	0.7184	0.70	0.71
	GPT-4o-mini	0.6560	0.6898	0.66	0.69
	GPT-4o	0.7062	0.7259	0.70	0.72
SFT	LLaMA3-8B	0.7361	0.7395	0.74	0.74

In evaluating the outcomes of our experimental comparisons, it is evident that the approach involving supervised fine-tuning (SFT) of LLaMA3-8B consistently provides superior performance across multiple metrics when compared to the zero-shot (ZS), zero-shot and chain-of-thought (ZS_CoT), and few-shot (FS) settings for unadapted models. As

summarized in Table 3, SFT demonstrates a Macro-F1 score of 0.7361, surpassing all baseline methods and highlighting the benefits of specialized domain adaptation. These findings derive in part from the motivations that guided our experiment design, namely the hypothesis that traffic crash narratives require careful domain grounding and that purely general-purpose prompting may not suffice in capturing nuanced, context-sensitive crash conditions.

A closer look at the ZS and FS settings indicates subtle differences. GPT-40 shows stronger performance in the zero-shot condition than the other baseline models, most notably in Macro-F1 (0.7067) and accuracy (0.7229). However, when LLaMA3-8B is given a minimal number of examples in the few-shot setting, its performance moves closer to that of GPT-40, demonstrating the model's capacity to incorporate contextual cues with only limited domain examples. The results in the ZS_CoT setting indicate that chain-of-thought prompting does not uniformly enhance outcomes; the performance declines for some models. This pattern suggests that while reflective reasoning can be beneficial in some contexts, introducing additional reasoning steps here may pose risks of over-interpreting or extraneous text generation that does not align with domain-consistent interpretations of crash data.

A point of particular interest is the comparable or even superior performance of fine-tuned LLaMA3-8B compared to a significantly larger model like LLaMA3-70B. The smaller model, when tuned on domain-specific data, not only reduces computational overhead but also captures key details of traffic crash environments, including interactions among vehicles, environmental features, and driver attributes. This highlights the potential of fine-tuned smaller models to perform competitively in specialized domains, offering an efficient alternative to large-scale models.

These results validate our initial hypothesis that transforming structured data into domain-focused narratives prior to model tuning enhances performance. The observed gain from SFT supports the notion that foundation models benefit from specialized retraining to internalize the nuanced complexity of traffic crashes. Moreover, the varying performance across prompting strategies underscores the importance of aligning LLM behavior with domain-specific reasoning paradigms. Through a systematic comparison of zero-shot, chain-of-thought, few-shot, and SFT approaches, our experiments demonstrate that, while general-purpose LLMs show strong baseline competence, supervised fine-tuning more effectively captures crash-specific scenarios. This finding charts a clear path for future work, where incorporating additional tuning data or exploring even more refined prompting techniques could further enhance both predictive accuracy and explainability in real-world traffic safety applications.

6 CrashSage: Interpretability and Attribution

In traffic safety modeling and analysis, interpretability is crucial for building trust and understanding the factors driving model predictions. Our CrashSage framework employs gradient-based attribution techniques to highlight the most influential terms in crash narratives that contribute to traffic crash severity inferences. This approach not only enhances transparency but also provides valuable insights into the complex interplay among factors affecting crash outcomes.

6.1 Individual Incident Inspection: Word-Level Explanations of an Accident

Our word-level attribution method aggregates token-level importance scores derived from gradient-based techniques to identify which words and phrases most significantly influence the model's severity classification. In detail, our implementation identifies token boundaries within the LLM's tokenization scheme and combines scores of sub-word tokens into coherent words, providing more intuitive interpretations for domain experts. These importance scores quantify how much each word contributes to the final prediction, enabling analysts to understand which aspects of a crash narrative were most decisive in the model's reasoning process.

6.1.1 Example 1: No apparent or minor injury crash

For illustration, consider a *no apparent or minor injury* crash example analyzed by our CrashSage. Figure 3 displays the crash narrative text where each word is associated with attribution scores. These scores represent the word-level attribution values derived from our model's analysis. Words receiving higher attribution scores are identified as having greater influence on the model's classification decision regarding crash severity. This word-level attribution approach provides a more interpretable view of which textual elements most significantly contributed to the severity assessment.

To better visualize the relative attribution of each word or phrase, we present the results using a color-coded heatmap. To minimize visual clutter, the visualization uses a binary color mapping scheme that clearly differentiate between high and low levels of attribution significance. Elements with higher attribution values are displayed in red, while those with lower but still notable attribution values appear in green. This divergent color scheme creates an intuitive visual hierarchy that emphasizes textual segments according to their influence on model predictions.

```
Example Narrative with Word-Level Attribution:
On[1.92] June[1.96] 29,[2.66] 2022,[4.28] at[1.68] 8:00[1.00] PM,[2.00] a[1.63] traffic[1.00] accident[1.68] occurred[1.39] on[1.00] Alternate[1.00] Route[1.00] 097ARi[1.68] in[1.00] Chelan,[3.39] Washington.[2.48]
[...] under dusk[2.28] conditions.[2.00] [...] rural[1.00] two-lane[2.00] road[1.00] [...]
[...] an [BRAND 1][2.00] [MODEL 1] vehicle[3.20] manufactured[1.46] in 2005,[3.68] [...] 35-year-old[2.68] male,[1.68]
[...] time[0.00] of[1.42] the[0.00] crash.[3.39] [...]
[...] 36-year-old[3.68] female,[0.00] [...][0.00] of[0.00] 47.9512[2.00] [...] hit-and-run[3.68] incident.[3.11]
```

Figure 3: Word-level attribution visualization for a *no apparent or minor injury* crash example. Terms with higher attribution scores (shown in brackets) have greater influence on the model's prediction. High-influence terms are highlighted in **bold** with color intensity representing attribution strength: gray (minimal).

The attribution visualization in Figure 4 reveals several key patterns in how the CrashSage evaluates crash narratives. Temporal markers ("On," "June," "29," "2022," "PM") receive moderate to high attribution scores, indicating the importance of time-related information in severity assessment. Location identifiers ("Chelan," "Washington") similarly show strong influence, suggesting geographical context impacts prediction.

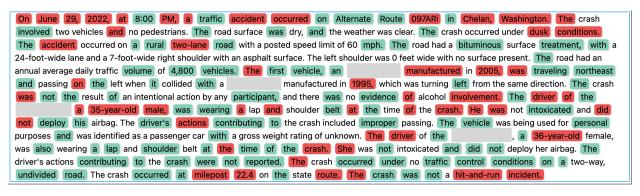


Figure 4: Exemplar visualization of attribution scores for a crash classified as "No apparent or minor injury." Colors indicate attribution strength: red indicates strongly influential words (high positive attribution), while green denotes moderately influential terms. This heat map visualization reveals which textual elements most significantly impact the model's severity classification decision. In this figure, sensitive vehicle information (e.g., make and model) has been redacted using gray boxes. This redaction was performed manually and does not alter or affect the substantive content or analytical integrity of the narrative.

Environmental conditions, particularly "dusk" and "two-lane" road configuration, demonstrate notable attribution weights, aligning with safety research that identifies limited visibility and road type as significant risk factors. Vehicle-specific details ("[BRAND 1] [MODEL 1] vehicle," "2005") receive high attribution scores, likely reflecting the model's attention to vehicle age and model. The most substantial attributions appear for driver characteristics ("35-year-old male," "36-year-old female") and incident type indicators, particularly the negated "hit-and-run incident" phrase, which appears critical to the model's classification decision.

We further employ a semi-automated pipeline with GPT-40, described in section 4.4.2,to summarize crash narratives with word-level attribution scores, wherein the system prompt for the LLM specifies a concise JSON-based output structure, organizing into five key categories: environmental, vehicle/occupant, behavioral, infrastructure, and unusual/standout factors. Raw crash text, augmented with word-level attribution scores, is processed by the GPT-40. The core advantage lies in pairing domain-specific prompting with robust file handling, leading to precisely formatted results that highlight each crash's salient risk factors by categories. A structured summary of the same example is presented below.

Environmental Factors

Summary: The crash occurred on June 29, 2022, at 8:00 PM under clear weather and dry road conditions. It was dusk at the time, and the location was a rural two-lane road in Chelan, Washington.

Key Factors:	
2022	(4.28)
Chelan	(3.39)
29	(2.66)
Washington	(2.48)
dusk	(2.28)
8:00 PM	(2.00)
June	(1.96)

Vehicle and Occupant Factors

Summary: The crash involved two vehicles: a 2005 [BRAND 1] [MODEL 1] driven by a 35-year-old male and a 1995 [BRAND 2] [MODEL 2] driven by a 36-year-old female. Both drivers were wearing seat belts, and no airbags were deployed. Neither driver was intoxicated.

Key Factors:	
2005	(3.68)
36-year-old	(3.68)
[BRAND 1] [MODEL 1] vehicle	(3.20)
1995	(3.00)
35-year-old	(2.68)
[BRAND 2] [MODEL 2] vehicle	(1.00)

Driver Behavioral Factors

Summary: The driver of the [BRAND 1] [MODEL 1] vehicle was improperly passing when the crash occurred. There was no evidence of alcohol involvement, and the crash was not intentional.

Key Factors:				
no evidence of alcohol				

involvement (2.00)improper passing (1.00)not intentional (1.00)

Infrastructure Factors

Summary: The crash occurred on Alternate Route 097ARi, a rural two-lane road with a posted speed limit of 60 mph. The road had a bituminous surface treatment, a 24-foot-wide lane, and a 7-foot-wide right shoulder. The left shoulder was 0 feet wide. The location had an AADT of 4,800 vehicles and no traffic control.

Key Factors:	
two-lane	(2.00)
Alternate Route 097ARi	(1.68)
AADT 4,800	(1.00)
no traffic control	(1.00)

Unusual/Standout Factors

Summary: The crash was not a hit-and-run incident, and the specific location at milepost 22.4.

Key Factors:	
hit-and-run	(3.68)
milepost 22.4	(2.00)

6.1.2 Example 2: Serious injury and fatal crash

```
Example Narrative with Word-Level Attribution (Serious/Fatal Crash):

On[0.00] September[1.22] 5,[2.68] 2020,[3.61] at[1.68] 1:00[2.62] PM,[3.02] a[1.00] rear-end[2.28] collision[1.92] occurred[1.00] on[1.63] State[0.00] Route[1.00] 542i[3.11] (MAINLINE)[4.37] in[0.00] Whatcom,[3.00] Washington.[2.00] [...] two-lane[2.00] road[1.68] [...] 2001[2.00] [BRAND 1] [0.00] [MODEL 1] vehicle,[3.05] [...] BRAND 2[0.00] MODEL 2[3.00] motorcycle,[2.00] [...]

[...] motorcycle's[2.68] rear-end[0.00] collision[1.00] [...] alcohol[0.00] involvement.[2.00] [...] [BRAND 1] [0.00] MODEL 2[4.31] [...] intoxication[2.68] [...] actions.[2.00] [...]
```

Figure 5: Word-level attribution visualization for a *serious injury or fatal* crash example. Terms with higher attribution scores (shown in brackets) have greater influence on the model's prediction. High-influence terms are highlighted in **bold** with attribution strength represented by numerical values. This visualization demonstrates how the model attends to different factors when assessing a severe crash compared to minor injury cases.

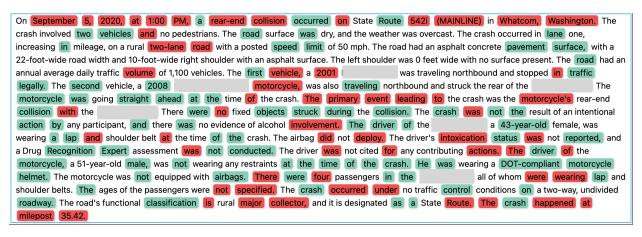


Figure 6: Example visualization of attribution scores for a crash classified as "Serious injury or fatal accident." Colors indicate attribution strength: red indicates strongly influential words (high positive attribution), while green denotes moderately influential terms. Notable high-attribution elements include specific temporal and location identifiers ("2020," "PM," "542i," "MAINLINE"), vehicle details ("[MODEL 1]," "motorcycle"), and precise geographic locations. In this figure, sensitive vehicle information (e.g., make and model) has been redacted using gray boxes. This redaction was performed manually and does not alter or affect the substantive content or analytical integrity of the narrative.

Similarly, the attribution visualization in Figure 6 reveals distinctive patterns for this serious/fatal crash. While temporal and location information remain important as in the minor injury case, this example shows particularly strong attribution to specific roadway identifiers ("542i," "MAINLINE") and precise geographic coordinates. The motorcycle involvement ("[MODEL 2]," "motorcycle's") receives substantial weight, consistent with traffic safety research identifying motorcycles as associated with higher crash severity outcomes. Notable also is the high attribution for "intoxication" even though it's negated in the text, suggesting the model considers this factor critically important when assessing crash severity.

The attribution pattern in this serious/fatal crash example reveals notable differences compared to the minor injury case previously analyzed. While both examples show significant attribution to temporal and location information, this severe crash narrative demonstrates particularly strong influence from specific roadway identifiers ("542i," "MAINLINE") and precise geographic "milepost 35.42".

The motorcycle involvement receives substantial attribution weight, with high scores for "[MODEL 1][3.00]" and "motorcycle's[2.68]," aligning with traffic safety research that consistently identifies motorcycles as associated with increased crash severity. Vehicle type information ("[BRAND 1] [MODEL 1][4.31]") shows remarkably high attribution, possibly indicating the model has learned relationships between vehicle types and crash outcomes from its training data.

Noteworthy is the high attribution score for "intoxication[2.68]" despite it being negated in the text, suggesting the model places significant importance on this factor's presence or absence when assessing crash severity. The phrase

"rear-end[2.28]" collision type also receives substantial attribution, reflecting its relevance to severity outcomes in motorcycle-involved crashes.

This gradient-based attribution analysis demonstrates how the CrashSage identifies distinctive risk patterns when evaluating crash severity. For serious/fatal crashes, the model focuses intensely on vehicle types (particularly motorcycles), specific road identifiers, and precise location data, while utilizing different attribution patterns for minor injury incidents. The visualization provides transparent insight into the model's decision-making process, revealing how it integrates multiple contextual elements: temporal, environmental, vehicular, and human factors, rather than focusing on isolated aspects. This comprehensive approach mirrors the multifaceted evaluation process used by human safety experts, while offering computational precision in identifying combinations of factors that contribute to different crash outcomes.

Similarly, the semi-automated pipeline is applied to summarize this serious injury/fatal crash narrative, with the resulting output shown below.

Environmental Factors

Summary: The crash occurred on a dry road surface under overcast weather conditions during the daytime (1:00 PM) on a rural two-lane road in Whatcom, Washington.

Key Factors:		
2020	(3.61)	
PM	(3.02)	
Whatcom	(3.00)	
5	(2.68)	
1:00	(2.62)	
Washington	(2.00)	
September	(1.22)	

Vehicle and Occupant Factors

Summary: The crash involved a 2001 [BRAND 1] [MODEL 1] and a 2008 [BRAND 2] [MODEL 2] motorcycle. The [BRAND 1] [MODEL 1] vehicle had a 43-year-old female driver and four passengers, all wearing seat belts. The motorcycle was driven by a 51-year-old male wearing a DOT-compliant helmet but no other restraints. The [BRAND 1] [MODEL 1]'s airbag did not deploy.

Key Factors:		
[BRAND 1]	(3.05)	
[BRAND 2]	(3.00)	
[MODEL 1]	(3.00)	
intoxication	(2.68)	
[MODEL 2]	(2.36)	
2001	(2.00)	
airbag	(1.65)	
2008	(1.00)	

Driver Behavioral Factors

Summary: The motorcycle rear-ended the [BRAND 1] [MODEL 1], which was legally stopped in traffic. There was no evidence of alcohol involvement or intentional actions by either driver. The motorcycle driver was not cited for any contributing actions.

Key Factors:	
motorcycle's	(2.68)
intoxication	(2.68)
rear-end	(2.28)
contributing actions	(2.00)
collision	(1.92)

Infrastructure Factors

Summary: The crash occurred on State Route 542i, a rural two-lane road with a posted speed limit of 50 mph. The road had an asphalt concrete surface and an annual average daily traffic volume of 1,100 vehicles.

Key Factors:	
MAINLINE	(4.37)
State Route 542i	(3.11)
State Route	(2.59)
two-lane	(2.00)
volume	(1.68)

Unusual/Standout Factors

Summary: The crash location's milepost were highly detailed and had high attribution scores.

Key Factors:	
35.42	(4.52)
milepost	(3.36)

6.2 Co-occurrence analysis of high-score factors from different aspects

To reveal relationships between influential crash factors, we conducted analysis, focusing on high-attribution elements identified through our gradient-based attribution approach. The process begins with factor extraction, constrained to the top five factors for each of the four aspects: environmental, driver behavioral, vehicle/occupant, and infrastructure. Subsequently, co-occurring factor pairs are identified and visualized using a Sankey diagram.

Semantic grouping was applied to consolidate conceptually similar terms that appeared with varied phrasing throughout the dataset. For instance, different temporal references were unified under broader descriptors such as "time of day". This semantic grouping process enhanced interpretability while preserving the underlying semantic relationships between factors across various crash instances.

The resulting data structure captures the co-occurrence patterns in reference to four fundamental safety aspects (i.e., environmental conditions, driver behaviors, vehicle attributes, and infrastructure features), with each node representing a distinct factor and connecting links quantifying co-occurrence frequency between factor pairs. This graphic visualization reveals how these specific aspects interact in complex crash scenarios. The structure illuminates critical interdependencies, such as correlations between alcohol-related behaviors and specific temporal or roadway characteristics, providing a holistic view of crash dynamics.

Figure 7 presents these relationships through a Sankey diagram where safety aspects are color-coded for clarity, and connection strengths are represented by flow widths proportional to co-occurrence frequency. The diagram reveals complex interdependencies among crash factors spanning environmental, behavioral, vehicle/occupant, and infrastructure aspects. Notably, driver behavioral factors, particularly intoxication status and alcohol-related impairment, emerge as central nodes with extensive links to both environmental and infrastructure elements, underscoring the critical role of driver condition in crash severity. Temporal factors, such as time of day, show strong associations with impairment-related behaviors, indicating diurnal patterns in high-risk driving. Additionally, geographic variability is reflected in distinct connection patterns between behavioral factors and specific locations, pointing to regional heterogeneity in crash factor distributions.

Several factor combinations stand out as particularly relevant to crash severity outcomes. The co-occurrence pathway between alcohol-related impairment and excessive speed suggests a synergistic relationship that likely amplifies crash energy and impact forces. Links between intoxication status and restraint utilization indicate a potential behavioral coupling that compounds injury risk through both increased crash likelihood and decreased protection. The visualization also reveals how certain environmental conditions interact with infrastructure characteristics, particularly evident in connections between wet weather and specific road surface attributes, which together create compounded effects on vehicle handling and control.

Vehicle and occupant factors primarily function as intermediate outcome nodes within the graph. Restraint systems and airbag status appear as downstream factors from driver behaviors, suggesting that safety equipment utilization is influenced by driver characteristics. The connection between vehicle year and restraint technologies indicates the progressive integration of advanced occupant protection systems. The position of restraint-related factors as major

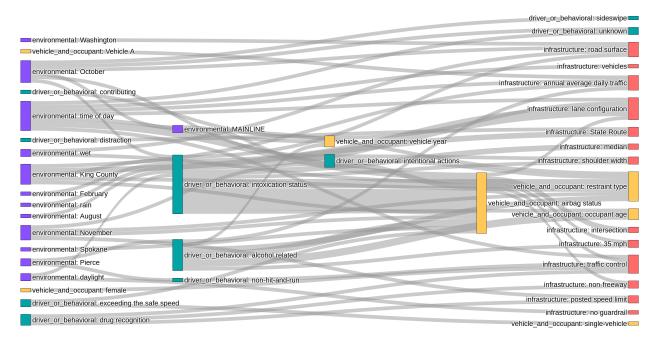


Figure 7: Aspect-level risk factor association visualized via a Sankey diagram. This visualization shows the relationships between identified high-attribution terms and their categorization into broader safety aspects. The thickness of each link indicates the relative frequencies of co-occurrence between factor categories.

nodes with multiple connections across the graph underscores their persistent influence across diverse crash scenarios, confirming their fundamental role in severity mitigation regardless of initiating factors.

Infrastructure elements exhibit multifaceted interactions with both antecedent and outcome variables. Lane configuration, with its numerous cross-category connections, exemplifies the complex role of road design in shaping crash dynamics. Road surface characteristics serve as bridging factors between environmental conditions and driver behaviors, suggesting a mediating influence on crash risk. Additionally, traffic control systems display strong association with driver decision-making, reflecting their impact on behavioral responses. Collectively, these interconnections highlight how the roadway environment influences and responds to driver behavior patterns, creating feedback loops that affect crash outcomes.

This diagram analysis demonstrates that crash severity outcome arise from intricate interactions among diverse factors across different aspects. The co-occurrence patterns suggest that effective traffic safety interventions require multifaceted approaches addressing both human factors and infrastructure design. Special attention is warranted for central behavioral nodes, which act as critical links between environmental conditions and crash outcomes.

Rather than viewing risk factors in isolation, the sankey diagram highlights their interrelationships and potential synergistic effects on crash severity outcomes. This integrated perspective offer road safety researchers and practitioners with an evidence-based foundation for developing multifaceted interventions that account for interconnected dynamics of crash development, thereby supporting more effective and systematic improvements in transportation systems.

7 Conclusion and Discussions

This paper introduces CrashSage, a novel LLM-based framework designed to address key challenges in traffic safety analysis. By transforming traditional crash records into coherent textual narratives, followed by context-aware data augmentation, our approach significantly mitigates information loss inherently associated with conventional tabular representations. Building upon enriched narratives, we fine-tune the LLaMA3-8B model to infer crash severity while generating interpretable insights grounded in domain-specific context.

Our experimental results show that the supervised fine-tuned LLaMA3-8B model outperforms comparative baselines, including zero-shot, zero-shot chain-of-thought, and few-shot prompting strategies. Moreover, the integration of gradient-based attribution methods enhances interpretability by uncovering the complex interplay among crash-related factors. This enables domain experts and decision-makers to identify the most influential elements (e.g., speed,

intoxication, infrastructure). This level of interpretability is especially valuable for informing targeted traffic safety interventions.

Despite its demonstrated strengths, we acknowledge several limitation of the proposed framework. As a purely text-based approach, it cannot capture other modalities, such as visual cues from roadside cameras or vehicle dash-cam footage, which can be crucial for understanding the time-critical driver action or vehicle status. Additionally, while our gradient-based explanations offer transparency, they rely on approximations of model behavior and may not fully capture the full depth of the model's internal representations or latent reasoning processes.

Looking ahead, we identify three promising directions for future work. First, integrating video and sensor data into the LLM pipeline could enhance crash narratives with real-time spatiotemporal context, improving model accuracy and robustness. Second, incorporating interpretability-driven constraints into the fine-tuning process may strengthen the consistency and reliability of explanations across diverse scenarios. Third, extending CrashSage from retrospective analyses to proactive risk estimation, where evolving traffic events trigger real-time alerts, offers significant operational benefits if deployed across large-scale transportation networks.

In summary, our work demonstrates the potential of LLMs to bridge structured data with natural language reasoning for advancing traffic safety research. By improving modeling accuracy, interpretability, and real-time applicability, CrashSage lays the groundwork for a more insightful, transparent, and actionable approach to crash analysis, empowering transportation agencies to make informed, data-driven decisions aimed at reducing road injuries and fatalities.

References

- [1] National Center for Statistics and Analysis. Early estimates of motor vehicle traffic fatalities and fatality rate by sub-categories in 2022. Crash•Stats Brief Statistical Summary DOT HS 813 448, National Highway Traffic Safety Administration, April 2023.
- [2] Shahrior Pervaz, Tanmoy Bhowmik, and Naveen Eluru. An econometric framework for integrating aggregate and disaggregate level crash analysis. *Analytic methods in accident research*, 39:100280, 2023.
- [3] Yichuan Peng, Mohamed Abdel-Aty, Qi Shi, and Rongjie Yu. Assessing the impact of reduced visibility on traffic crash risk using microscopic data and surrogate safety measures. *Transportation research part C: emerging technologies*, 74:295–305, 2017.
- [4] Thomas F Golob and Wilfred W Recker. Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. *Journal of transportation engineering*, 129(4):342–353, 2003.
- [5] Naveen Eluru and Chandra R Bhat. A joint econometric analysis of seat belt use and crash-related injury severity. *Accident Analysis & Prevention*, 39(5):1037–1049, 2007.
- [6] Peter T Savolainen, Fred L Mannering, Dominique Lord, and Mohammed A Quddus. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis & Prevention*, 43(5):1666–1676, 2011.
- [7] Hao Zhen, Oscar Lares, Jeffrey Cooper Fortson, Jidong J Yang, Wei Li, and Eric Conklin. Unraveling the dynamics of single-vehicle versus multi-vehicle crashes: a comparative analysis through binary classification. *Applied Computing and Intelligence*, 4(2):349–369, 2024.
- [8] Ali Behnood and Fred L Mannering. An empirical assessment of the effects of economic recessions on pedestrianinjury crashes using mixed and latent-class models. *Analytic methods in accident research*, 12:1–17, 2016.
- [9] Donald Mathew Cerwick, Konstantina Gkritza, Mohammad Saad Shaheed, and Zachary Hans. A comparison of the mixed logit and latent class methods for crash severity analysis. *Analytic Methods in Accident Research*, 3:11–27, 2014.
- [10] Naveen Eluru and Shamsunnahar Yasmin. A note on generalized ordered outcome models. *Analytic methods in accident research*, 8:1–6, 2015.
- [11] Shamsunnahar Yasmin, Naveen Eluru, and Abdul R Pinjari. Analyzing the continuum of fatal crashes: A generalized ordered approach. *Analytic methods in accident research*, 7:1–15, 2015.
- [12] Shamsunnahar Yasmin, Naveen Eluru, Chandra R Bhat, and Richard Tay. A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. *Analytic methods in accident research*, 1:23–38, 2014.
- [13] Naveen Eluru, Morteza Bagheri, Luis F Miranda-Moreno, and Liping Fu. A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings. *Accident Analysis & Prevention*, 47:119–127, 2012.

- [14] Yingge Xiong, Justin L Tobias, and Fred L Mannering. The analysis of vehicle crash injury-severity data: A markov switching approach with road-segment heterogeneity. *Transportation research part B: methodological*, 67:109–128, 2014.
- [15] Fred Mannering, Chandra R Bhat, Venky Shankar, and Mohamed Abdel-Aty. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic methods in accident research*, 25:100113, 2020.
- [16] Md Adilur Rahim and Hany M Hassan. A deep learning based traffic crash severity prediction framework. *Accident Analysis & Prevention*, 154:106090, 2021.
- [17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
- [18] Oscar Lares, Hao Zhen, and Jidong J. Yang. Feature group tabular transformer: a novel approach to traffic crash modeling and causality analysis. *Applied Computing and Intelligence*, 5(1):29–56, 2025.
- [19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.
- [20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.
- [21] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. OpenAI, 2018.
- [22] Pooja Dubey, Ananya Deshpande, William Zhao, Sophia Kolek, Dasha Sorokin, Kshitij Chintala, Vedanuj Arun, Ahmed Awadallah, Ryan Cheng-Yue, Yanping Chuang, et al. Llama 3: An updated open foundation language model. *arXiv preprint arXiv:2404.08647*, 2024.
- [23] Yucheng Shi, Hehuan Ma, Wenliang Zhong, Qiaoyu Tan, Gengchen Mai, Xiang Li, Tianming Liu, and Junzhou Huang. Chatgraph: Interpretable text classification by converting chatgpt knowledge to graphs. In 2023 IEEE International Conference on Data Mining Workshops (ICDMW), pages 515–520. IEEE, 2023.
- [24] Yucheng Shi, Tianze Yang, Canyu Chen, Quanzheng Li, Tianming Liu, Xiang Li, and Ninghao Liu. Searchrag: Can search engines be helpful for llm-based medical question answering? *arXiv preprint arXiv:2502.13233*, 2025.
- [25] Yucheng Shi, Shaochen Xu, Tianze Yang, Zhengliang Liu, Tianming Liu, Quanzheng Li, Xiang Li, and Ninghao Liu. Mkrag: Medical knowledge retrieval augmented generation for medical question answering. arXiv preprint arXiv:2309.16035, 2023.
- [26] Hao Zhen, Yucheng Shi, Yongcan Huang, Jidong J Yang, and Ninghao Liu. Leveraging large language models with chain-of-thought and prompt engineering for traffic crash severity analysis and inference. *Computers*, 13(9):232, 2024.
- [27] Zhiwen Fan, Pu Wang, Yang Zhao, Yibo Zhao, Boris Ivanovic, Zhangyang Wang, Marco Pavone, and Hao Frank Yang. Learning traffic crashes as language: Datasets, benchmarks, and what-if causal analyses. *arXiv* preprint *arXiv*:2406.10789, 2024.
- [28] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [29] AI@Meta. Llama 3 model card. 2024.
- [30] OpenAI. Gpt-4o (version 2024-11-20) [large language model], 2024. https://openai.com/index/hello-gpt-4o/.
- [31] OpenAI. Gpt-4o-mini (version 2024-07-18) [large language model], 2024. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.
- [32] Anthropic. Claude. https://www.anthropic.com/claude, 2023. Accessed: 2024.
- [33] OpenAI. Gpt-4 technical report. https://arxiv.org/abs/2303.08774, 2023.
- [34] Yilong Ren, Yue Chen, Shuai Liu, Boyue Wang, Haiyang Yu, and Zhiyong Cui. Tpllm: A traffic prediction framework based on pretrained large language models. *arXiv* preprint arXiv:2403.02221, 2024.
- [35] Ramya Jonnala, Gongbo Liang, Jeong Yang, and Izzat Alsmadi. Exploring the potential of large language models in public transportation: San antonio case study. *arXiv preprint arXiv:2501.03904*, 2025.

- [36] Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, et al. Usable xai: 10 strategies towards exploiting explainability in the llm era. *arXiv* preprint arXiv:2403.08946, 2024.
- [37] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, page 276. Association for Computational Linguistics, 2019.
- [38] Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, 2019.
- [39] Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, 2020.
- [40] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.
- [41] Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2341–2369, 2024.
- [42] Yucheng Shi, Qiaoyu Tan, Xuansheng Wu, Shaochen Zhong, Kaixiong Zhou, and Ninghao Liu. Retrieval-enhanced knowledge editing in language models for multi-hop question answering. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 2056–2066, New York, NY, USA, 2024. Association for Computing Machinery.
- [43] Yucheng Shi, Quanzheng Li, Jin Sun, Xiang Li, and Ninghao Liu. Enhancing cognition and explainability of multimodal foundation models with self-synthesized data. *arXiv preprint arXiv:2502.14044*, 2025.
- [44] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- [45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- [46] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020.