Article

Leveraging Large Language Models with Chain-of-Thought and Prompt Engineering for Traffic Crash Severity Analysis and Inference

Hao Zhen [1], Yucheng Shi [2], Yongcan Huang [1], Jidong J. Yang [1,*] and Ninghao Liu [2]

[1] School of Environmental, Civil, Agricultural, and Mechanical Engineering, University of Georgia, Athens, GA 30605, USA; hao.zhen@uga.edu (H.Z.); yh15555@uga.edu (Y.H.)

[2] School of Computing, University of Georgia, Athens, GA 30605, USA; yucheng.shi@uga.edu (Y.S.); ninghao.liu@uga.edu (N.L.)

* Correspondence: jidong.yang@uga.edu

Abstract: Harnessing the power of Large Language Models (LLMs), this study explores the use of three state-of-the-art LLMs, specifically GPT-3.5-turbo, LLaMA3-8B, and LLaMA3-70B, for crash severity analysis and inference, framing it as a classification task. We generate textual narratives from original traffic crash tabular data using a pre-built template infused with domain knowledge. Additionally, we incorporated Chain-of-Thought (CoT) reasoning to guide the LLMs in analyzing the crash causes and then inferring the severity. This study also examine the impact of prompt engineering specifically designed for crash severity inference. The LLMs were tasked with crash severity inference to: (1) evaluate the models' capabilities in crash severity analysis, (2) assess the effectiveness of CoT and domain-informed prompt engineering, and (3) examine the reasoning abilities with the CoT framework. Our results showed that LLaMA3-70B consistently outperformed the other models, particularly in zero-shot settings. The CoT and Prompt Engineering techniques significantly enhanced performance, improving logical reasoning and addressing alignment issues. Notably, the CoT offers valuable insights into LLMs' reasoning process, unleashing their capacity to consider diverse factors such as environmental conditions, driver behavior, and vehicle characteristics in severity analysis and inference.

Keywords: traffic safety; crash severity inference; crash causality reasoning; crash narrative generation; textual context; transformer; large language models; Chain-of-Thought; prompt engineering; zero-shot learning; few-shot learning

1. Introduction

Traffic safety research plays a crucial role in enhancing road safety by examining the root causes of accidents, identifying hazardous behaviors or factors, and proposing effective countermeasures [1]. Despite advancements in vehicle safety, enhancements in road design, and the implementation of various policies, traffic safety remains a significant challenge. One important aspect of road safety research is understanding contributing factors leading to different crash severity outcomes, which is essential for mitigating crash consequences.

The challenge of traffic accident modeling stems from their multifaceted nature, involving intricate interplay among diverse factors, such as human behavior, vehicle dynamics, traffic conditions, environmental factors, and roadway characteristics. Traffic safety research has been primarily focused on understanding causality using observational data, due to the impracticality of conducting controlled experiments in this field [1]. Traditional statistical and econometric methods have long been studied in traffic safety domain [2–6] for causality understanding. These classic methods suffer from several limitations, including constraints imposed by specific functional forms and distributional assumptions [1], as

well as subtle confounding effects, also referred as heterogeneous treatment effects [1,7,8], which often lead to an incomplete or misleading understanding of influencing factors.

Another limitation of statistical and econometric methods lies in the fact that they were designed around and can only consume structured data with traditional numeric or categorical coding and a limited number of features. These methods can not effectively handle unstructured textual data or passages of narratives. Due to recent advancements in AI and the abundance of narrative data captured in crash reports, natural language processing (NLP) methods have been applied in text mining of crash narratives [9–11]. In previous works, researchers are required to collect a large amount of high-quality, labeled crash reports for model training. However, this process is time-consuming and costly. Additionally, low-quality training data and poorly chosen training parameters can lead to undesirable performance. In contrast, large language models (LLMs) offer a distinct advantage by leveraging their immense knowledge acquired from extensive pre-training with vast datasets, effectively addressing these challenges. Motivated by their superior capability to comprehend and generate human-like text, this study aims to investigate whether LLMs can process complex and unstructured data in traffic safety domain to enable elaborate case-specific analysis.

Despite the release of numerous LLMs, represented by the GPT family [12] and LLaMA family [13], their ability for traffic crash analysis and reasoning remains unexplored. Applying LLMs to crash analysis presents two major challenges: (1) it requires LLMs to fully understand the domain knowledge and potential causality behind crash events. However, LLMs, typically built on transformer architectures, are often regarded as "black-box" models, making it difficult to interpret their decision-making processes. (2) while LLMs possess extensive real-world knowledge acquired from the pre-training stage, they are not specialized in analyzing textual data in crash reports. This creates an alignment gap between the model's original intent and the specific requirements of this specific task.

To address the first challenge, we propose to leverage the Chain-of-Thought (CoT) technique to enhance the reasoning capabilities of LLMs [14]. This technique guides the model through a structured reasoning process, helping it better understand the detailed knowledge and causality behind crash data. Additionally, the CoT approach provides explainable reasoning steps for each intermediate result, making the model's decisions more transparent and easier to interpret. By incorporating CoT, we aim to improve the LLMs' performance in crash severity modeling, leveraging their capability to effectively process the complex and diverse data relevant to crash analysis.

To address the second challenge, we propose to use prompt engineering (PE) and few-shot learning (FS) to better align the LLMs with the specific requirements of the target task: crash severity modeling and analysis. PE can tailor the input prompts to guide the LLMs toward more relevant and reliable analysis, while few-shot learning can provide the models with specific examples to improve their understanding and performance in the subject domain. By combining these techniques, we aim to bridge the alignment gap and enhance the models' ability to effectively analyze textual descriptions in crash reports.

To demonstrate the efficacy of our approach, we explore three state-of-the-art LLMs, specifically GPT-3.5-turbo, LLaMA3-8B, and LLaMA3-70B, for crash severity inference, framing it as a multi-class classification task. In our experiments, we utilized textual narratives derived from crash tabular data as input for crash severity analysis with LLMs. Additionally, we incorporated CoT to guide the LLMs in analyzing potential crash causes and subsequently inferring severity outcome. We also examine prompt engineering specifically designed for crash severity inference. We task LLMs with crash severity inference to (1) assess their capability for crash severity analysis; (2) evaluate the effectiveness of CoT and domain-informed PE; and (3) examine their reasoning ability within the CoT framework.

The experimental setup involves several strategies, including plain zero-shot and few-shot settings, zero-shot with Chain-of-Thought (ZS_CoT), zero-shot with prompt engineering (ZS_PE), zero-shot with both prompt engineering and Chain-of-Thought

(ZS_PE_CoT), and few-shot with prompt engineering (FS_PE). The LLMs evaluated include GPT-3.5-turbo, LLaMA3-8B, and LLaMA3-70B, with specific hyperparameters to ensure consistent and reliable results. We compare the performance of these models and settings to determine the most effective approach for the crash severity inference task.

2. Methods

2.1. Model Descriptions

In our approach, we leverage the reasoning ability of LLMs for domain-specific (i.e., traffic safety) analysis and modeling. Specifically, we utilize two state-of-the-art LLMs as our base models, including GPT-3.5-turbo [12] and LLaMA-3 [13]. Both models are auto-regressive language models, which generate text by predicting the next word or subword in a sequence based on the previous words or subwords, one step at a time. This approach allows the model to create coherent and contextually relevant text. In this auto-regressive setting, the joint probability is expressed as the product of sequential conditional probabilities in Equation (1):

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid x_1, x_2, \ldots, x_{i-1}), \tag{1}$$

where $P(x_i \mid x_1, x_2, \ldots, x_{i-1})$ represents the conditional probability of the *i*-th word given the preceding words. This auto-regressive modeling framework, designed to handle large context windows and trained on extensive datasets, empowers the model to generate fluent and context-aware sequences. Some details about these two models are provided below for direct reference.

GPT-3.5-turbo [12], developed by OpenAI, is designed for a variety of applications, including advanced text generation, coding assistance, and more. Trained on a vast corpus of internet text, including diverse sources such as books and websites, it has billions of parameters, allowing for nuanced understanding and generation of text. In our experiments, we use the gpt-3.5-turbo-0125 version.

LLaMA-3 [13], developed by Meta, is another state-of-the-art LLM designed for efficient and scalable language understanding. It is pre-trained on over 15T tokens that cover diverse internet text sources. LLaMA-3 is available in sizes of 8 billion and 70 billion parameters, making it flexible for various applications.

Both models employ techniques like supervised fine-tuning (SFT) and Reinforcement Learning with Human Feedback (RLHF) to align their outputs with human preferences, ensuring the model is helpful and safe [15]. Specifically, SFT fine-tunes a pre-trained model on specific datasets with human instructions, ensuring it understands relevant vocabulary, patterns, and knowledge, thereby improving accuracy and relevance. RLHF, on the other hand, refines the model using feedback from human experts, allowing it to adapt to complex real-world scenarios and prioritize critical safety aspects, enhancing both adaptability and reliability.

2.2. In-Context Learning

In-context learning is a promising approach for traffic safety analysis and modeling, where a LLM learns to perform a specified task by observing examples of the task within the input context, without any explicit fine-tuning or gradient updates [16]. The LLM leverages its pre-existing knowledge and the provided examples to generate appropriate inference for new, unseen instances of the task. In-context learning encompasses both zero-shot and few-shot learning [17], with the number of examples ranging from zero to a few. In the context of traffic safety analysis and modeling, in-context learning can be applied to various tasks, such as crash severity inference.

Zero-shot learning in traffic safety analysis refers to the setting where the model is expected to perform a task without any examples or explicit training for that specific task. The model relies solely on its pre-existing knowledge to make inference. While few-shot learning involves providing the model with a few examples of the target task before

asking it to perform the same task on new instances. The model learns from these few examples and adapts its behavior accordingly. For instance, in crash severity inference, the model may be provided with a few examples of crashes and their corresponding severity outcomes. The model then learns from these examples and reasons the severity outcome of a new, unseen crash. Few-shot learning allows the models to quickly adapt to new tasks or scenarios with minimal training data for diverse real-world applications [18].

2.3. Chain-of-Thoughts (CoT)

A notable distinction between LLMs and conventional machine learning algorithms is the capacity of foundation LLMs to process variably structured input data, specifically prompts, during the inference phase [14]. For inference, LLMs generally precede or prioritize the information provided in the input prompts over the large implicit knowledge gained from the pre-training stage. Consequently, this leads to clear, comprehensive content. This content could encompass domain-specific knowledge, contextual background, or detailed step-by-step reasoning guidance. By doing so, LLMs can be encouraged to disclose their reasoning processes during inference, thereby enhancing the explicability of their outputs. In this paper, we utilize the CoT technique to enable LLMs to generate step-by-step reasoning before providing the final answer, improving their performance on complex reasoning tasks.

Although it is widely recognized that LLMs excel at generating responses reminiscent of human conversation, they often have opacity issues in their reasoning processes. This opacity hinders users' ability to evaluate the trustworthiness of the responses, particularly in scenarios that necessitate elaborate reasoning.

Formally, let f_θ be a language model and $X = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n)\}$ be an input prompt, where (x_i, y_i) denotes the *i*-th example question-response pair. In a standard question-answering scenario, the model output is given by: $y_n = \arg\max_Y p_\theta(Y|x_1, y_1, x_2, y_2, \ldots, x_n)$, where p_θ is the predicted probability of the language model f_θ. This setting, however, does not provide insights into the reasoning process behind the answer y_n. The CoT method extends this by including human-crafted explanations e_i for each example, resulting in a modified input format $X = (x_1, e_1, y_1), (x_2, e_2, y_2), \ldots, (x_n)$. The model then outputs both the explanation e_n and the answer y_n:

$$(e_n, y_n) = \arg\max_Y p_\theta(Y|x_1, e_1, y_1, x_2, e_2, y_2, \ldots, x_n). \tag{2}$$

However, in practice, it is difficult to obtain sufficient high-quality explanation examples for crash severity classification, while low-quality explanations can harm the CoT performance. Therefore, following strategies in [19], we simply ask LLMs to think step by step to conduct traffic safety analysis and inference. Some template examples are presented in the Section 4. Besides allowing for a more transparent and understandable interaction with LLMs, the CoT approach is also practically useful in several key aspects: (1) Reducing Errors in Crash Risk Assessment: By breaking down complex traffic scenarios, CoT can better understand and reason over specific cases, thus reducing errors in crash risk analysis [14,20,21]. (2) Providing Adjustable Intermediate Steps for Crash Analysis: CoT enables the outlining of traceable intermediate steps within the crash analysis process. This can be helpful in identifying potential biases or errors in the model's reasoning and allow for more reliable crash risk assessment [22]. By leveraging the CoT approach in traffic safety analysis, we can enhance the interpretability and reliability of crash risk assessment.

2.4. Prompt Engineering (PE)

Prompt engineering is a technique for crafting and refining input prompts for LLMs to enhance their performance on specific tasks. It is akin to "asking the right question (prompt) to get the desired answer (response)". For traffic safety analysis, carefully designed prompts can significantly enhance LLMs' ability to analyze complex scenarios and provide meaningful insights. Specifically, in the context of crash severity inference, one

critical category is 'Fatal accidents'. However, LLMs often exhibit a tendency to avoid assigning this label to relevant cases. This behavior stems from their alignment training, which generally encourages them to avoid discussing unpleasant subject or potentially unsafe topics [23,24]. Such design constraints present a challenge in accurately inferring severe outcomes for traffic incidents. To address this issue, we found it necessary to rephrase the original 'Fatal accidents' label using alternative terms. This "soft" approach allows us to maintain inference accuracy while respecting LLMs' aligned parameters. The specific modifications and their impacts on inference performance are discussed in detail in the Section 4.

We understand that while some literature considers CoT a subset of PE, in this study, we distinguish between the two to emphasize the unique contributions of each approach to traffic safety analysis. We explicitly define Prompt Engineering as the technique of crafting and refining input prompts to guide the model toward more relevant and reliable outputs. In contrast, Chain-of-Thoughts is considered as a specific reasoning technique that encourages the model to break down complex problems into intermediate steps, providing a more transparent and structured reasoning process.

3. Data

In this section, we first discuss the dataset employed for the study. We then explain how we convert the crash tabular data to coherent descriptive narratives. Finally, we discuss our experimental settings and evaluation methods.

3.1. Dataset

Our empirical analysis utilizes data sourced from CrashStats data from Victoria, Australia spanning from 2006 to 2020. The crash database contains records of vehicles involved in crashes. A four-point ordinal scale is used to code the severity level of each accident, including: (1) non-injury accident, (2) other injury (minor injury) accident, (3) serious injury accident, and (4) fatal accident. Each sample denotes a vehicle involved in a crash with driver's information. After data prepossessing, the final dataset has an extremely low representation of non-injury accidents (only four instances), accounting for less than 0.001%. Consequently, these four non-injury accidents are merged to the category of "Minor or non-injury accidents". As a result, the dataset contains 197,425 minor or non-injury accidents, 89,925 serious injury accidents, and 4760 fatal accidents.

The traffic accident attributes considered in our empirical study include crash characteristics, driver traits, vehicle details, roadway attributes, environmental conditions, and situational factors (see Table 1).

Table 1. Traffic accident attributes.

Variables	Description
Crash characteristics	
ACCIDENT_TYPE	The type of accident.
EVENT_TYPE	Type of incident event.
VEHICLE_1_COLL_PT	Collision point on the first vehicle involved in the event.
VEHICLE_2_COLL_PT	Collision point on the second vehicle involved in the event.
OBJECT_TYPE	Object involved in the specific accident event.
DCA	The definitions for classifying accidents.
ACCIDENT_MONTH	The month in which the accident occurred, derived from "ACCIDENT_DATE".
TIME_PERIOD	The period the accident occurred, derived from "ACCIDENT_TIME".
DAY_OF_WEEK	The day of the week the accident occurred.
LGA_NAME	The name of local government areas.

Table 1. *Cont.*

Variables	Description
REGION_NAME	The region where the accident occurred.
DEG_URBAN_NAME	The type of urbanized area for the crash site.
Driver characteristics	
DRIVER_SEX	The sex of the driver.
AGE_GROUP	The age group of the driver, derived from "DRIVER_AGE".
ROAD_USER_TYPE	The role of the person was at the time of the accident.
Vehicle characteristics	
VEHICLE_TYPE	The type or category of vehicle.
VEHICLE_WEIGHT	The weight or mass of the vehicle. The unit of measurement is kilograms.
NO_OF_WHEELS	The number of wheels that the vehicle has.
SEATING_CAPACITY	The number of seats in the vehicle.
FUEL_TYPE	The type of fuel used by the vehicle.
VEHICLE_AGE	The age of the vehicle when the accident occurred.
VEHICLE_BODY_STYLE	The body type of the vehicle.
TRAILER_TYPE	The type of trailer towed by the vehicle involved in the accident.
Roadway attributes	
ROAD_TYPE	Type of the highest priority road at the intersection or the road the accident occurred.
ROAD_GEOMETRY	The layout of the road where the accident occurred.
SPEED_ZONE	The speed zone at the location of the accident.
ROAD_SURFACE_TYPE	The type of road surface: 1: Paved 2: Unpaved 3: Gravel 9: Not known.
ROAD_TYPE_INT	The type or suffix of the intersecting road.
COMPLEX_INT_NO	Whether or not the segment is part of a complex intersection.
Environmental factors	
LIGHT_CONDITION	The light condition or level of brightness at the time of the accident.
SURFACE_COND	Road surface condition: dry, wet, muddy, snowy, icy, unknown.
SURFACE_COND_SEQ	Starts with 1 and incremented by 1 if more than one road surface condition.
ATMOSPH_COND	Atmospheric condition.
ATMOSPH_COND_SEQ	1 and incremented by 1 if more than one atmospheric condition is entered.
Situational factors	
HELMET_BELT_WORN	Whether or not the person was wearing a helmet or seatbelt at the time of the accident.
NO_OF_VEHICLES	The number of vehicles involved in the accident.
LAMPS	Whether the lamps or headlights for the vehicle were alight (on).
VEHICLE_MOVEMENT	The actual movement of the vehicle before the accident.
TRAFFIC_CONTROL	The type of traffic control measure in the location where the accident occurred.
NO_PERSONS	The number of people involved in the accident.
NO_OCCUPANTS	The number of occupants or people in the vehicle at the time of the accident.
SUB_DCA	SUB_DCA code and description of the accident.
SUB_DCA_SEQ	Starts with 1 and is incremented by 1 if more than one sub_dca is entered.
DRIVER_INTENT	The intent of the driver initially.

3.2. Textual Narrative Generation

To get coherent, informative passages enriched with domain-specific knowledge, we use a simple yet effective template to convert the raw structured tabular data into detailed, human-readable textual narratives, encapsulating vital information about traffic accidents, which can be better consumed by LLMs. This process is depicted in Figure 1.
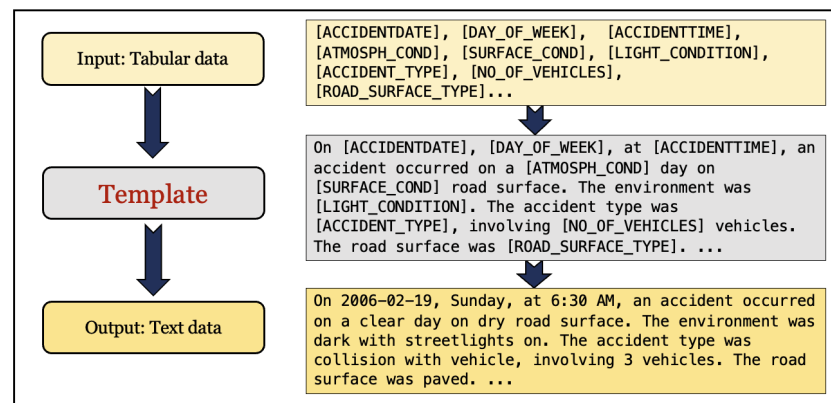
Figure 1. Illustration of textual narrative generation.

The primary objective is to augment the applicability and relevance of tabular data as input for LLMs, facilitating more context-aware inference for tasks like accident severity assessment. Furthermore, road safety engineers can supplement it with established facts or domain-specific knowledge for further enhancement. For instance, "Children and elders are typically more vulnerable in accidents without seat belts". However, it is important to note that, for this study, we do not delve into the intricate design of scientific knowledge in traffic safety.

4. Experiments

4.1. Experiments Design

In this work, we tackle the crash severity inference problem as a classification task. The inputs encompass various crash related attributes, including environmental conditions, driver characteristics, crash details, and vehicle features. The original data is in tabular format including categorical and numerical fields. We then transform the tabular data into consistent textual narrative with a simple template, detailed in the preceding subsection. The objective is to estimate severity outcomes of crashes with the state-of-the-art LLM models, such as GPT-3.5-turbo, LLaMA3-8B, and LLaMA3-70B.

LlaMA3s models are open-source foundation models, while the GPT model is a close-source and non-free model. Considering the cost and our goal of evaluating crash severity inference performance of foundation models, we randomly draw 50 samples from each of the three severity outcome categories, resulting in a total of 150 samples. These samples are used to demonstrate the potential of LLMs in enhancing crash analysis and reasoning.

The strategies outlined in Table 2 includes: Zero-shot and Few-shot settings coupled with different techniques. We use ZS and FS to denote the plain zero-shot and few-shot setting without prompt engineering and Chain-of-Thought. Other settings are zero-shot with Chain-of-Thought (ZS_CoT), zero-shot with prompt engineering (ZS_PE), zero-shot with prompt engineering and Chain-of-Thought (ZS_PE_CoT), and few-shot with prompt engineering (FS_PE). It worth noting that some literature treats Chain-of-Thought (CoT) as a special form of Prompt Engineering (PE). In this paper, we make distinction between PE and CoT to highlight the advantages of CoT over a basic PE approach in the context of traffic safety analysis.

With these experiments, we aim to determine: (1) the accident severity inference performance of LLMs in a plain zero-shot setting; (2) whether CoT enhances the performance through its reasoning process (ZS_CoT vs ZS; ZS_PE vs ZS_PE_CoT); (3) whether PE improves performance in zero-shot and few-shot settings (ZS_PE vs ZS; FS_PE vs FS); (4) whether few-shot learning boost performance compared to the zero-shot setting (ZS vs. FS). Accordingly, we tested six prompts to automatically infer the severity outcome of crashes. The details of these prompts are presented in the following section.

Table 2. Experiments.

Experiments Setting				
	Plain	Chain-of-Thought	Prompt Engineering	Prompt Engineering with Chain-of-Thought
Zero-shot	ZS	ZS_CoT	ZS_PE	ZS_PE_CoT
Few-shot	FS	/	FS_PE	/
Models				
Sampling Strategy (temperature, top_p)	GPT-3.5-turbo Greedy (0, 0.01)	LLaMA3-8B Greedy -	LLaMA3-70B Greedy -	

The experiments were conducted using GPT-3.5-turbo, LLaMA3-8B, and LLaMA3-70B. For GPT-3.5-turbo and LLaMA3, the hyperparameters are configured with temperature = 0 and top_p = 0.0001 for crash severity inference, aiming to produce consistent and reliable results through greedy decoding. Additionally, the LLaMA3 models are set with $do\_sample = FALSE$ to generate deterministic outputs, ensuring reproducibility.

4.2. Prompts for LLMs

In this section, we explained in detail how we design the prompts for different experiments in Table 2.

4.2.1. Zero-Shot

The prompt designed for plain zero-shot setting is demonstrated in Figure 2. The provided prompt tasks each LLM as a professional road safety engineer with classifying the severity of a traffic crash in Victoria, Australia, based on a detailed description of a crash. The engineer is required to categorize the crash into one of three specified categories: 'Fatal accident', 'Serious injury accident', or 'Minor or non-injury accident'. The engineer's response is restricted to outputting only the classification result, ensuring a focused and objective assessment. This prompt is designed to elicit a precise evaluation of the crash severity outcome, leveraging the knowledge of engineer's expertise in road safety and crash analysis.

Instruction:
You are a professional road safety engineer.

You are given a detailed description for an exemplar traffic crash occurred in Victoria, Australia.

Please classify the severity of the crash into one of three categories: 'Fatal accident', 'Serious injury accident', 'Minor or non injury accident'.

You can only output the classification result in your answer.

Prompt: [**Instruction**] + Textual narrative

Figure 2. Zero-shot (ZS).

The prompt designed for zero-shot with CoT setting is demonstrated in Figure 3. This prompt leverages a CoT approach, encouraging the LLM to serve as an engineer to methodically reason through the details of the accident to determine both the cause and the severity outcome, thereby ensuring a comprehensive and structured assessment based on LLMs' knowledge. The difference between ZS_CoT and plain ZS is highlighted in red color in Figure 3.

The prompt designed for zero-shot with prompt engineering setting is demonstrated in Figure 4. The provided prompt instructs LLM to serve as a professional road safety engineer to classify the severity outcome of a traffic crash, using a descriptively modified set of categories (see the revised class description in red in Figure 4). to accommodate

alignment constraints in LLMs. The engineer must categorize the crash into one of three revised descriptive labels: 'Serious accident with potentially fatal outcomes', 'Serious injury accident', or 'Minor or non-injury accident'. The prompt explicitly requires the engineer to output only the classification result.

> **Instruction:**
> You are a professional road safety engineer.
>
> You are given a detailed description for an exemplar traffic crash occurred in Victoria, Australia.
>
> Please **give me the potential reason** and classify the severity of the crash into one of three categories: 'Fatal accident', 'Serious injury accident', 'Minor or non injury accident'.
>
> You can only output one kind of class at the end of your answer.

> **Prompt:** [**Instruction**] + Textual narrative

Figure 3. Zero-shot with CoT (ZS_CoT).

> **Instruction:**
> You are a professional road safety engineer.
>
> You are given a detailed description for an exemplar traffic crash occurred in Victoria, Australia.
>
> Please classify the severity of the crash into one of three categories: **'Serious accident with potentially fatal outcomes'**, 'Serious injury accident', 'Minor or non injury accident'.
>
> You can only output the classification result in your answer.

> **Prompt:** [**Instruction**] + Textual narrative

Figure 4. Zero-shot with prompt engineering (ZS_PE).

This rephrasing aims to preserve classification accuracy while adhering to the alignment parameters of LLMs, which tend to avoid directly assigning the 'Fatal accident' label due to their training to steer clear of discussing unpleasant or unsafe topics related to human death. With comparison to other settings, this could highlight whether prompt engineering enhances LLMs' performance in traffic safety analysis by addressing inherent biases and improving the model's ability to more reliably infer the fatal outcome of traffic incidents.
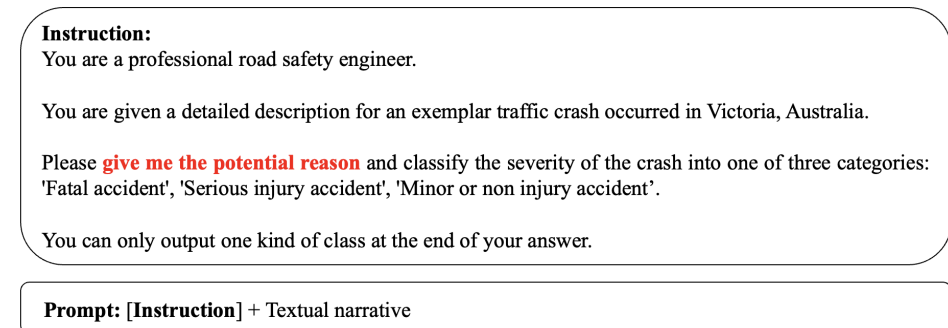
The prompt designed for Zero-shot with PE & CoT setting is shown in Figure 5. In this prompt design, we not only included the CoT, by requiring a logical deduction from cause reasoning to severity outcome classification, but also changed the class label of 'Fatal accident' to the soft version of 'Serious accident with potentially fatal outcomes'. The differences between ZS_PE_CoT and plain ZS are highlighted in red in Figure 5.

> **Instruction:**
> You are a professional road safety engineer.
>
> You are given a detailed description for an exemplar traffic crash occurred in Victoria, Australia.
>
> Please **give me the potential reason** and classify the severity of the crash into one of three categories: **'Serious accident with potentially fatal outcomes'**, 'Serious injury accident', 'Minor or non injury accident'.
>
> You can only output one kind of class at the end of your answer.

> **Prompt:** [**Instruction**] + Textual narrative

Figure 5. Zero-shot with prompt engineering & CoT (ZS_PE_CoT).

4.2.2. Few-Shot

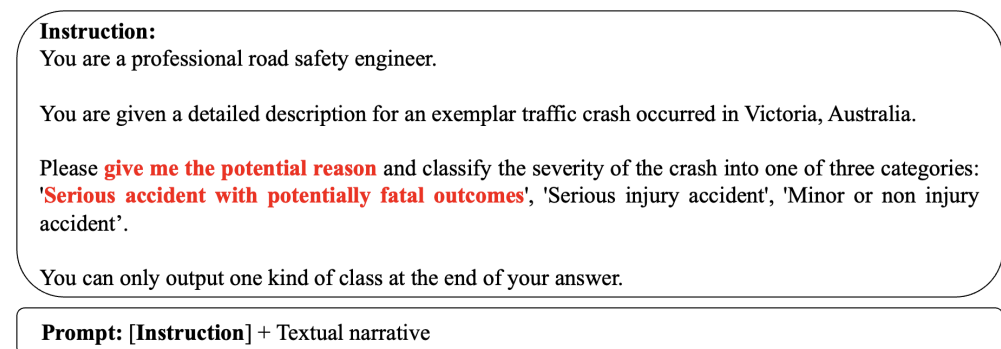The prompt designed for the plain few-shot setting is shown in Figure 6. In this paper, the few-shot setting refers to a three-shot scenario, where three examples, one from each severity category, are provided for the LLMs to infer from.

Instruction:
You are a professional road safety engineer.

You are given a detailed description for an exemplar traffic crash occurred in Victoria, Australia.

Please classify the severity of the crash into one of three categories: 'Serious accident with potentially fatal outcomes', 'Serious injury accident', 'Minor or non injury accident'.

You can only output the classification result in your answer.

On 2020-09-11 00:00:00, Friday, at 22:00:00, an accident occurred on a Clear day with Dark No streetlights condition. The road surface condition is dry. The accident type is Struck animal, event type is Collision, and the object involved in the accident is Wombat. It happened at a Not at intersection with 1.0 road surface, involving 1 vehicle. The speed zone for the location where the accident happened is 100 km/hr. The traffic control when the accident happened was No control. The DCA code for this accident is 167 and the description for the DCA code is STRUCK ANIMAL. The sub-DCA code description is Tree (shrub/scrub).
The number of vehicles involved in the accident is 1 and the number of occupants in the vehicle at the time of the accident is 1.0. The driver involved in the primary vehicle, identified as A, is a female of age 60.0. The driver was wearing a seatbelt. The driver is not a driver of oversize or overmass large vehicles. The actual movement of the vehicle prior to the accident is Going straight ahead. When the accident happened, the driver was driving a 15 year old Car. The fuel type of this vehicle is Petrol. The vehicle weighs nan kilograms and has 4.0 wheels. The lamp or headlights for the vehicle was on. The vehicle has 5.0 seats with body style 'Sedan - Car with external boot'. The trailer type is Not applicable. The accident took place on SOUTH GIPPSLAND in region EASTERN REGION in area of RURAL_VICTORIA. It happened on road SOUTH GIPPSLAND near the intersection with road BOOLARRA-FOSTER. It did happen on a complex intersection. The vehicles involved were identified as A and with collision points on Front and respectively. The accident did not involve any pedestrians (No pedestrian movement).
#Classification#
Serious accident with potentially fatal outcomes

You are a professional road safety engineer.

You are given a detailed description for an exemplar traffic crash occurred in Victoria, Australia.

Please classify the severity of the crash into one of three categories: 'Serious accident with potentially fatal outcomes', 'Serious injury accident', 'Minor or non injury accident'.

You can only output the classification result in your answer.

On 2006-01-14 00:00:00, Saturday, at 15:55:00, an accident occurred on a Clear day with Day condition. The road surface condition is dry. The accident type is Collision with vehicle, event type is Collision, and the object involved in the accident is Not Applicable. It happened at a Not at intersection with 1.0 road surface, involving 2 vehicles. The speed zone for the location where the accident happened is 100 km/hr. The traffic control when the accident happened was No control. The DCA code for this accident is 120 and the description for the DCA code is HEAD ON (NOT OVERTAKING). The sub-DCA code description is Road curved at mid-block.
The number of vehicles involved in the accident is 2 and the number of occupants in the vehicle at the time of the accident is 1.0. The driver involved in the primary vehicle, identified as B, is a male of age 63.0. The driver was wearing a seatbelt. The driver is not a driver of oversize or overmass large vehicles. The actual movement of the vehicle prior to the accident is Going straight ahead. When the accident happened, the driver was driving a 9 year old Car. The fuel type of this vehicle is Unknown. The vehicle weighs nan kilograms and has nan wheels. The vehicle was likely don't have any headlights. The vehicle has nan seats with body style 'Sedan - Car with external boot'. The trailer type is Not known what is being towed. The accident took place on MURRINDINDI in region NORTHEASTERN REGION in area of RURAL_VICTORIA. It happened on road MAROONDAH near the intersection with road ROUCH. It did happen on a complex intersection. The vehicles involved were identified as A and B with collision points on Front and Towed unit respectively. The accident did not involve any pedestrians (No pedestrian movement).
#Classification#
Serious injury accident

You are a professional road safety engineer.

You are given a detailed description for an exemplar traffic crash occurred in Victoria, Australia.

Please classify the severity of the crash into one of three categories: 'Serious accident with potentially fatal outcomes', 'Serious injury accident', 'Minor or non injury accident'.

You can only output the classification result in your answer.

On 2006-01-25 00:00:00, Wednesday, at 14:15:00, an accident occurred on a Clear day with Day condition. The road surface condition is dry. The accident type is Collision with a fixed object, event type is Collision, and the object involved in the accident is Not Applicable. It happened at a Not at intersection with 1.0 road surface, involving 2 vehicles. The speed zone for the location where the accident happened is 50 km/hr. The traffic control when the accident happened was No control. The DCA code for this accident is 146 and the description for the DCA code is REVERSING INTO FIXED OBJECT/PARKED VEHICLE. The sub-DCA code description is Buildings.
The number of vehicles involved in the accident is 3 and the number of occupants in the vehicle at the time of the accident is 2.0. The driver involved in the primary vehicle, identified as B, is a male of age 83.0. The driver was wearing a seatbelt. The driver is not a driver of oversize or overmass large vehicles. The actual movement of the vehicle prior to the accident is Reversing. When the accident happened, the driver was driving a 5-year-old Car. The fuel type of this vehicle is Unknown. The vehicle weighs nan kilograms and has nan wheels. The vehicle was likely don't have any headlights. The vehicle has nan seats with body style 'Sedan - Car with external boot'. The trailer type is Not applicable. The accident took place on BAW BAW in region EASTERN REGION in area of TOWNS. It happened on road HAMP near the intersection with road HOPE. It did happen on a complex intersection. The vehicles involved were identified as B and A with collision points on Front and Rear, respectively. The accident did not involve any pedestrians (No pedestrian movement).
#Classification#
Minor or non injury accident

Prompt: [Instruction] + Textual narrative

Figure 6. Few shot (FS).

The only difference between the prompt for Few-shot with Prompt Engineering (FS_PE) and that of the plain few-shot is that we substitute "Fatal accidents" with "Serous accidents with potentially fatal outcomes".

4.3. Evaluation Metrics

Following standard practice in the context of multi-class classification, we adopt two commonly used classification metrics: Macro-Accuracy, and F1-score. Additionally, we include class-specific accuracies. These metrics are briefly discussed below.

(1) **Accuracy.** Accuracy measures the proportion of correctly classified instances in the test dataset. It is calculated as:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \tag{3}$$

where:

- Correct Predictions: The number of correctly classified instances in the test dataset.
- Total Predictions: The total number of instances in the test dataset.

It should be noted that we first calculate the accuracy for each class and then calculate the macro-accuracy as the average of these class accuracies.

(2) **F1-score.** F1-score is defined as the harmonic mean of precision and recall, computed as:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The F1-score reported in the following section (Section 5) is at the macro level, which is an averaged F1 of all classes.

Precision quantifies the accuracy of positive predictions for a specific class, computed as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{5}$$

where:

- True Positives: The number of correctly predicted instances of the class.
- False Positives: The number of instances wrongly classified into the class.

Recall, also known as sensitivity or true positive rate, measures the ability of the model to correctly identify instances of a specific class. It is calculated as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{6}$$

where:

- False Negatives: The number of instances of the class wrongly classified as something else.

5. Findings

5.1. Exemplar Responses of LLMs to Crash Severity Inference Queries

The exemplar responses of GPT-3.5, LLaMA3-8B, and LLaMA3-70B in each of the six settings, including ZS, FS, ZS_CoT, FS_PE, ZS_PE, and ZS_PE_CoT (outlined in Table 2), are shown in Figure 7. It demonstrates that the LLMs can effectively respond to the severity inference task, delivering expected results. Note that the examples in Figure 7 only showcase correct severity inferences.

Given the prompt for each setting (see Section 4.2), each model can directly answer or ultimately summarize its estimated severity for the given accident as one of the defined categories.

In the plain Zero-shot and Few-shot settings, the models respond directly with one of the three class labels, i.e., "Minor or non-injury accident", "Serious injury accident", or "Fatal accident". Similarly, in the ZS_PE and FS_PE settings, the models respond directly as "Minor or non-injury accident", "Serious injury accident", or "Serious accident with potentially fatal outcomes".
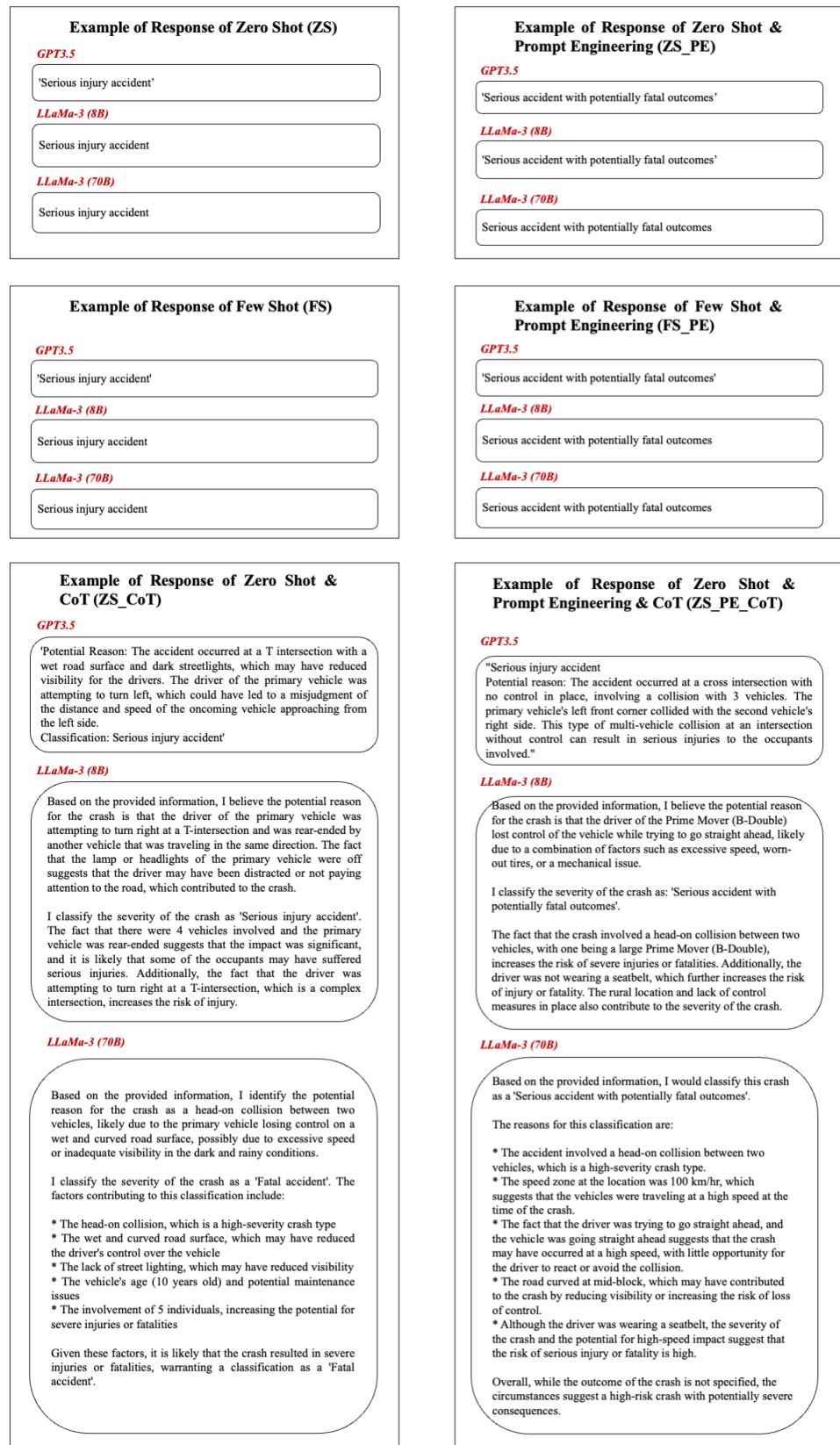

Example of Response of Zero Shot (ZS)

GPT3.5

'Serious injury accident'

LLaMa-3 (8B)

Serious injury accident

LLaMa-3 (70B)

Serious injury accident

Example of Response of Zero Shot & Prompt Engineering (ZS_PE)

GPT3.5

'Serious accident with potentially fatal outcomes'

LLaMa-3 (8B)

'Serious accident with potentially fatal outcomes'

LLaMa-3 (70B)

Serious accident with potentially fatal outcomes

Example of Response of Few Shot (FS)

GPT3.5

'Serious injury accident'

LLaMa-3 (8B)

Serious injury accident

LLaMa-3 (70B)

Serious injury accident

Example of Response of Few Shot & Prompt Engineering (FS_PE)

GPT3.5

'Serious accident with potentially fatal outcomes'

LLaMa-3 (8B)

Serious accident with potentially fatal outcomes

LLaMa-3 (70B)

Serious accident with potentially fatal outcomes

Example of Response of Zero Shot & CoT (ZS_CoT)

GPT3.5

'Potential Reason: The accident occurred at a T intersection with a wet road surface and dark streetlights, which may have reduced visibility for the drivers. The driver of the primary vehicle was attempting to turn left, which could have led to a misjudgment of the distance and speed of the oncoming vehicle approaching from the left side.
Classification: Serious injury accident'

LLaMa-3 (8B)

Based on the provided information, I believe the potential reason for the crash is that the driver of the primary vehicle was attempting to turn right at a T-intersection and was rear-ended by another vehicle that was traveling in the same direction. The fact that the lamp or headlights of the primary vehicle were off suggests that the driver may have been distracted or not paying attention to the road, which contributed to the crash.

I classify the severity of the crash as 'Serious injury accident'. The fact that there were 4 vehicles involved and the primary vehicle was rear-ended suggests that the impact was significant, and it is likely that some of the occupants may have suffered serious injuries. Additionally, the fact that the driver was attempting to turn right at a T-intersection, which is a complex intersection, increases the risk of injury.

LLaMa-3 (70B)

Based on the provided information, I identify the potential reason for the crash as a head-on collision between two vehicles, likely due to the primary vehicle losing control on a wet and curved road surface, possibly due to excessive speed or inadequate visibility in the dark and rainy conditions.

I classify the severity of the crash as a 'Fatal accident'. The factors contributing to this classification include:

* The head-on collision, which is a high-severity crash type
* The wet and curved road surface, which may have reduced the driver's control over the vehicle
* The lack of street lighting, which may have reduced visibility
* The vehicle's age (10 years old) and potential maintenance issues
* The involvement of 5 individuals, increasing the potential for severe injuries or fatalities

Given these factors, it is likely that the crash resulted in severe injuries or fatalities, warranting a classification as a 'Fatal accident'.

Example of Response of Zero Shot & Prompt Engineering & CoT (ZS_PE_CoT)

GPT3.5

"Serious injury accident
Potential reason: The accident occurred at a cross intersection with no control in place, involving a collision with 3 vehicles. The primary vehicle's left front corner collided with the second vehicle's right side. This type of multi-vehicle collision at an intersection without control can result in serious injuries to the occupants involved."

LLaMa-3 (8B)

Based on the provided information, I believe the potential reason for the crash is that the driver of the Prime Mover (B-Double) lost control of the vehicle while trying to go straight ahead, likely due to a combination of factors such as excessive speed, worn-out tires, or a mechanical issue.

I classify the severity of the crash as: 'Serious accident with potentially fatal outcomes'.

The fact that the crash involved a head-on collision between two vehicles, with one being a large Prime Mover (B-Double), increases the risk of severe injuries or fatalities. Additionally, the driver was not wearing a seatbelt, which further increases the risk of injury or fatality. The rural location and lack of control measures in place also contribute to the severity of the crash.

LLaMa-3 (70B)

Based on the provided information, I would classify this crash as a 'Serious accident with potentially fatal outcomes'.

The reasons for this classification are:

* The accident involved a head-on collision between two vehicles, which is a high-severity crash type.
* The speed zone at the location was 100 km/hr, which suggests that the vehicles were traveling at a high speed at the time of the crash.
* The fact that the driver was trying to go straight ahead, and the vehicle was going straight ahead suggests that the crash may have occurred at a high speed, with little opportunity for the driver to react or avoid the collision.
* The road curved at mid-block, which may have contributed to the crash by reducing visibility or increasing the risk of loss of control.
* Although the driver was wearing a seatbelt, the severity of the crash and the potential for high-speed impact suggest that the risk of serious injury or fatality is high.

Overall, while the outcome of the crash is not specified, the circumstances suggest a high-risk crash with potentially severe consequences.

Figure 7. Exemplar responses of LLMs in different settings.

In contrast, in the CoT settings (ZS_CoT and ZS_PE_CoT), the models return longer responses by reasoning first and then making inference of the severity outcome of each accident. Generally, the GPT-3.5 model's responses are more concise.

5.2. Severity Inference Performance of the LLMs with Different Strategies

The performance metrics of GPT-3.5, LLaMA3-8B, and LLaMA3-70B for the crash severity inference task under the six settings (see Table 2) are presented in Table 3.

Table 3. Performance of Models on Crash Severity Inference Task.

Model	Macro F1-Score	Macro-Accuracy	Fatal Accident	Serious Injury Accident	Minor or Non-Injury Accident
ZS					
GPT-3.5	0.1812	0.3400	0.00	1.00	0.02
LLaMA3-8B	0.1818	0.3400	0.00	1.00	0.02
LLaMA3-70B	0.4541	0.4533	0.44	0.34	0.58
ZS_CoT					
GPT-3.5	0.2073	0.3533	0.00	1.00	0.06
LLaMA3-8B	0.2496	0.3533	0.00	0.88	0.18
LLaMA3-70B	<u>0.4747</u>	<u>0.4733</u>	0.40	0.64	<u>0.38</u>
ZS_PE					
GPT-3.5	0.3798	0.4533	<u>0.62</u>	0.72	0.02
LLaMA3-8B	0.3120	0.4000	0.34	0.86	0.00
LLaMA3-70B	0.4755	0.4933	0.60	0.66	0.22
ZS_PE_CoT					
GPT-3.5	0.3509	0.4200	0.68	0.56	0.02
LLaMA3-8B	0.4033	0.4533	0.60	0.68	0.08
LLaMA3-70B	0.3581	0.4267	<u>0.62</u>	0.64	0.02
FS					
GPT-3.5	0.2514	0.3667	0.04	0.96	0.10
LLaMA3-8B	0.4068	0.4267	0.22	0.72	0.34
LLaMA3-70B	0.4131	0.4200	0.26	0.64	0.36
FS_PE					
GPT-3.5	0.2576	0.3667	0.18	0.92	0.00
LLaMA3-8B	0.2928	0.3933	0.08	<u>0.98</u>	0.12
LLaMA3-70B	0.3856	0.4600	0.56	0.80	0.02

Bold values are the best per column. <u>Underlined</u> values are the second best per column.

The results reveal varied performance across models and settings in inferring crash severity outcomes. LLaMA3-70B consistently exhibited superior performance, particularly with zero-shot prompt engineering (ZS_PE), achieving the highest macro F1-score (0.4755) and macro-accuracy (49.33%). Furthermore, LLaMA3-70B attained the second-best performance in macro F1-score (0.4747) and macro-accuracy (47.33%) under the zero-shot with Chain-of-Thought (ZS_CoT) setting. These findings suggest that both prompt engineering and Chain-of-Thought methodologies contribute positively to model performance. Nevertheless, no single technique demonstrated consistent superiority across all severity categories. For fatal accidents, GPT-3.5 with zero-shot prompt engineering and Chain-of-Thought (ZS_PE_CoT) exhibited the highest accuracy (68%). In contrast, for "serious injury" accidents, GPT-3.5 and LLaMA3-8B in the plain zero-shot setting (ZS), as well as GPT-3.5 in the zero-shot with Chain-of-Thought scenario (ZS_CoT), achieved 100% accuracy. However, it is crucial to note that in these settings, these models performed poorly for fatal and "minor or non-injury" accidents, indicating an inherent bias toward the intermediate severity category of "serious injury".

Interestingly, LLaMA3-70B with the basic zero-shot approach demonstrated the best inference performance for "minor or non-injury" accidents (58% accuracy) while maintaining a relatively balanced performance across fatal and serious injury accidents. This suggests a robust generalization capability of LLaMA3-70B across crash severity categories.

The implementation of prompt engineering, particularly in zero-shot settings, generally enhanced performance across models. This improvement was especially pronounced for fatal accident classification, where rephrasing the "Fatal accidents" label to the soft

version of "Serious accident with potentially fatal outcomes" facilitated maintenance of classification accuracy while adhering to LLMs' aligned behaviors.

These results underscore the complexity inherent in crash severity inference task, as no single approach consistently outperformed others across all metrics and severity categories. These findings highlight the need for careful selection of models and methodologies based on specific task requirements and the emphasis of severity categories in the application context.

5.3. Effectiveness of Prompt Engineering (PE) and Chain-of-Thought (CoT)

Figure 8 demonstrated the performance gains by CoT and PE as compared to the plain zero-shot setting. Both ZS_CoT and ZS_PE consistently demonstrate enhanced performance in terms of Macro F1-score and Macro-accuracy across all three models evaluated. This improvement underscores the efficacy of CoT and PE in boosting model performance in zero-shot scenarios.
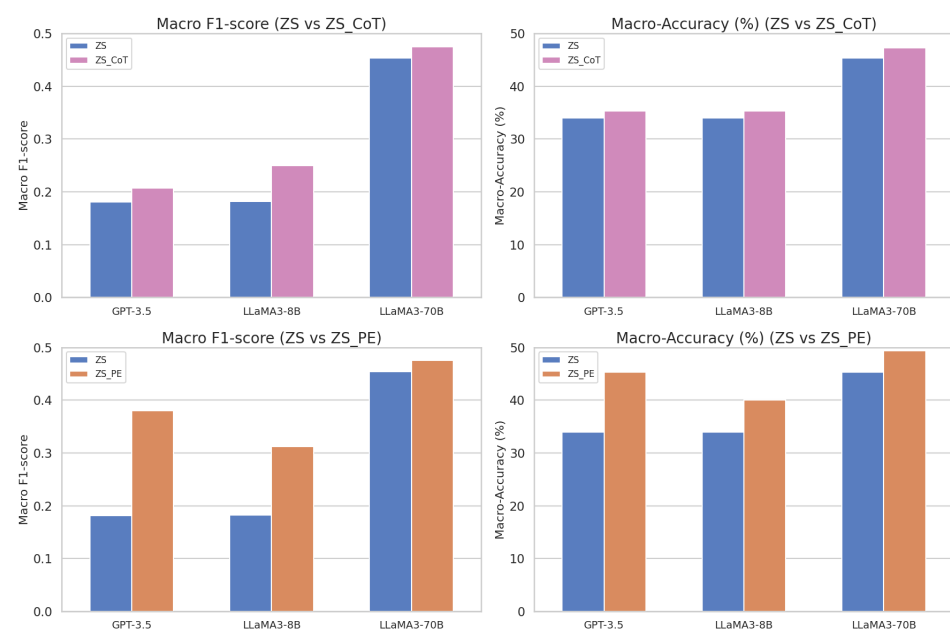
Figure 8. Effect of PE or CoT separately.

Notably, the implementation of PE (ZS_PE) yields more substantial improvements relative to CoT (ZS_CoT). This differential in enhancement suggests that, within the context of this specific task, the reformulation of prompts may be particularly effective in guiding model outputs as compared to the structured reasoning approach with CoT. The consistent pattern of improvement across different model architectures and sizes indicates the broad applicability of these techniques in zero-shot learning paradigms.

As illustrated in Figure 8, CoT improves both Macro F1-score and Macro-Accuracy across all three models in the plain zero-shot (ZS) setting. Based on the results summarized in Table 3, GPT-3.5 and LLaMA3-8B show improved recognition of "Minor or non-injury" accidents. LLaMA3-70B demonstrates substantial gain in identifying "Serious injury" accidents and "Minor or non-injury" accidents, with only a slight reduction in performance for "Fatal" accidents. The use of CoT enables LLMs to better understand and reason through questions, leading to more reliable and explainable inferences.

The PE technique also leads to increased Macro F1-score and Macro-Accuracy across all three models compared to the plain zero-shot (ZS) setting, as depicted in Figure 8. Notably, it greatly enhances the models' ability to detect Fatal accidents by simply softening the label description from "Fatal accident" to "Serious accident with potentially fatal outcomes", resulting in more balanced performance across severity categories. Compared to the zero-shot baseline, GPT-3.5 with PE attains a remarkable increase in Fatal accident

detection, with accuracy rising from 0% to 62%. Similarly, LLaMA3-8B and LLaMA3-70B show increases in fatal accident accuracy from 0% to 34%, 44% to 60%, respectively.

These improvements may stem from the fact that PE directs LLMs to concentrate more specifically on accident severity classification, potentially addressing any initial tendency to be overly cautious or generalized in their responses. This targeted guidance enables the models to make more precise distinctions among accident severity categories.

Figure 9 shows the comparative performance of the three models across incremental settings from ZS to ZS_PE and ZS_PE_CoT.
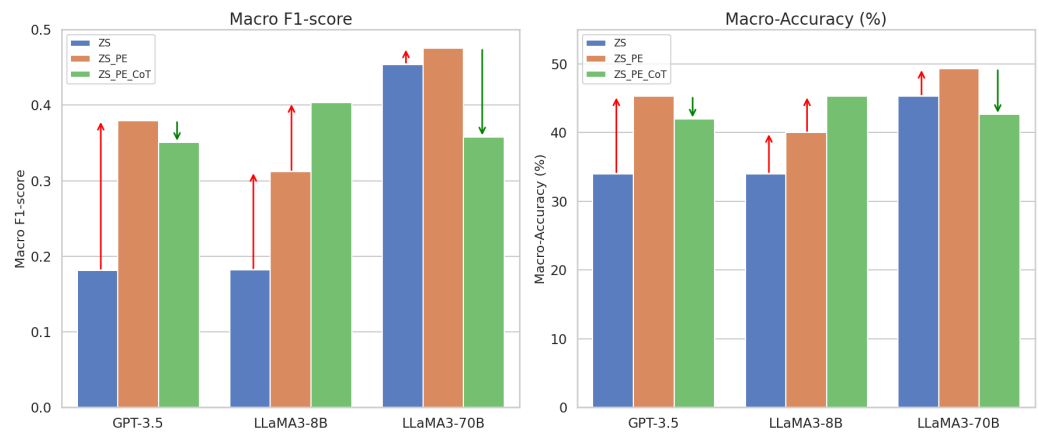
Figure 9. Performance comparison of models in ZS, ZS_PE, and ZS_PE_CoT.

In both the zero-shot (ZS) and zero-shot with PE (ZS_PE) settings, LLaMA3-70B consistently outperforms the other two models. In the ZS setting, LLaMA3-70B achieves a macro F1-score of 0.4541 and a macro-accuracy of 45.33%, significantly higher than both GPT-3.5 (0.1812 and 34.00%) and LLaMA3-8B (0.1818 and 34.00%). This performance advantage is maintained in the ZS_PE setting, where LLaMA3-70B shows further improvement with a macro F1-score of 0.4755 and a macro-accuracy of 49.33%, compared to GPT-3.5 (0.3798 and 45.33%) and LLaMA3-8B (0.3120 and 40.00%).

However, the performance dynamics shift in the setting of zero-shot with PE and CoT (ZS_PE_CoT). In this setting, LLaMA3-8B leads the performance with a macro F1-score of 0.4033 and a macro-accuracy of 45.33%, surpassing both GPT-3.5 (0.3509 and 42.00%) and LLaMA3-70B (0.3581 and 42.67%). GPT-3.5 and LLaMA3-70B experience slightly decreased performance under the ZS_PE_CoT setting as compared to the ZS_PE setting. In contrast, LLaMA3-8B show improved macro-F1-score from 0.3120 to 0.4033 as well as increased macro-accuracy from 0.4000 to 0.4533. This shift in performance suggests that the combination of PE and CoT reasoning are particularly more beneficial to small models, such as LLaMA3-8B, than large models like GPT-3.5 and LLaMA3-70B for the crash severity inference task.

In the ZS_PE_CoT setting, all three models, GPT-3.5, LLaMA3-8B, and LLaMA3-70B, demonstrate improved recognition of fatal accidents, evidenced in Table 3. This enhancement indicates that the combination of CoT and PE is particularly beneficial for identifying more severe crashes than less severe ones.

5.4. Zero-Shot vs. Few-Shot Learning

In the FS setting, the inclusion of three examples improves both the macro F1-score and macro-accuracy compared to the ZS setting, boosting classification accuracy of GPT-3.5 and LLaMA3-8B for "Fatal accident" and "Minor or non-injury accident". However, this comes at the expense of accuracy for "Serious injury accident". This indicates a potential trade-off in classification performance across severity categories or a decrease of the model bias in the ZS setting.

Moreover, smaller models like LLaMA3-8B generally benefit more from few-shot learning than larger models, such as GPT-3.5 and LLaMA3-70B, as evidenced by a notable

increase in Macro F1 score from 0.1818 to 0.4068. Nevertheless, LLaMA3-70B, being a larger model, performs slightly better in the zero-shot settings, suggesting it may have gained some general knowledge in traffic safety domain during the pre-training stage, where the zero-shot prompting can draw upon such knowledge.

In contrast, the effects of PE in the FS setting exhibit more variations across models. GPT-3.5 demonstrates improvements in Macro F1-score and Fatal accident accuracy, and LLaMA3-70B shows remarkably improved inference accuracy for "fatal accident" and "Serious injury accident". Conversely, LLaMA3-8B shows decreased macro F1-score and macro-accuracy, indicating that PEin the few-shot setting may not be equally beneficial for models of different sizes.

It important to note that we did not explore the aspect regarding the choice of the examples in the FS setting, which might have a varying effect on different models.

6. Discussions

In this section, we focus on the reasoning abilities of LLMs with CoT by analyzing their responses using word clouds. Additionally, We highlight the limitations of the current work and suggest directions for future reseearch.

6.1. Can LLMs with CoT Yield Logical Reasoning for Their Inference Outcomes?

CoT is a unique technique to augment LLM's reasoning capability. But how reasonable is the reasoning by LLMs with CoT? We aim to examine the responses of LLMs with CoT from the view of a traffic safety engineer. Specifically, the responses of LLaMA3-70B in the ZS_CoT setting are evaluated due to its best performance in this setting (see Table 3).

As a qualitative assessment, three word clouds are drawn separately with respect to the three severity categories, i.e., "Minor/non-injury", "Serious injury", and "Fatal" accidents (see Figures 10–12). Note that only the correct inferred responses are used in creating these word clouds, where the bigger sizes of words indicate their higher frequencies in the LLMs' responses. These visualizations offer insights into the LLM's conceptualization and reasoning processes regarding accident causation and factors considered during the severity inference.

In the three word clouds (Figures 10–12), some words consistently appear regardless of severity outcomes, including "collision", "intersection", "vehicle", and "driver". This suggests a core set of concepts that the LLM associates with traffic accidents. Additionally, LLaMA3-70B demonstrates consideration of diverse factors in its accident analysis, including:

- Crash-related factors (e.g., "rear-end collision", "pedestrian", "opposite directions", "corner")
- Environmental conditions (e.g., "wet road surface", "rain", "dark", "stop-go")
- Driver behavior (e.g., "failing to yield", "misjudgment", "turning", "give way", "excessive speed")
- Driver characteristics (e.g., "male" "older driver", "age")
- Vehicle factors (e.g., "bus", "headlights", "seatbelt")
- Road design elements (e.g., "traffic lights", "intersection", "curved road", "t intersection")

The word cloud for "Minor or non-injury" accidents (Figure 10), is characterized by distinct terms suggesting lower impact and preventative measures, such as "Low speed" and "slow", indicating the LLM associates lower velocities with less severe outcomes. "Seatbelt" and "failed" suggest a focus on minor infractions and safety equipment. "Stop" and "time" may relate to issues at intersections or reaction times. "Rear-end collision", and "corner" also are prominent, indicating that LLM associates rear-end collision or accident impact on corners of vehicles with less severe outcomes.
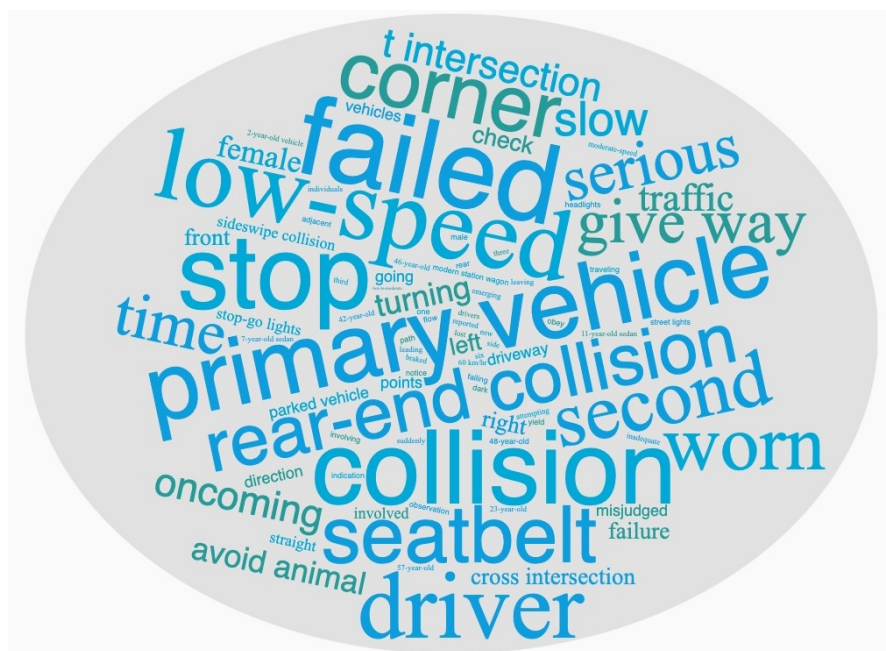
Figure 10. Word cloud for correctly inferred "Minor or non-injury accident" in the ZS_CoT setting.
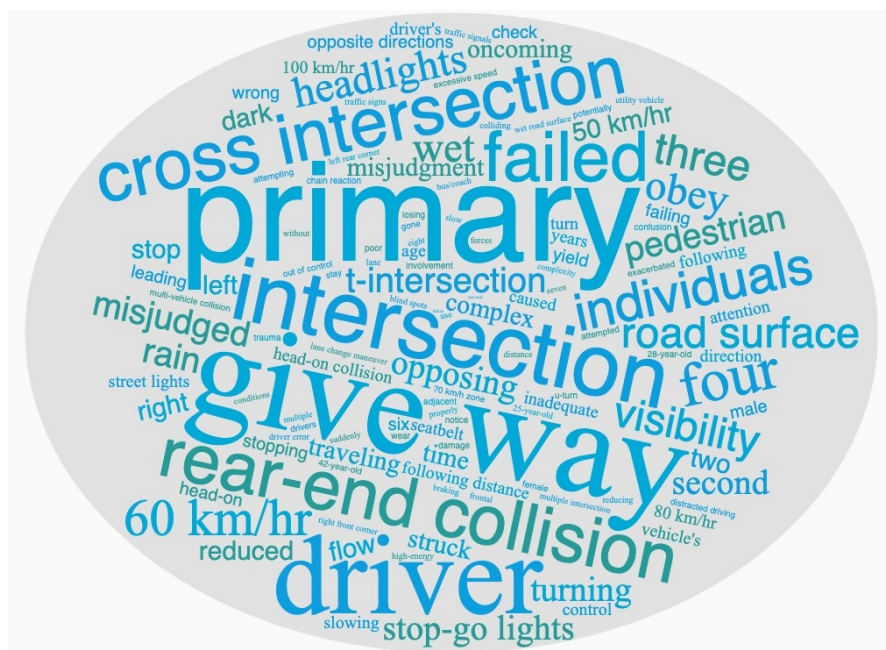

Figure 11. Word cloud for correctly inferred "Serious injury accident" in the ZS_CoT setting.

The word cloud for "Serious injury accident", shown in Figure 11, is characterized by terms suggesting moderate impact, such as those relating to drivers (e.g., "give way", "misjudged") and locations (e.g., "intersection", "cross intersection"). "Wet", "rain", and "visibility" emerge, highlighting adverse weather or environmental conditions. "60 km/h" and "50 km/h" reveals relatively high speeds are associated with more severe outcomes. "Four" and "three" appear, indicating potentially more vehicles or individuals involved in an accident with severe outcomes. "Rear-end collision" appears again, which combines with "intersections" and the higher speeds (e.g., "60 km/h"), could lead to "Serious injury" accidents.
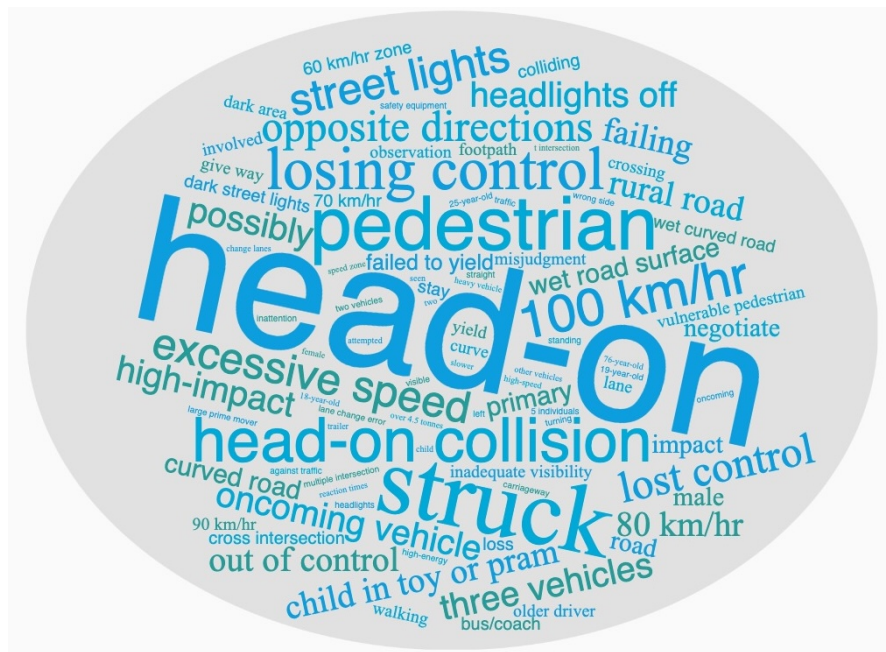
Figure 12. Word cloud for correctly inferred "Fatal accident" in the ZS_CoT setting.

The word cloud for "Fatal accident" reveals a marked shift towards high-energy impact scenarios (see Figure 12). "Head-on" and "high-impact" are among the most prominent terms. Specifically, higher speeds and speeding (e.g., "100 km/h", "excessive speed") are associated with fatal accidents. "Pedestrian" gains significance, reflecting heightened vulnerability. "Losing control" and "out of control" suggest more catastrophic situations or driver errors. "Rural road", "curved road" are also important terms identified in the word cloud. Furthermore, three distinct examples are provided in Figure 13 to illustrate the CoT's reasoning process.

Example 1:
Based on the provided information, a potential reason for the crash is a head-on collision between two vehicles traveling in opposite directions, likely due to one of the drivers failing to stay in their lane or losing control of their vehicle.

I would classify the severity of this crash as a 'Fatal accident'. The fact that it was a head-on collision at a speed zone of 100 km/hr, with no control in place, suggests a high-impact crash with a high likelihood of severe injuries or fatalities. Additionally, the fact that the accident occurred on a rural road with no street lights and at evening time may have reduced visibility, increasing the severity of the crash.

Example 2:
Based on the provided information, the potential reason for the crash is that the pedestrian, who was likely crossing the road, was not visible to the driver due to the dark street lights being off and the vehicle's headlights also being off. The driver, who was attempting to go straight ahead, failed to detect the pedestrian in time, resulting in the collision.

I would classify the severity of this crash as a 'Fatal accident'. The fact that the pedestrian was struck by the vehicle and the lack of visibility due to the dark conditions suggest a high likelihood of severe injuries or fatalities. Additionally, the speed zone of 80 km/hr and the lack of control measures in place at the intersection may have contributed to the severity of the crash.

Example 3:
Based on the provided information, the potential reason for the crash is a head-on collision between two vehicles traveling in opposite directions, likely due to one of the drivers failing to negotiate a curved section of the road or losing control of their vehicle.

I would classify the severity of this crash as a 'Fatal accident'. The fact that it was a head-on collision, which is one of the most severe types of crashes, and involved a heavy vehicle (rigid) over 4.5 tons, suggests a high likelihood of severe injuries or fatalities. Additionally, the crash occurred at a speed zone of 60 km/hr, which is a relatively high speed, increasing the risk of severe consequences.

Figure 13. Output examples for fatal accidents from LLaMA3-70B in ZS_CoT setting.

6.2. Limitations and Future Research

One of the primary limitations of this study is the relatively small sample used, including only 150 instances (50 for each of the three severity categories). This limited dataset may not fully capture the variability and complexity of diverse real-world crash scenarios, potentially affecting the generalizability of the findings.

For future research, several directions can be explored to further enhance the performance and applicability of LLMs in crash analysis and modeling:

- Expanding the dataset to include a larger and more diverse set of samples will allow for a more comprehensive evaluation of the models' capabilities and improve the robustness of the results.
- Fine-tuning LLMs with more extensive and domain-specific data (e.g., crash reports and databases) can significantly enhance their domain knowledge to better understand the nuances and specificities of traffic accidents, leading to more accurate and reliable reasoning and inference.
- Investigating explanation methods in conjunction with LLMs can yield more interpretable and trustworthy results.

7. Conclusions

In conclusion, this study demonstrates the efficacy of LLMs in crash severity reasoning and inference using textual narratives of crash events constructed from structured tabular data. Our comprehensive evaluation of modern LLMs (GPT-3.5-turbo, LLaMA3-8B, and LLaMA3-70B) across different settings (zero-shot, few-shot, CoT, and PE) yields insightful findings. LLaMA3-70B consistently outperformed other models, especially in zero-shot settings. CoT and PE techniques lead to enhanced performance, improving logical reasoning and addressing alignment issues.

Notably, the use of CoT provided valuable insights into LLM reasoning processes, revealing their capacity to consider multiple factors such as environmental conditions, driver behavior, and vehicle characteristics in the crash severity inference task. These findings collectively suggest that LLMs hold considerable promise for crash analysis and modeling. Future research may explore other safety applications beyond the severity analysis and inference.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, J.J.Y., H.Z. and N.L.; data curation, H.Z. and Y.S.; methodology, H.Z., Y.S. and J.J.Y.; software, H.Z.; formal analysis, H.Z., Y.S. and J.J.Y.; writing—original draft preparation, H.Z., Y.S. and Y.H.; writing—review and editing, J.J.Y. and N.L.; visualization, H.Z.; supervision, J.J.Y. and N.L. All authors reviewed the results and approved the final version of the manuscript.

References

1. Mannering, F.; Bhat, C.R.; Shankar, V.; Abdel-Aty, M. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Anal. Methods Accid. Res.* **2020**, *25*, 100113. [CrossRef]
2. Golob, T.F.; Recker, W.W. Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. *J. Transp. Eng.* **2003**, *129*, 342–353. [CrossRef]
3. Eluru, N.; Bhat, C.R. A joint econometric analysis of seat belt use and crash-related injury severity. *Accid. Anal. Prev.* **2007**, *39*, 1037–1049. [CrossRef] [PubMed]
4. Lord, D.; Mannering, F. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transp. Res. Part Policy Pract.* **2010**, *44*, 291–305. [CrossRef]

5. Savolainen, P.T.; Mannering, F.L.; Lord, D.; Quddus, M.A. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accid. Anal. Prev.* **2011**, *43*, 1666–1676. [CrossRef] [PubMed]
6. Karlaftis, M.G.; Golias, I. Effects of road geometry and traffic volumes on rural roadway accident rates. *Accid. Anal. Prev.* **2002**, *34*, 357–365. [CrossRef] [PubMed]
7. Zhang, Z.; Akinci, B.; Qian, S. Inferring heterogeneous treatment effects of work zones on crashes. *Accid. Anal. Prev.* **2022**, *177*, 106811. [CrossRef] [PubMed]
8. Pervez, A.; Lee, J.; Huang, H. Exploring factors affecting the injury severity of freeway tunnel crashes: A random parameters approach with heterogeneity in means and variances. *Accid. Anal. Prev.* **2022**, *178*, 106835. [CrossRef] [PubMed]
9. Goh, Y.M.; Ubeynarayana, C. Construction accident narrative classification: An evaluation of text mining techniques. *Accid. Anal. Prev.* **2017**, *108*, 122–130. [CrossRef] [PubMed]
10. Zhang, X.; Green, E.; Chen, M.; Souleyrette, R.R. Identifying secondary crashes using text mining techniques. *J. Transp. Saf. Secur.* **2020**, *12*, 1338–1358. [CrossRef]
11. Das, S.; Datta, S.; Zubaidi, H.A.; Obaid, I.A. Applying interpretable machine learning to classify tree and utility pole related crash injury types. *IATSS Res.* **2021**, *45*, 310–316. [CrossRef]
12. OpenAI. GPT-3.5 Turbo Updates. 2023. Available online: https://platform.openai.com/docs/models/gpt-3-5-turbo (accessed on 14 June 2024).
13. AI@Meta. Llama 3 Model Card. 2024. Available online: https://www.meta.ai (accessed on 14 June 2024).
14. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
15. Wu, X.; Yao, W.; Chen, J.; Pan, X.; Wang, X.; Liu, N.; Yu, D. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning. *arXiv* **2023**, arXiv:2310.00492.
16. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
17. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
18. Yin, W. Meta-learning for few-shot natural language processing: A survey. *arXiv* **2020**, arXiv:2007.09604.
19. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large language models are zero-shot reasoners. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 22199–22213.
20. Qin, C.; Zhang, A.; Zhang, Z.; Chen, J.; Yasunaga, M.; Yang, D. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? In Proceedings of the The 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023.
21. Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; Smola, A. Multimodal Chain-of-Thought reasoning in language models. *arXiv* **2023**, arXiv:2302.00923.
22. Lyu, Q.; Havaldar, S.; Stein, A.; Zhang, L.; Rao, D.; Wong, E.; Apidianaki, M.; Callison-Burch, C. Faithful Chain-of-Thought Reasoning. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023), Bali, Indonesia, 1–4 November 2023; Association for Computational Linguistics: Bali, Indonesia, 2023.
23. Wang, Y.; Zhong, W.; Li, L.; Mi, F.; Zeng, X.; Huang, W.; Shang, L.; Jiang, X.; Liu, Q. Aligning large language models with human: A survey. *arXiv* **2023**, arXiv:2307.12966.
24. Shen, T.; Jin, R.; Huang, Y.; Liu, C.; Dong, W.; Guo, Z.; Wu, X.; Liu, Y.; Xiong, D. Large language model alignment: A survey. *arXiv* **2023**, arXiv:2309.15025.