nature communications

Article

https://doi.org/10.1038/s41467-025-64574-w

SafeTraffic Copilot: adapting large language models for trustworthy traffic safety assessments and decision interventions

Received: 17 March 2025

Accepted: 17 September 2025

Published online: 07 October 2025

Check for updates

Yang Zhao[1,2,5], Pu Wang[1,2,5], Yibo Zhao[1], Hongru Du[1,2,3] & Hao Frank Yang[1,2,4]

Predicting expected traffic crashes and designing targeted interventions are highly challenging due to the inherent complexity of crash data and persistent concerns over the prediction trustworthiness. We introduce *SafeTraffic Copilot* that adapts Large Language Models (LLMs) to perform expected crash prediction as a text-reasoning task, then attribute critical features for targeted safety interventions. Within the *Copilot*, *SafeTraffic LLM* is customized then fine-tuned on the textualized *SafeTraffic Event* dataset, which consists of 66,205 real-world crash cases with 14.5 million words from five U.S. states. Across multiple prediction tasks including crash type, severity, and number of injuries, *SafeTraffic LLM* demonstrates a 33.3% to 45.8% improvement in average F1-score over existing works. To interpret these results and inform safety interventions, we introduce *SafeTraffic Attribution*, a sentence-level feature-attribution framework enabling conditional "what-if" risk analysis. Findings reveal that alcohol-impaired driving is the leading factor for severe crashes, with impairment-related and aggressive behaviors contributing nearly three times more risk than other behaviors. Furthermore, *SafeTraffic Attribution* identifies critical features during fine-tuning, guiding crash data collection strategies for continual improvement. *SafeTraffic Copilot* enables prediction and reasoning of conditional crash risks through foundation models, thereby supporting traffic safety improvements and offering clear advantages in generalization, adaptation, and trustworthiness.

Road traffic injury remains a stubborn public health crisis in the United States. In 2022 alone, 42,795 people were killed on U.S. roads—one of the highest per capita fatality rates in the developed world[1]. Despite decades of counter-measures, the fatality curve continues to rise, especially in the United States (as shown in Fig. 1a), underscoring an urgent need for new data-driven techniques that can uncover the mechanisms of crashes and inform decisive policy action. Expected crash prediction models (hereafter referred to as crash prediction) offer a principled way to learn from historical data and isolate the factors that most strongly elevate risk[2].

The current approaches to crash prediction are broadly categorized into macroscopic, statistical-level analyses, and microscopic, event-level investigations. Macroscopic models offer a general understanding of safety performance, identifying high-risk areas and

[1]Department of Civil and Systems Engineering, Johns Hopkins University, Baltimore, MD, USA. [2]Center for Systems Science and Engineering, Johns Hopkins University, Baltimore, MD, USA. [3]Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA. [4]Johns Hopkins Data Science and AI Institute, Johns Hopkins University, Baltimore, MD, USA. [5]These authors contributed equally: Yang Zhao, Pu Wang.
e-mail: haofrankyang@jhu.edu

https://doi.org/10.1038/s41467-025-64574-w

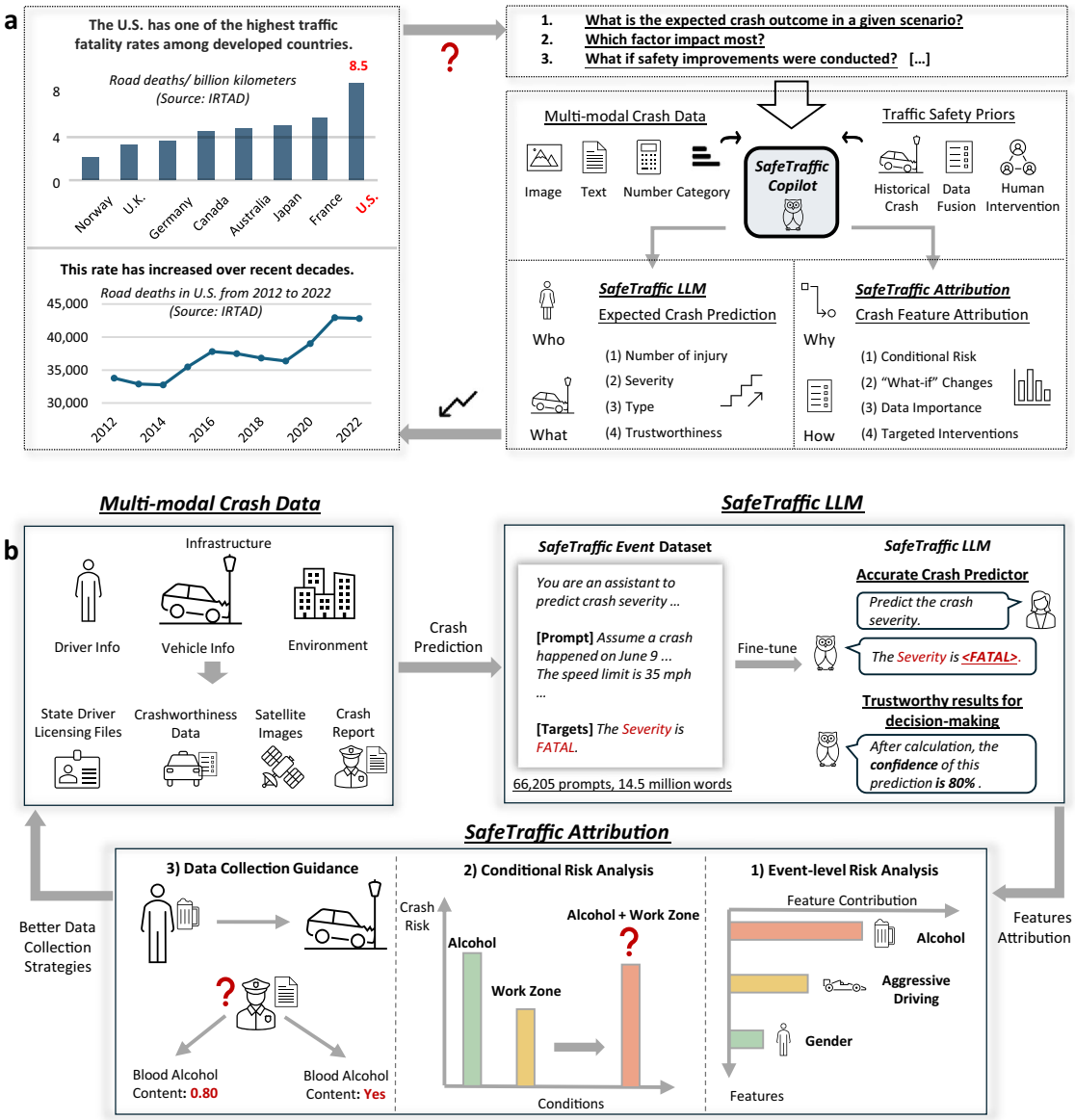


Fig. 1 | Overview of the proposed *SafeTraffic Copilot*. **a** The U.S. faces one of the highest crash risks among developed countries, with a rising trend. However, analyzing and addressing this issue is challenging due to the heterogeneous factors involved in crash events, including traffic conditions, human behavior, environmental impacts, and driver characteristics. To tackle this, we propose *SafeTraffic Copilot*, a framework designed for two key tasks: (1) Predicting crash outcomes and (2) Attributing crash factors for conditional risk analysis. By addressing questions such as why crashes occur and how to mitigate crash risks, *SafeTraffic Copilot* seeks to deliver optimal policies for safety improvement. **b** The *SafeTraffic Copilot* workflow incorporates multi-modal data, including driver behavior, vehicle details, infrastructure, and environmental conditions, represented through textual reports, satellite imagery, and other formats. Leveraging an AI-expert cooperative method, the crash data is transformed into textual prompts, resulting in the *SafeTraffic Event* dataset comprising 66,205 cases. *SafeTraffic LLM* is created with accurate and trustworthy forecasting abilities for further analysis. Building on this pipeline, *SafeTraffic Attribution* operates across three dimensions: (1) Event-level risk analysis to identify feature contributions, (2) Conditional risk analysis to assess state-level risks under varying conditions, and (3) Data collection guidance to optimize the data acquisition process. The results of *SafeTraffic Attribution* provide actionable insights to enhance data analysis and collection, fostering a more comprehensive understanding of crash data and events.

temporal trends, but they lack the granularity to explain the specific circumstances of a crash—the who, what, and why[3–6]. While microscopic models, often employing machine learning, aim to predict crash consequences under specific traffic conditions, they have struggled with precision and generalization[4,7–10]. A fundamental challenge lies in effectively integrating the multi-modal data associated with a crash event, spanning textual narratives, numerical data, images, and driver histories, and interpreting the complex interplay of contributing factors, thereby limiting their utility in designing effective safety policies.

The recent emergence of foundation models, especially Large Language Models (LLMs), presents a transformative opportunity to mitigate these enduring challenges by leveraging their advanced capabilities in processing and reasoning from complex, multi-modal information[11–15]. These models can synthesize and interpret vast, unstructured data, such as the narrative descriptions in crash reports, and align them with structured data like roadway characteristics and driver histories, offering a more holistic understanding than was previously possible. However, adapting these powerful generative models for the discriminative task of crash outcome prediction introduces its own set of technical hurdles. The primary challenge is methodological: generative LLMs with extensive output vocabularies must be re-engineered to reliably predict outcomes within a set of well-defined,

finite categories (e.g., crash severity levels)[16]. This adaptation raises major concerns about the trustworthiness and calibration of their predictions, which is crucial for high-stakes applications like public safety[17]. Furthermore, the inherent "black box" nature of these models poses a major obstacle to achieving the interpretability required for targeted safety improvements[18–20]. While initial studies have explored LLMs for traffic safety, they have been limited to prompt engineering and have not addressed the crucial interpretability gap, which is essential for robust decision support and for answering the critical why and how questions of crash causation[21–23].

In this research, we introduce *SafeTraffic Copilot*, a LLM-driven framework that shifts the paradigm from aggregate-level statistics to granular, event-level crash prediction and understanding (see Fig. 1b). By reframing crash prediction as a text-based reasoning task, *SafeTraffic Copilot* is designed to address the key challenges of data integration, model generalization, and feature attribution. The framework consists of three integrated components: *SafeTraffic Event* dataset, for unifying multi-modal crash data; *SafeTraffic LLM*, for accurate outcome prediction; and *SafeTraffic Attribution*, for conditional risk analysis. This approach allows us to not only forecast the when, where, who, and what of a crash but also to provide a deep, interpretable understanding of why it occurred and how similar risks can be mitigated, offering a unified approach for targeted and effective data-driven safety interventions.

Results

This study shifts traffic safety analysis from aggregate-level to event-level crash prediction by developing *SafeTraffic Copilot*, a customized LLM framework that integrates multi-modal crash data into the broader semantic context to forecast consequences and attribute features with interpretability. Using *SafeTraffic Event* dataset (66,205 textual prompts; over 14 million words) and framing outcome prediction as token generation, *SafeTraffic LLM* delivers a 33.3% to 45.8% average F1 improvement over competitive baselines across multiple crash-consequence tasks. By embedding traffic-safety priors and explicitly targeting the number of injuries, crash severity, and crash type via special tokens, *SafeTraffic Copilot* yields accurate and trustworthy predictions, where the accuracy increases with confidence, achieving over 70% accuracy when the confidence score exceeds 60%, and 95% precision for fatal-crash predictions at the same threshold. Our proposed textual feature-attribution module provides event- and state-level insight—it simultaneously uncovers what drives risk in a specific crash, enabling conditional intervention, and what information drives model quality at scale, guiding strategic data-collection policies for long-term accuracy. Specifically, the proposed sentence-level Shapley scores identify high-risk scenarios (such as "alcohol-impaired," "work-zone," "inappropriate behaviors," etc.) for actionable "what-if" analysis. For example, combining alcohol impairment with a work-zone setting nearly doubles the likelihood of a severe crash compared with sober driving under identical conditions. Further, summing Shapley contributions across the entire fine-tuning dataset ranks data fields by their marginal impact on prediction accuracy and confidence score, thereby guiding first-responders toward the details that matter and streamlining continuous model updates. For example, unit information (driver and vehicle attributes) accounts for more than 40% of the contribution to crash severity prediction, highlighting these fields as priority entries for future crash documentation.

Conditional expected crash prediction

In our research, we define crash prediction as expected (conditional) crash prediction, a definition consistent with the agencies and established literature[4,24–26]. Officially, expected crash prediction is defined as the task of estimating expected crash characteristics, including crash type, severity, number of injuries, and their likelihood of occurrence, under specified conditions. These estimations are based on expected traffic conditions and relevant contextual information, such as roadway attributes, environmental conditions, traffic volumes, and driver behaviors.

The prediction targets consist of three variables with belonged confidence score: *Number of Injury*, *Severity*, and *Crash Type* (see Fig. 2)[27,28]. Specifically, *Number of Injury* task predicts the number of people injured in the given crash event. We define the number of injuries as the total number of non-fatal injuries, including possible, minor, and serious injuries, obtained by subtracting fatalities from the total number of injured people in a crash. *Number of Injury* task is treated as a classification task with four categories: *zero*, *one*, *two*, and *three or more than three*, where crashes involving more than two injured people are grouped into a single category due to the limited number of such cases. The *Severity* task assesses the level of injury severity in a crash, classified into five levels from *no apparent injury* to *fatal*. *Type* task predicts the type of crash, such as the *rear-end collision* or *collision with object*, with 14 crash type categories in the Washington dataset and 16 in the Illinois dataset. Detailed information on the defined targets and confidence scores is available in "SafeTraffic Event dataset construction" and "Expected crash prediction confidence score calculation" in the "Methods" section.

Original multi-modal crash data

Our cleaned dataset comprises crash data from Washington State in 2022, totaling 16,188 records, and from Illinois in 2022, totaling 42,715 records, after excluding cases with missing key attributes related to vehicle or crash object status. Primary sources include the Highway Safety Information System (HSIS) crash data[29], the state crash report, and satellite images[30]. HSIS is a multistate database that contains crash, roadway inventory, and traffic volume data for a select group of States. The HSIS crash data contains four major components: crash data, infrastructure data, vehicle data, and person data. Crash data provides detailed descriptions of crashes, such as location, time, and injury severity. Infrastructure data includes information about road layouts and traffic characteristics, such as road level and speed limits. Vehicle data contains details such as manufacturing year and reported defects of the involved vehicles, while person data captures demographic and other relevant details about drivers and passengers, such as age and gender. Satellite images complement the HSIS data by providing additional visual context, including information about lanes, intersections, and other roadway attributes. Further information on raw data formats and types is available in "Raw data" in the "Methods" section.

In addition to the Washington and Illinois datasets, we also collect and process 2250 crash cases from Maine, 2250 from Ohio, and 2802 from North Carolina to evaluate the model's zero-shot generalization. Their raw data sources align with those used for Illinois and Washington. Using the Illinois State prompt template (due to its closer data format), we converted the records into prompt format, keeping shared features and setting unmatched ones to null. Differences in data representation (e.g., driver alcohol levels as a numeric value in Illinois versus a textual descriptor in Maine) introduce unseen values and distributions, enabling evaluation of the model's zero-shot generalization.

Developing *SafeTraffic LLM* for predicting crashes

To leverage the multi-modal crash data described above for crash prediction, we developed the *SafeTraffic Copilot* crash outcomes prediction pipeline, which transforms crash outcomes prediction into a text-based reasoning task. To achieve this, the raw crash data is organized into the textual *SafeTraffic Event* dataset, which is then used to fine-tune the *SafeTraffic LLM*. Figure 2 presents an overview of the proposed pipeline.

The *SafeTraffic Event* dataset is created through an AI-expert cooperative textualization process, organizing multi-modal raw data

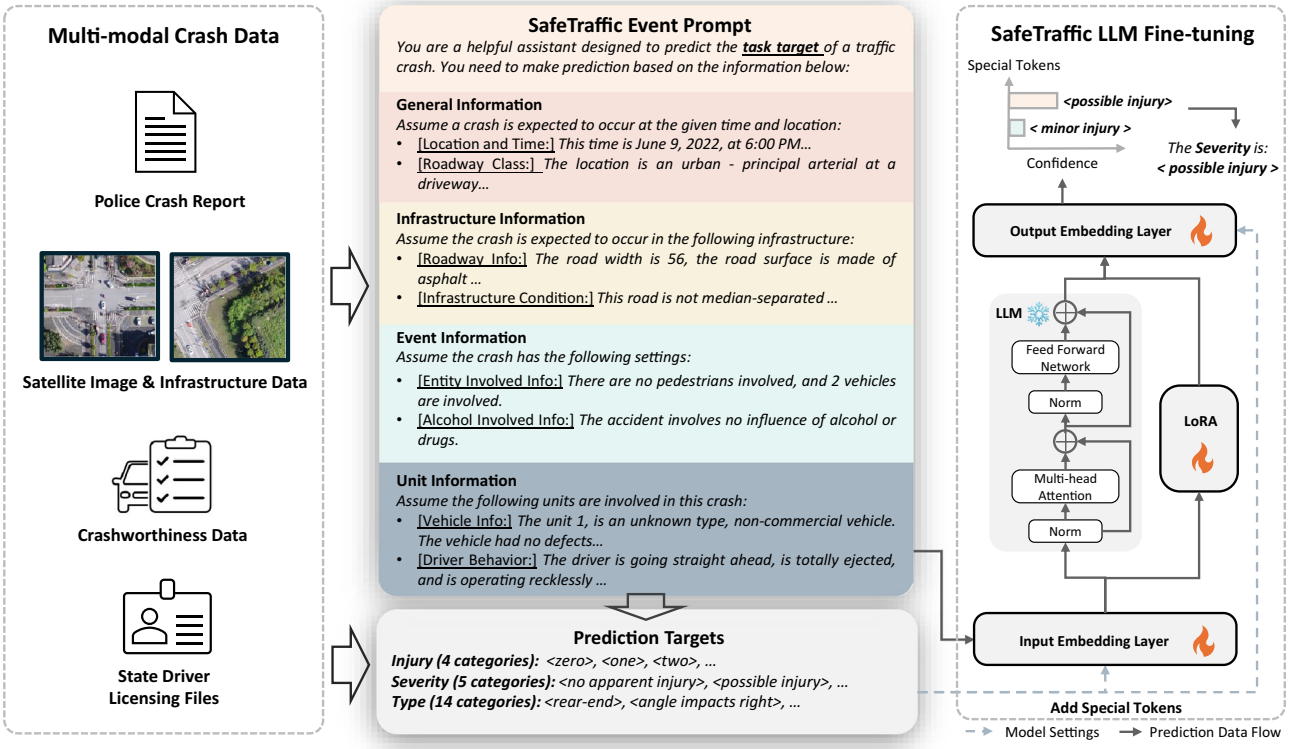


Fig. 2 | *SafeTraffic Copilot* crash outcomes prediction pipeline. Multi-modal crash data is collected and organized into textual prompts through an AI-expert cooperative process. The Highway Safety Information System (HSIS) crash data, satellite images, and infrastructure data are used to extract general and infrastructure information, including the crash time, location, the road level, and so on. The vehicle data and person data are converted into the event information and the unit information, including vehicle movements, driver characteristics (e.g., age, gender, alcohol use), vehicle attributes (e.g., manufacture year), and so on. *SafeTraffic Event* dataset is created with three prediction targets: *Number of Injury*, *Severity*, and *Type*. The *Number of Injury* task predicts the number of people injured in the crash event, the *Severity* task estimates the severity level of the crash, such as *no apparent injury* or *fatal*, and the *Type* task classifies type of crash, such as *single vehicle with object* or *angle impacts right* (The crash event outcomes classification are provided in Supplementary Table 4 and Supplementary Table 5). The *SafeTraffic LLM* is fine-tuned using the *SafeTraffic Event* dataset. To reframe the crash outcomes prediction from a classification task to a language inference task, *SafeTraffic LLM* is fine-tuned by adding prediction targets as special tokens in its vocabulary and adjusting parameters using Low-Rank Adaptations (LoRA)[37], a lightweight fine-tuning technique that injects trainable rank-decomposed matrices into each layer without updating the full model.

for effective crash prediction. The detailed information about the raw data feature engineering and the textualization process is available in "SafeTraffic Event dataset construction" in the "Methods" section. As shown in Fig. 2, the constructed prompts are divided into five parts: one system prompt and four content parts, with each content part containing approximately 100 words. These parts include:

- System prompt: provides an introduction and task-specific instructions.
- General information: includes general information about the time and location of the prediction region and the roadway category.
- Infrastructure information: describes road infrastructure, encompassing static features like the number of lanes and speed limits, as well as dynamic elements such as work zones, lighting, and road surface conditions.
- Event information: contains detailed descriptions of crash events, such as the number of vehicles involved and their directions of movement.
- Unit information: provides vehicle and individual details relevant for crash prediction, such as airbag status and the driver's age.

To organize and merge the mentioned information for each crash event, we perform feature engineering and textualization, structure the textualized data as input, and process labels corresponding to the three targeted crash prediction tasks from real-world reports. The complete prompt examples are presented in Fig. 3. Ultimately, after filtering out data items with missing information, the *SafeTraffic Event* dataset merges the complementary information from multi-modal data sources and contains 66,205 crash records with approximately 14.5 million words. These records are split into training, validation, and test sets in a 7:1.5:1.5 ratio for Washington and Illinois datasets, while the remaining datasets from three other states are used exclusively for cross-region training-free generalization evaluation.

With *SafeTraffic Event* dataset, we can adapt LLMs for expected crash prediction. Although vanilla LLMs like Llama 3[13] possess broad general knowledge and strong text-reasoning capabilities, they demonstrate limited effectiveness on crash prediction tasks without the fine-tuning process (see Supplementary Section 3.1). To address this, we developed *SafeTraffic LLM*, a specialized model fine-tuned on the processed *SafeTraffic Event* dataset. This fine-tuning process enhances the LLM's comprehension of crash events and enables accurate outcome prediction. Specifically, additional special tokens are introduced into the LLM vocabulary as prediction targets (*Number of Injury*, *Severity*, and Crash *Type*), fine-tuning the model to generate these tokens during prediction. The details of the fine-tuning are provided in "*SafeTraffic LLM*" in the "Methods" section.

Prediction performance and trustworthiness

We evaluate the performance of *SafeTraffic LLM* and compare its performance with other baselines (see "Adopted baselines" in the

Example Prompt - #EC22961

*You are a helpful assistant designed to predict the **task target** of a traffic crash. You need to make prediction based on the information below:*

General Information

This incident occurred on February 23, 2022, at 2:00 PM, in the city of Bremerton, Kitsap County, on the 303 route increasing milepost direction at milepost 1.87. The location is an Urban - Principal Arterial, not at an intersection and not related to a driveway. The type of roadway is classified as an Urban Multilane Undivided Non-Freeway. The level of access control is Non Limited Access Least Restrictive, the speed limit is 30, and the average annual daily traffic is 37000.

Infrastructure Information

The road width is 52 feet, the road surface is made of Asphalt, the right and left shoulders width is unknown, and the surface type of the left shoulder is unknown. This road does not have a median-separated area, there is no barrier in the median and the median width is unknown. The condition of the road is unknown regarding work zone status, but it is known that the accident occurred during daylight and the road surface condition was dry.

Event Information

There were no pedestrians involved, 3 vehicles involved. The accident has no influence of alcohol or drugs. There were no objects involved. Vehicle1 was moving North, in the direction of increasing milepost, Vehicle2 was also moving North, in the direction of increasing milepost. The first vehicle was moving straight when the second vehicle was stopped in traffic, legally standing.

Unit Information

The unit 1 is a Vanette Under 10,000 lb, non-commercial vehicle. The airbag was not deployed. The vehicle had no defects. The driver was going straight ahead, was not ejected, and was distracted by an unknown factor. Person 1: Motor Vehicle Driver, Female, 47, Restraint use is unknown. The unit 2 is a Vanette Under 10,000 lb, non-commercial vehicle. The airbag was not deployed. The vehicle had no defects. The driver had stopped for traffic, was not ejected, and no violations or factors contributed to the incident. Person 1: Motor Vehicle Driver, Female, 26, Restraint use is unknown. The unit 3 is a Vanette Under 10,000 lb, non-commercial vehicle. The airbag was not deployed. The vehicle had no defects. The driver was going straight ahead, was not ejected, and was distracted by an unknown factor. A drug recognition expert was not requested. Person 1: Motor Vehicle Driver, Female, 50, Restraint use is unknown.

Targets

Please predict the *Injury* number of the crash choosing from the following tokens (4 options available).
Assistant: <ZERO>

Please predict the *Severity* of the crash choosing from the following tokens (5 options available).
Assistant: <NO APPARENT INJURY>

Please predict the crash *Type* of the crash choosing from the following tokens (14 options available).
Assistant: <REAR END COLLISIONS>

Fig. 3 | Example prompt structure and content in the *SafeTraffic Event* dataset. An example prompt of an expected crash prediction event from the dataset collected in Washington State.

"Methods" section). The fine-tuning process is based on two vanilla LLMs with different sizes: Llama 3.1 8B and Llama 3.1 70B. Accuracy, precision, and F1-score are used as the evaluation metrics; the detailed information is available in "Evaluation metrics" in the "Methods" section.

SafeTraffic LLM provides the highest accurate and reliable prediction results across all crash types, severity, and injury counts, even in zero-shot scenarios. Table 1 compares the performances of *SafeTraffic LLM* and adopted baselines. The results show that the *SafeTraffic LLM* outperforms all the baselines in each task setting with an average F1-score improvement of 33.3%–45.8% across multiple tasks. *SafeTraffic LLM* performs well on both the Washington and Illinois datasets, demonstrating its stability across diverse geographical regions. Moreover, as shown in the confusion matrix in Fig. 4a, b, beyond aggregated metrics, *SafeTraffic LLM* demonstrates a more balanced prediction distribution and achieves higher accuracy across individual categories. In contrast, as shown in Fig. 4c, d, existing machine learning models tend to predict the dominant categories (e.g., *zero* under *Number of Injury* prediction task, *no apparent injury* under *Severity* prediction task, and the complete confusion matrix is shown in Supplementary Fig. 7).

SafeTraffic LLM provides trustworthy crash predictions, where a higher confidence score links to higher accuracy. *SafeTraffic LLM* tailors LLMs for discriminative crash outcomes prediction tasks, generating predictions accompanied by confidence scores that represent the probabilities associated with specific special tokens

https://doi.org/10.1038/s41467-025-64574-w

Table 1 | Performance comparison of the three expected crash prediction tasks

Dataset	Model	Number of injuries			Severity			Crash type			Rank (Avg.)
		Accuracy	Precision	F1-score	Accuracy	Precision	F1-score	Accuracy	Precision	F1-score	
Washington	RandomForest[42]	0.522	0.649	0.545	0.628	0.546	0.549	0.740	0.398	0.274	3 (4.44)
	AdaBoost[44]	0.495	0.245	0.328	0.492	0.245	0.328	0.563	0.249	0.302	10 (9.11)
	CatBoost[46]	0.495	0.245	0.328	0.492	0.245	0.328	0.715	0.400	0.329	7 (8.11)
	DecisionTree[43]	0.495	0.245	0.328	0.528	0.428	0.372	0.628	0.406	0.323	6 (6.67)
	LogisticRegression[45]	0.495	0.245	0.328	0.492	0.245	0.328	0.547	0.401	0.309	9 (8.67)
	XGBoost[41]	0.566	0.665	0.469	0.534	0.428	0.367	0.739	0.413	0.298	4 (4.56)
	National Baseline[49]	0.343	0.555	0.424	0.353	0.547	0.429	/	/	/	/
	TabNet[48]	0.504	0.584	0.352	0.510	0.424	0.399	0.684	0.649	0.655	5 (4.78)
	BERT[47]	0.491	0.241	0.323	0.484	0.412	0.363	0.411	0.335	0.333	8 (8.56)
	SafeTraffic LLM 8B	0.622	0.630	0.618	0.640	0.636	0.634	0.756	0.763	0.755	2 (2.22)
	SafeTraffic LLM 70B	0.630	0.682	0.649	0.648	0.644	0.644	0.760	0.775	0.759	1 (1.00)
Illinois	RandomForest[42]	0.462	0.554	0.383	0.430	0.452	0.338	0.610	0.670	0.632	3 (4.44)
	AdaBoost[44]	0.403	0.183	0.251	0.318	0.147	0.200	0.109	0.083	0.083	10 (10.00)
	CatBoost[46]	0.457	0.543	0.388	0.454	0.446	0.404	0.535	0.656	0.579	4 (4.78)
	DecisionTree[43]	0.426	0.514	0.410	0.417	0.398	0.361	0.504	0.624	0.548	7 (6.22)
	LogisticRegression[45]	0.413	0.439	0.410	0.360	0.385	0.355	0.379	0.477	0.400	8 (7.33)
	XGBoost[41]	0.442	0.575	0.340	0.405	0.419	0.278	0.678	0.694	0.683	5 (5.22)
	National Baseline[49]	0.369	0.136	0.199	0.442	0.195	0.271	/	/	/	/
	TabNet[48]	0.460	0.568	0.369	0.404	0.224	0.265	0.710	0.666	0.677	6 (5.67)
	BERT[47]	0.444	0.197	0.273	0.417	0.341	0.321	0.302	0.290	0.292	9 (8.00)
	SafeTraffic LLM 8B	0.529	0.529	0.533	0.578	0.584	0.571	0.701	0.768	0.721	2 (2.11)
	SafeTraffic LLM 70B	0.534	0.587	0.543	0.554	0.561	0.548	0.727	0.767	0.737	1 (1.44)
North Carolina (zero-shot)	Llama 3[23]	0.270	0.213	0.221	0.351	0.339	0.294	0.512	0.527	0.398	2 (2.11)
	BERT[47]	0.304	0.093	0.142	0.345	0.269	0.243	0.271	0.316	0.158	3 (2.89)
	SafeTraffic LLM 8B	0.566	0.582	0.545	0.536	0.570	0.508	0.667	0.750	0.675	1 (1.00)
Maine (zero-shot)	Llama 3[23]	0.118	0.045	0.060	0.311	0.645	0.408	0.478	0.638	0.544	3 (2.44)
	BERT[47]	0.636	0.469	0.518	0.536	0.476	0.503	0.074	0.182	0.106	2 (2.33)
	SafeTraffic LLM 8B	0.617	0.550	0.575	0.597	0.585	0.579	0.686	0.826	0.686	1 (1.22)
Ohio (zero-shot)	Llama 3[23]	0.347	0.205	0.183	0.222	0.407	0.108	0.570	0.376	0.453	2 (2.33)
	BERT[47]	0.345	0.119	0.177	0.575	0.331	0.420	0.276	0.551	0.320	3 (2.56)
	SafeTraffic LLM 8B	0.510	0.602	0.529	0.498	0.630	0.544	0.702	0.730	0.706	1 (1.11)

We present quality metrics along with model rankings by averaging the column-wise rank. For baseline models, we use the same input features as those used in *SafeTraffic LLM*. For models that only accept numerical inputs, we discretize or group the input features accordingly. For models that accept textual inputs, such as BERT[47], we directly use the prompt-based inputs for fine-tuning. Detailed descriptions of the data sources and hyperparameter configurations are provided in "Hyperparameter settings" in the "Methods" section. In supervised fine-tuning experiments on the Washington and Illinois datasets, *SafeTraffic LLM* outperforms all other methods, with *SafeTraffic LLM* 70B achieving the best performance. Additionally, we evaluate *SafeTraffic LLM*'s generalization ability by testing it on the North Carolina, Maine, and Ohio datasets after fine-tuning on the Illinois dataset. *SafeTraffic LLM* demonstrates strong generalization capabilities in zero-shot experiments compared with other baseline models (see "Adopted baselines" in the "Methods" section for baseline model settings in zero-shot experiments).

https://doi.org/10.1038/s41467-025-64574-w

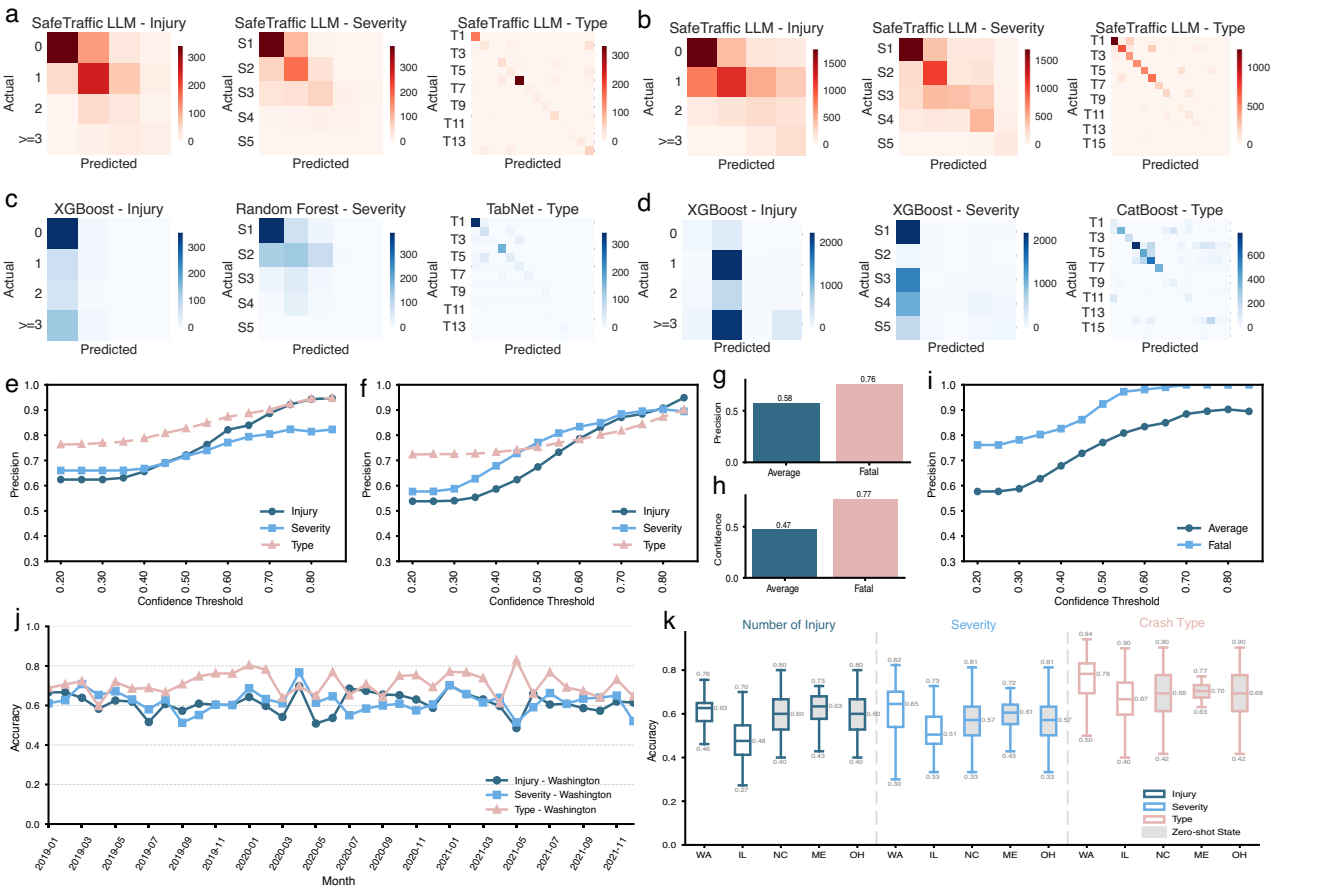


Fig. 4 | *SafeTraffic LLM* provides predictions with trustworthiness. The confusion matrices generated by the *SafeTraffic LLM* for both the **a** Washington and **b** Illinois datasets clearly demonstrate improved prediction results (we select the best results for each task based on the F1-score). In contrast, baseline models tend to predict the most frequent category across both the **c** Washington and **d** Illinois datasets (we show baseline models with the best F1-score. The performances for other baseline models can be found in Supplementary Fig. 7). Meanwhile, *SafeTraffic LLM* produces trustworthy predictions for both the **e** Washington and **f** Illinois datasets. Higher confidence levels in the model's predictions correspond to an increased likelihood of accuracy. Furthermore, **g** The *SafeTraffic LLM* achieves higher precision for fatal-crash predictions. **h** Fatal-crash predictions also exhibit higher confidence in the Illinois dataset. The Washington dataset is not shown due to limited fatal cases. **i** For fatal crashes, the *SafeTraffic LLM* achieves near-perfect precision (97.61%) when the confidence score exceeds 0.6, indicating that the *SafeTraffic LLM* is highly accurate and trustworthy for fatal crashes. **j** The 3-year temporal comparison of monthly prediction accuracy across tasks (2019–2021). The used *SafeTraffic LLM* was fine-tuned on the 2022 Washington dataset and evaluated on the 2019–2021 Washington datasets to assess its temporal generalization capability. The central line represents the median; the box spans from the 25th to 75th percentiles; whiskers extend to 1.5 × IQR. **k** The prediction performance at the county level, aggregated in a box plot. *SafeTraffic LLM* demonstrates stable performance at the county level for fine-tuning tasks in Illinois (IL) and Washington (WA), as well as zero-shot tasks in Maine (ME), North Carolina (NC), and Ohio (OH). States evaluated in zero-shot settings are highlighted with a gray background. Source data are provided as a Source data file.

(see "Expected crash prediction confidence score calculation" in the "Methods" section for the calculation details of the confidence score). Figure 4e, f illustrates the trend of accuracy in relation to the confidence scores of *SafeTraffic LLM* 's predictions for the Washington and Illinois datasets. The results indicate that our model achieves greater accuracy at higher confidence levels. For instance, for the *Number of Injury* prediction task in the Washington dataset, when the model's confidence score exceeds 0.40, the accuracy rises above 0.65, and with confidence scores over 0.60, the accuracy surpasses 0.80. This relationship is even more pronounced for fatal-crash predictions (see Fig. 4g–i). The strong positive correlation between confidence scores and accuracy showcases the quantifiable trustworthiness of the *SafeTraffic Copilot*. By providing reliable confidence scores alongside predictions, the framework empowers informed decision-making in real-world applications.

SafeTraffic LLM exhibits reliable spatial and temporal generalization capabilities. For spatial generalization, we evaluated *SafeTraffic LLM* by training on the Illinois dataset and testing on three unseen states: Maine, North Carolina, and Ohio. Without any additional fine-tuning, *SafeTraffic LLM* achieved average F1-scores of 0.576 in North Carolina, 0.613 in Maine, and 0.593 in Ohio, which is comparable to its performance in Illinois (see Table 1). Beyond state-level evaluation, we also assessed generalization at the county level. As shown in Fig. 4j, most counties exhibit an accuracy variation within 10%–20%, demonstrating the model's stable performance across regions. For temporal generalization, we evaluate the model fine-tuned on 2022 Washington data using data from 2019 to 2021 in the same state. As shown in Fig. 4k, the model maintains stable performance across months for all three tasks: *Number of Injury*, *Severity*, and *Type* prediction, with over 75% of the months falling within a ±10% accuracy range. Aggregated yearly performance is presented in Supplementary Section 3.3.

SafeTraffic Attribution framework

Understanding how *SafeTraffic LLM* generates accurate predictions and how various components of the input prompt influence the outcomes is fundamental to enabling evidence-based decision-making. In our analysis, we focus exclusively on severe crashes (i.e., fatal and

Fig. 5 | Single case feature-attribution results for *Severity* task. The left part displays the full prompt from **a** Washington and **b** Illinois, with different colors representing various semantic text sequences. The right part illustrates the feature contribution assigned to each text sequence. Positive contributions signify a supportive role in the model's prediction, whereas negative contributions indicate a detracting influence. The absolute value of these contributions represents the importance of each sequence to the model's output.

serious injury crashes) to identify the contributing factors behind these events. As discussed above, the *SafeTraffic LLM* 's confidence score strongly correlates with its predictive accuracy for severe crashes. Consequently, the confidence score associated with severe crash predictions (hereafter referred to as the confidence score) can be used as an indicator of crash risk level: a higher confidence score corresponds to greater prediction accuracy for severe crashes, which in turn reflects a higher likelihood that the crash is severe (rather than minor or no apparent injury) in the real world. Notably, the *SafeTraffic LLM* 's confidence scores tend to be lower than their corresponding accuracy values, indicating that using the confidence score is a conservative estimate of risk.

Within the *SafeTraffic Attribution* framework, a sentence-based feature contributions calculation method was proposed to identify how each sentence contributes to the LLM's outputs based on Shapley theory, which is recognized as a systematic and equitable method for attributing the contribution of each feature to a model's output[31,32], thereby revealing crash-related factors at the event level

(see "*SafeTraffic Attribution*" in the "Methods" section for details). In essence, each feature's contribution represents its share of responsibility for the model's confidence in a particular prediction. The sum of all feature contributions equals the confidence score itself. Figure 5 illustrates sentence-level feature contributions for the severity of individual crash events, using one crash from Washington and one from Illinois as examples. In the Washington crash example (Fig. 5a), *Driver Behavior* (e.g., reckless driving or speeding) is the primary factor contributing to serious injury crashes, with the feature contribution of 0.258. *Person Info* (e.g., no seatbelt use) also shows a substantial impact with the feature contribution of 0.149. By contrast, *Dynamic Info* (daylight and dry roads) lowers the probability of a crash with serious injuries with a negative feature contribution of −0.009. While in the Illinois example (Fig. 5b), an elevated *BAC* (Blood Alcohol Content, with feature contribution of 0.284) and the presence of a *Work Zone* (feature contribution of 0.462) notably increase the likelihood of fatal-crash outcomes. More additional sentence-level feature-attribution analysis can be found in Supplementary Sections 4.1 and 4.2. The following sections utilize *SafeTraffic Attribution* framework to examine feature importance from two perspectives: (1) at the inference stage, to identify key factors influencing crash predictions under various conditions and high-risk scenarios, and (2) at the fine-tuning stage, for which data are more critical for model learning.

Factor attribution at the inference stage for conditional risk analysis

Conditional analysis evaluates crash outcomes across various scenarios, such as driving with or without alcohol consumption, to quantify the risk factors associated with each scenario. Severe crashes (serious injuries and fatal crashes) were prioritized in the conditional analysis due to their critical importance for traffic safety. These crashes, particularly fatal ones, were predicted accurately and reliably by *SafeTraffic LLM* (see Fig. 4g–i). Five key contributing factors were identified for this conditional analysis: *Driver BAC* (BAC = 0 mg/dL or not offered/BAC < 80 mg/dL/BAC ≥ 80 mg/dL), *Roadway Type* (Highway/not highway), *Work Zone* (Work zone/not work zone), *User Type* (Pedalcyclist or pedestrian/not pedalcyclist or pedestrian), and *Driver Behavior* (Aggressive driving / impairment-related behavior/traffic rules violations/improper driving/others). Collectively, these factors accounted for an average of 79.33% of the model's overall attribution in predicting serious and fatal crashes (see Fig. 6b). A summary of key findings is provided:

- The BAC record emerges as a critical determinant in predicting serious and fatal crashes. Among all contributing factors, BAC accounts for 25.26% of the total contribution to serious and fatal-crash prediction (see Fig. 6b). Notably, its contribution substantially increases when a driver consumes alcohol, irrespective of the amount. When drivers are under the influence of alcohol even if their BAC does not exceed the legal intoxication limit of 80 mg/dL[33,34], this factor's feature contribution still reaches around 0.45, surpassing that of most other factors in many cases (see Fig. 6a). Conversely, when a driver's BAC is recorded as "zero or not offered," its contribution approaches zero, indicating minimal impact on the model's predictions.
- Driving in a work zone is already risky under sober conditions, but alcohol consumption greatly increases the danger, making it one of the most hazardous scenarios for severe injury crashes. As shown in Fig. 6a, driving in a work zone while sober ("Work Zone-Yes" and "BAC = 0 or not offered") contributes little to severe crash outcomes, with an average feature contribution of 0.03. However, after consuming alcohol (whether "BAC is higher than 80 mg/dL" or "BAC less than 80 mg/dL"), the work zone feature contribution rises more than seven times to an average of 0.22. Furthermore, the overall crash risk increases substantially when driving in a work zone after drinking, as indicated by an average risk level of 0.78, compared to 0.44 under sober conditions. These findings indicate that work zones become especially hazardous when alcohol consumption is involved, creating one of the highest-risk scenarios for severe crash outcomes. Potential drunk driving warnings and risk mitigation strategies shall be closely linked with work-zone areas.
- Aggressive and impairment-related behaviors pose nearly three times the risk for severe crash outcomes compared to other driver behaviors. As illustrated in Fig. 6c, aggressive driving emerges as the most important contributor among driver behaviors, with a median feature contribution of 0.195. Impairment-related behavior, including driving under the influence of alcohol or drugs, also has a substantial influence, with a median feature contribution of 0.154. In comparison, other improper driver behaviors, such as traffic rule violations (median feature contribution of 0.055) and distractions like mobile phone use (categorized under improper driving, with median feature contribution of 0.015), show below-average contributions to serious and fatal crashes. The "other" category, which includes normal driving and unknown behaviors, has the smallest impact, with a feature contribution of 0.007.
- The co-occurrence of risk factors substantially increases the expected crash risk level. As illustrated in Fig. 6a, c, our analysis reveals a strong correlation between the number of risk factors present in a crash and the expected risk level for severe crash outcomes. High-risk factors are defined as those with a 75th percentile contribution exceeding 0.2 in Fig. 6c, including driving after drinking (BAC < 0 mg/dL), driving in work zones, pedestrian-involved crashes, and high-risk driver behaviors (aggressive or impairment-related). When no risk factor is involved, the average risk level for severe crash outcomes is estimated at 0.47. This value increases to 0.59 with one risk factor, rises to 0.68 with two, and reaches 0.73 when three risk factors co-occur. These findings underscore the need for transportation agencies to implement comprehensive, multi-dimensional interventions, particularly in scenarios characterized by overlapping high-risk conditions.

Factor attribution at training stage for effective data collection and model development

Event information and unit information are the most important components for the model training. While feature contributions at the inference stage reveal which features drive critical crash outcomes, understanding feature contributions during training provides deeper insights into which data components most effectively enhance model accuracy. As shown in Fig. 6d, the feature contributions of each component in the Washington and Illinois datasets are shown, demonstrating their impact on the model's performance during training (see Supplementary Table 10 for detailed results and "*SafeTraffic Attribution*" in the "Methods" section for calculation details). The results indicate that in both the Washington and Illinois datasets, for the *Severity* task, the unit information describing attributes of the primary entities involved in the crash has the highest contribution to the model's performance (0.314 in Washington, 0.248 in Illinois). For the Crash *Type* Prediction task, the event information, which provides information on the vehicle's movement prior to the crash, has the highest contribution (0.388 in Washington, 0.283 in Illinois), followed by the unit information (0.257 in Washington, 0.279 in Illinois) and other components.

Discussions

Transforming crash prediction into a text-reasoning task unlocks the full richness of multi-modal safety data. By integrating crash narratives, satellite and incident images, and infrastructure attributes into a unified textual prompt, foundational language models can jointly reason

https://doi.org/10.1038/s41467-025-64574-w

Fig. 6 | Conditional risk analysis for the serious injury and fatal crashes. Higher confidence scores in *SafeTraffic LLM* 's predictions correspond to greater accuracy, allowing the confidence score (calculated as the sum of feature contributions for all data components) to serve as an indicator of risk level for serious and fatal crashes. **a** The aggregated average crash risk level under different feature combinations. Cases are grouped based on their combinations of conditions (indicated by dark dots). For each group, the average crash risk level and the average contributions of five key features are calculated and visualized, including driver behavior, user type, whether driving occurred within a work zone, roadway type, and Blood Alcohol Content (BAC, in mg/dL). We use 80 mg/dL as a key threshold, based on the legal limit for ethanol concentration[33,34]. The detailed data is provided in Supplementary Table 11. **b** The overall average contribution of each feature. Taking BAC as an example, each case yields a BAC contribution value, and the mean of these values across all cases represents the overall contribution of BAC. The inner circle represents the average absolute contribution, and the outer circle shows its proportional impact on crash risk. **c** Average feature contribution for each factor under specific values. Boxes are colored pink if the median exceeds the overall median for that factor, and blue otherwise. Each box is annotated with its median value. The central line represents the median; the box spans from the 25th to 75th percentiles; whiskers extend to 1.5 × IQR. **d** Feature contributions of different data components during the training stage for the Washington and Illinois datasets. Source data are provided as a Source data file.

over behavioral cues ("alcohol-impaired," "work-zone"), pre-crash trajectories, and environmental context instead of treating them as disconnected numbers. Our AI-expert co-designed prompts let LLMs outperform conventional baselines while attribution scores reveal which factor combinations, e.g., impairment plus work zones, are the most elevated risk, guiding both targeted interventions and smarter data-collection priorities. This multi-modal-to-text paradigm therefore signals a powerful solution: integrating diverse crash information streams through foundational models not only boosts predictive accuracy but also yields transparent, actionable insights for continuous safety improvements.

Integrating rich data with a powerful foundation-model engine shifts prediction from simple distribution fitting to situational-aware reasoning, yielding transparent and trustworthy insights that generalize across regions. In both Washington and Illinois, accuracy exceeds 70% once confidence surpasses 60%, with a near-linear rise in precision thereafter—giving practitioners a clear, quantifiable handle on uncertainty, as shown in Fig. 4e, f. Furthermore, the proposed *SafeTraffic Attribution* components turn this trust into action: it ranks textual, visual, and categorical cues by their contribution to confidence, highlighting the levers that most elevate risk. Notably, alcohol-impaired driving raises the

severe-crash confidence score by 0.47, underscoring its critical policy relevance.

The *SafeTraffic LLM* 's conditional attribution engine pinpoints which factor combinations truly drive crash risk with trust, ranking scenarios by danger and revealing actionable "what-ifs." In data-rich settings, it reliably surfaces high-risk pairings, e.g., alcohol use in work zones or aggressive driving under impairment, guiding focused counter-measures such as on-site BAC checks or behavior-targeted education. Crucially, its probabilistic confidence signals rise with accuracy, lending quantifiable trustworthiness to every recommendation. When data are sparse, the same framework generalizes through simulation: analysts can toggle rarely observed variables (pedestrian presence, freeway geometry, etc.) and still obtain credible risk shifts. This blend of calibrated confidence and flexible what-if analysis empowers agencies to design precise, evidence-based traffic-safety interventions before crashes happen.

Aggregated data-attribution analysis pinpoints the crash-record elements that matter most, offering a generalizable blueprint for smarter, future-proof data collection and quality control. Currently, each state designs its own crash-report form, preventing a common standard and stifling national-scale analytics. During the fine-tuning stage, unit-level details, such as driver behavior, vehicle attributes, and event-level cues, such as vehicle movement, weather, and roadway conditions, emerged as the strongest predictors of injury severity. Prioritizing complete, high-resolution capture of alcohol use, vehicle defects, vulnerable-user status, and road context, therefore, maximizes model payoff, whereas gaps in these fields quickly erode performance. Moreover, the aggregated attribution analysis also provides a quantitative basis for assessing data quality, showing how missing or incomplete values in key components impair model performance. Feeding these insights back into template design lets agencies standardize richer, more consistent reporting protocols that fuel continual model refinement and transferability across regions, accelerating broad safety transformations without sacrificing generalizability.

Limitations and future work of the *SafeTraffic Copilot*. A primary limitation relates to the handling of multi-modal data. In the *SafeTraffic Copilot*, satellite images were processed into textual descriptions and incorporated into prompts. While this approach offers flexibility, advancements in multi-modal foundation models and increasing research on integrating multi-modal data with LLMs present promising alternatives[35]. Leveraging specialized image encoders or utilizing multi-modal foundation models for processing image data are compelling directions. Another potential limitation lies in the efficiency of model training and attribution. Fine-tuning LLMs and computing feature contributions have always required substantial resources and time. Although we employed LoRA fine-tuning and a stratified sampling technique to enhance efficiency[36], implementing the complete framework still demands substantial resources. This poses certain limitations when resources are scarce or in situations demanding rapid model deployment.

Methods

Raw data

The raw crash data used in this study were obtained from the HSIS[29] and Google Maps[30]. Data from the HSIS, sourced from multiple systems, encompasses a variety of formats, including categorical, numerical, and textual. In total, four main datasets were used:

- Crash data. This dataset captures the essential spatio-temporal and contextual attributes of each crash. It includes crash date, time, day of the week, and month, along with location details such as route number, milepost, and the surrounding area's classification (e.g., rural or urban). Higher-level planning attributes (e.g., roadway and functional classifications, intersection-related indicators) are also recorded. In addition, it documents the dynamic circumstances leading up to the event, including the number of vehicles and pedestrians involved, vehicle travel directions (increasing or decreasing milepost), and any maneuvers performed (e.g., lane changes, straight-line movement).
- Infrastructure data. This dataset details the physical and infrastructural features of the crash site. Key elements include the type of road surface (e.g., asphalt or concrete), average annual daily traffic (AADT), posted speed limits, and access control mechanisms. It also encompasses dimensions such as total road width, right and left shoulder widths, and median width (including median barriers if present), as well as road surface conditions (e.g., dry or wet) and ambient lighting at the time of the crash (e.g., daylight or dusk).
- Vehicle data. This dataset consolidates information on the vehicles involved in each crash, including vehicle type (e.g., passenger car or truck), intended use (e.g., commercial or private), mechanical condition (e.g., defects), and relevant driver actions (e.g., lane changes or stopping). Additional information on airbag deployment and occupant ejection status provides further granularity.
- Person data. This dataset compiles information about individuals involved in the crash, detailing demographic characteristics such as age, gender, and seating position. It also includes the use of safety equipment (e.g., seat belts or helmets) and any contributing factors, such as driver distraction or impairment.

The satellite images obtained from Google Maps serve as a supplementary data source to complement the HSIS dataset. Overall, we collected 16,188 crash event data from Washington State and 42,715 events from Illinois State for further analysis. We also collect and process 2250 events from Maine, 2250 from Ohio, and 2802 from North Carolina to evaluate the model's cross-region training-free generalization.

SafeTraffic Event dataset construction

To adapt the raw data for LLMs' fine-tuning process, we employ the feature engineering and textualization process to generate textual inputs (see Fig. 7). We followed the following process to generate a textual prompt from raw data entry:

- Data mapping and organization. For each crash, we associated the crash report with the involved vehicles and individuals using the crash ID, thus obtaining descriptions of the crash and the persons involved. The route ID and milepost were used to identify the specific road segment where the crash occurred, allowing us to gather related road and environment information from infrastructure data. The integrated data was then systematically organized into four categories: general information, infrastructure information, event information, and unit information, aligning with the components outlined above.
- Satellite images textualization. The HSIS datasets provide GPS coordinates for crash locations in Washington and Illinois. To address missing information, such as the number of road lanes, high-resolution satellite images (512 × 512 pixels at a zoom level of 19) were retrieved using these GPS coordinates via the Google Maps API. These images supplement the crash dataset with crucial infrastructure and environmental context. Descriptive textual annotations were generated from the satellite images using GPT-4, filling key gaps in the original dataset. These annotations include information such as the number of lanes at the crash site, whether the crash occurred at an intersection, and whether the surrounding area is residential. Image-related information enhances the model's performance; see Fig. 8 for details.
- Dimensionality reduction. Raw data include abundant attributes with rich and varied descriptions. However, some features suffer from insufficient distinction between attribute values due to the original classification's complexity. To address this, we performed

https://doi.org/10.1038/s41467-025-64574-w

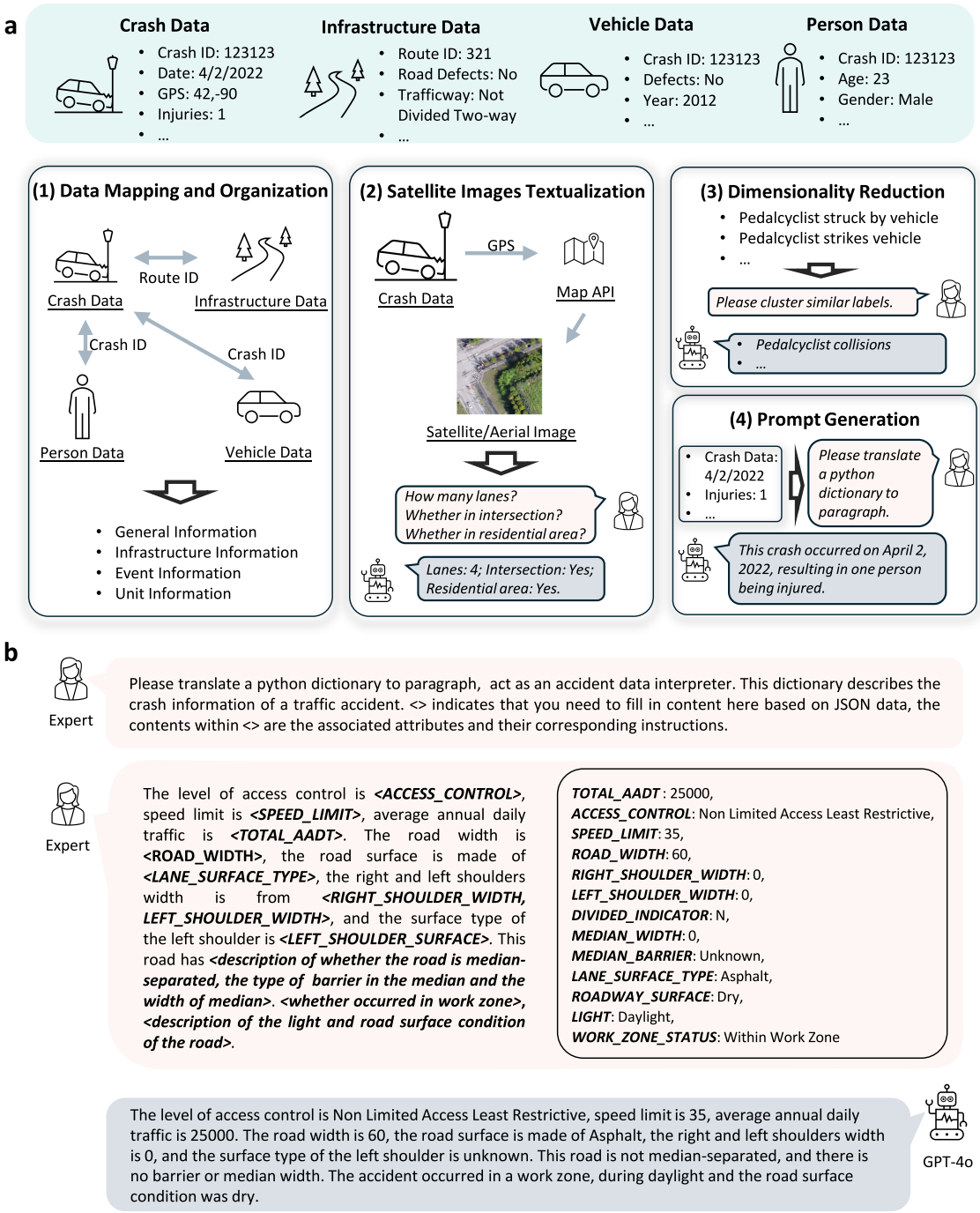
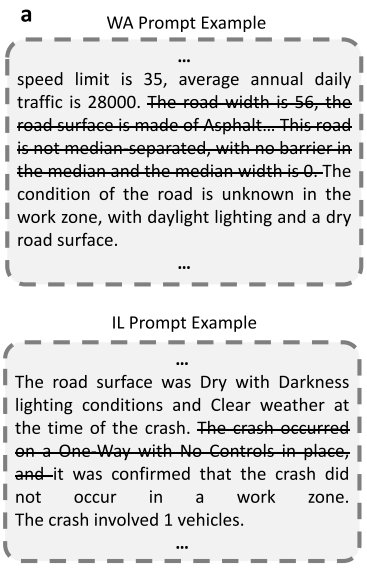


Fig. 7 | The construction procedure of *SafeTraffic Event* dataset. a Data processing. Four raw datasets from HSIS (crash, infrastructure, vehicle, and person data) are used to construct a prompt through four steps. (1) Data mapping and organization: Link the datasets and organize them into four parts: general, infrastructure, event, and unit. (2) Satellite image textualization: Retrieve satellite images via GPS coordinates using the Google Maps API, then employ GPT-4o to extract text-based information. (3) Dimensionality reduction: Combine targets with similar values using GPT-4o. (4) Prompt generation: Use the processed data from the previous steps to generate a prompt for each part. **b** AI-expert textualization. An example of the infrastructure information part of an event case in the Washington dataset is shown.

dimensionality reduction on these attributes by combining domain experts' insights with GPT-4o clustering results. For example, similar classifications like "pedalcyclist struck by vehicle" and "pedalcyclist strikes vehicle" were clustered under a broader category such as "pedalcyclist collisions." This process generalized the data and reduced redundancy. See Supplementary Table 6 for detailed information.

- Prompt generation using an AI-expert textualization method. To generate logically coherent and continuous textual data suitable for LLM training, we transformed each category of data into text format using GPT-4o[12]. All data are organized as key-value pairs, and we get four parts of the key-value pairs for each event case. Then GPT-4o is used to generate the text prompt for each section of the key-value pairs individually. For each part, we apply a straightforward prompt to GPT-4o, such as "Please translate a python dictionary to paragraph, act as a crash data interpreter." The text content is extracted from GPT-4o's response for each part, consisting of approximately 100 words.



b

State	Inputs	Injury			Severity			Type		
		Accuracy	Precision	F1-score	Accuracy	Precision	F1-score	Accuracy	Precision	F1-score
WA	T + I	0.622	0.630	0.618	0.640	0.636	0.634	0.756	0.763	0.755
	T	0.597	0.631	0.582	0.651	0.628	0.624	0.732	0.762	0.728
IL	T + I	0.529	0.529	0.533	0.578	0.584	0.571	0.701	0.768	0.721
	T	0.474	0.550	0.487	0.507	0.538	0.501	0.701	0.765	0.716

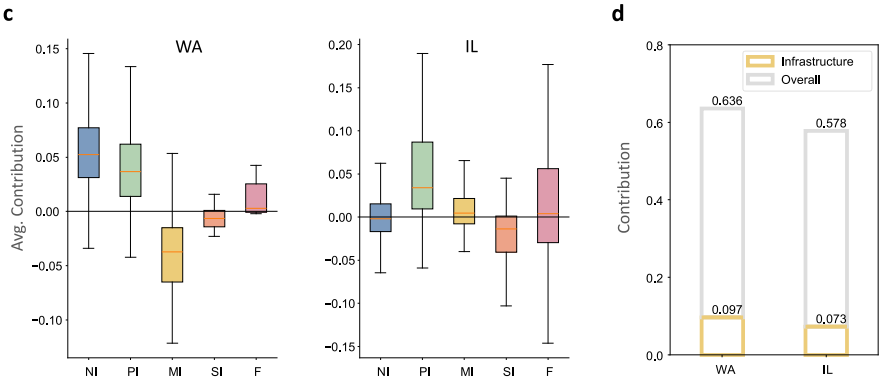


Fig. 8 | Analysis of the impacts of visual-textual information integration in *SafeTraffic LLM*. **a** Examples of prompt modifications with image-derived information removed. **b** Performance comparison for expected crash prediction on *Num. of Injury*, *Severity*, and *Crash Type* prediction tasks on the Illinois and Washington datasets, using T + I (Text + Image) and T (Text-only) input modalities. **c** Average contribution of image-derived information at the inference stage. **d** Contribution of the image-derived paragraph at the training stage. The central line represents the median; the box spans from the 25th to 75th percentiles; whiskers extend to 1.5 × IQR. Source data are provided as a Source data file.

By linking four parts of the text, we obtain a comprehensive textual description for each crash event case. The detailed process is shown in Fig. 7b. For the Maine, Ohio, and North Carolina datasets, the prompts were constructed by filling values into the Illinois template if the features are matched; otherwise, the features are set to None in the Illinois template.

We select three variables as the prediction targets: *Injury*, *Severity*, and crash *Type*. The three targets are defined as:

- The $Injury_i^{\mathcal{D}} \in \{f(l)|l=0,1,2,\cdots\}$, where i denotes the ith data in the dataset, $\mathcal{D} \in \{\mathcal{W},\mathcal{I}\}$ denotes the Washington dataset $\mathcal{W}$ or the Illinois dataset $\mathcal{I}$, l represents the number of people injured, and $f(l)$ denotes the label when the injured people is l.
- The $Severity_i^{\mathcal{D}} \in \{S_k|k=1,2,\cdots\}$, where S_k is the kth level of crash severity.
- The $Type_i^{\mathcal{D}} \in \{T_k^{\mathcal{D}}|k=1,2,,\cdots\}$, where $T_k^{\mathcal{D}}$ is the kth label of crash type in dataset $\mathcal{D}$.

We utilize these three variables to describe the crash result $CR_i^{\mathcal{D}}$. The crash outcome can be presented in the following format: $CR_i^{\mathcal{D}} = (n_i^{\mathcal{D}}, s_i^{\mathcal{D}}, t_i^{\mathcal{D}})$. For numerical variables, the function $f(l)$ describes the number of people injured in a crash as follows: "zero" if $l=0$, "one" if $l=1$, "two" if $l=2$, and "three and more than three" if $l \geq 3$, the values for S_k and $T_k^{\mathcal{D}}$ are provided in the Supplementary Table 4 and Supplementary Table 5.

SafeTraffic LLM

We fine-tune *SafeTraffic LLM* by adapting LLaMa 3.1[13] to crash prediction tasks to enhance the LLMs' capabilities in interpreting crash data, identifying critical factors, and conducting feature-attribution analysis to offer insights for crash prevention. In this section, we will introduce detailed information on the fine-tuning process.

During the fine-tuning of LLMs, a single input consists of three components: the system prompt, the user prompt, and the target prompt. The system prompt introduces the task, for example: "You are a helpful assistant designed to predict the severity of a traffic crash...". The user prompt comprises the four content parts detailed in "SafeTraffic Event dataset construction" section for each case. The target prompt represents the expected output. Examples of these prompts are shown in Fig. 3 and Supplementary Section 2.3. We tokenize the text inputs using LLaMA 3.1's tokenizer.

To adapt the LLM as a crash classifier, additional tokens have been incorporated into the tokenizer's vocabulary, and the detailed crash attribute categories are listed in Supplementary Table 4 and Supplementary Table 5. Specifically, for predicting the number of people *Injuries* of Washington dataset and Illinois dataset, we have introduced four special tokens: `<ZERO>`, `<ONE>`, `<TWO>`, and `<THREE AND MORE THAN THREE>`. Similarly, for predicting the Crash *Severity* of the Washington dataset and the Illinois dataset, we use five additional tokens: S_k, where $1 \leq k \leq 5$, corresponding to different levels of severity. The *Type* task differs slightly between the Washington and Illinois datasets. For Washington datasets, we utilize 14 special tokens: $T_k^{\mathcal{W}}$, where $1 \leq k \leq 14$, each representing a specific crash type. For Illinois datasets, we utilize 16 special tokens: $T_k^{\mathcal{I}}$, where $1 \leq k \leq 16$. The parameters of the input and output embedding layers are set as trainable, enabling the model to align the representations of these special tokens with the existing embedding space.

During the fine-tuning phase, the traffic forecasting task is framed as a next-token generation task. Given an input prompt x_i and its prediction target y_i, we construct the full prompt as $T_i = \text{concat}(x_i, y_i)$, where concat(·) denotes the concatenation operation that appends the target label y_i to the input x_i as a special token. The next-token generation process can be described as:

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

where T_i is the ith item in the training data, p_θ is the LLM, $t_j^{(i)}$ denotes the jth token in T_i. By maximizing the likelihood $p_\theta(T) = \prod_{i=1}^{N} p_\theta(T_i)$, the LLM's parameters are learned. Both the system prompt and the user prompt are masked for loss computation during training. We also used a uniform data sampling strategy during the training process to

facilitate the convergence of *SafeTraffic LLM* [16]. Through this process, the model learns to make predictions for a traffic crash.

Expected crash prediction confidence score calculation

The confidence score is a critical component that links model predictions to interpretability within the *SafeTraffic Copilot*. The confidence score quantifies the model's certainty in its prediction for a given input. Since we incorporate target labels as special tokens in the LLM's vocabulary and fine-tune the model to generate only these tokens as outputs, we define the confidence score based on the predicted token's probability. Specifically, given a textual input x_i and its corresponding label y_i, the confidence score $C(x_i)$ is defined as:

$$C(x_i) = \max_{y_i \in \mathcal{Y}} p_\theta(y_i|x_i) \tag{2}$$

where $\mathcal{Y}$ denotes the set of all possible labels (e.g., *fatal*, *serious injury*, etc., for crash severity prediction). $p_\theta(y_i|x_i)$ is the softmax probability assigned by the model to class y_i, which can be computed by applying the softmax function over the logits corresponding to the special tokens representing each label.

For a given threshold t, let N_t denote the number of samples with confidence scores greater than t. Among these, R_t samples are correctly classified. The accuracy at threshold t is then given by

$$\mathrm{Acc}_t = \frac{R_t}{N_t}. \tag{3}$$

By computing the accuracy at different thresholds t, we can plot the relationship between accuracy Acc_t and the threshold t, as shown in Fig. 4e, f.

Hyperparameter settings

In our experiments, we follow LoRA[37] to fine-tune LLaMA 3.1 models. Specifically, we update only the input and output layers directly, while all remaining layers are frozen and trained through LoRA. We use the AdamW optimizer[38] with a learning rate of 3e-4 and a batch size of 32 (with gradient accumulation over 8 steps). The models are trained on 8 NVIDIA A100 GPUs (80GB memory each) using DeepSpeed[39] for efficient distributed training.

Data split

We split the Washington and Illinois dataset into training, validation, and test sets in a 7:1.5:1.5 ratio. Since the Washington dataset contains relatively few crash events per year, we utilized as many reports as possible to ensure sufficient training data. However, the data distribution across different classes is highly imbalanced. For example, in the crash severity prediction task in the Washington dataset, the ratio of $\#S_1/\#S_5$ is nearly 100:1, where $\#S_k$ is the number of data with label S_k. The imbalanced data distribution presents a great challenge for the model's training and evaluation. During the fine-tuning, we used a uniform sampling strategy to train the model on this unbalanced data. Similarly, to facilitate the model's evaluation, for the validation set and test set, we removed most of the data with a crash severity category of S_1. Specifically, after processing, the dataset consisted of 16,188 records, with 11,332 used for training, 2428 for validation, and 2428 for testing. To balance the validation and test set for better evaluation, we removed 1428 S_1 data and used 1000 remaining data for the validation set and test set separately. Compared with Washington state, more crash records can be used in Illinois state to generate a dataset. As a result, we were able to balance all subsets, including the training, validation, and test sets. Ultimately, the Illinois dataset comprised 42,715 records, with 29,307 used for training, 6704 for validation, and 6704 for testing. See Supplementary Section 1.3 for the detailed distribution for each dataset.

Evaluation metrics

In evaluating the model performance as a classification task, we employ weighted accuracy, precision, and F1-score as metrics. In the context of a classification task, we have four notations: True Positive (*TP*), True Negative (*TN*), False Positive (*FP*), and False Negative (*FN*). Using these notations, we can represent the metrics as follows:

- Accuracy is one of the most commonly used measures for the classification performance, and it is defined as a ratio between the correctly classified samples to the total number of samples as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

- Precision represents the proportion of positive samples that were correctly classified to the total number of positive predicted samples, which reflects the performance of the prediction:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

- F1-score combines results on precision and recall. It is the harmonic mean of precision and recall, which can be calculated using the formula:

$$\text{F1} - \text{score} = \frac{2}{\text{Precision}^{-1} + \text{Recall}^{-1}} = 2 \cdot \left(\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}\right) \tag{6}$$

where Recall = $TP/(TP + FN)$.

Adopted baselines

We follow recent studies[40] and adopt machine learning models, including XGBoost[41], Random Forest (RF)[42], Decision Trees (DT)[43], Adaptive Boosting (AdaBoost)[44], Logistic Regression (LR)[45], and Categorical Boosting (CatBoost)[46]. We also include deep learning models such as BERT[47] and TabNet[48]. In addition, we consider the National Average[49], which predicts crash severity distributions using calibrated Severity Distribution Functions. For these models, the Bayesian optimization method (*BayesSearchCV*) is used to facilitate the identification of optimal hyperparameters, such as *max_depth* and *learning_rate*. To ensure a fair comparison across baseline models, we retained the original architecture and design of each model, modifying only the input data format when necessary. Detailed information on hyperparameter settings and input data preprocessing for all baseline models is provided in Supplementary Section 1.2.

The experiments on the North Carolina, Maine, and Ohio datasets are conducted under a zero-shot setting, where the model is fine-tuned only on the Illinois dataset and has never seen data from North Carolina, Maine, and Ohio during training. Traditional machine learning models perform poorly in this context due to their limited ability to adapt. Therefore, to ensure a fair comparison under the same conditions, we introduce two baseline methods:

- BERT[47]. Leveraging its pre-training on large corpora, BERT possesses a certain degree of generalization capability. In our experiments, we fine-tune BERT using prompts from the Illinois dataset and evaluate its zero-shot performance on the North Carolina and Maine datasets.
- CoT[50]. Chain-of-thought (CoT) reasoning enables language models to perform multi-step inference by generating intermediate reasoning steps before arriving at a final answer. Zhen et al.[23] explored the use of CoT for zero-shot crash severity prediction

and reported improved performance over standard LLM prompting. Following their approach, we apply CoT prompting to evaluate zero-shot performance on the North Carolina and Maine datasets.

SafeTraffic Attribution

To identify the feature contribution of each factor to the prediction results, this paper introduces and adapts the concept of Shapley values[31]. Shapley value is a concept from cooperative game theory that has been widely adopted in machine learning to interpret model predictions[51]. It provides a way to fairly allocate the contribution of each feature to the outcome of a predictive model. In essence, the Shapley value quantifies how much each feature contributes to a prediction by considering all possible combinations of features. Formally, the Shapley value φ of a feature (or player) i in a cooperative game is defined as:

$$\varphi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} [v(S \cup \{i\}) - v(S)], \tag{7}$$

where $N = \{1, 2, \ldots, n\}$ is the index set of n features, S is a subset of N, and $v(S)$ is the utility of the subset S, which represents a measurable value, such as accuracy or prediction score, achieved by the model using only the subset S of features.

The Shapley value is utilized in both the training and inference stages in *SafeTraffic Copilot*. During the training stage, it quantifies the contributions of four primary categories of information: general information, infrastructure information, event information, and unit information. During the inference stage, the Shapley value is applied to assess the contributions of individual sentences to the prediction outcomes.

Feature contributions at the training stage

The Shapley value is utilized to assess the influence of different components in the training set on the model during training. As outlined in "Developing *SafeTraffic LLM* for predicting crashes" in the "Results" section, the jth prompt T_j in the dataset P is divided into five parts: c_0: system prompt (i.e., "You are a helpful assistant designed to predict the severity of a traffic crash..."), c_1: general information, c_2: infrastructure information, c_3: event information, and c_4: unit information. We denote $p_j(k)$ as the c_k portion of p_j. Given an index set S, we can construct a variant $T_j(S)$ by concatenating the parts in S. For example, if $S = \{0, 1, 2\}$, then $T_j(S)$ contains c_0, c_1, and c_2. Formally,

$$T_j(S) = \text{concat}_{k \in S}\, T_j(k), \tag{8}$$

where concat denotes concatenation. The resulting dataset based on S is $P(S) = \{T_j(S) | j = 0, 1, \ldots, L\}$, where L is the dataset size.

Referring to Equation (7), the contribution of part c_i at training, φ_i^{train}, is

$$\varphi_i^{\text{train}} = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} \cdot [v(P(S \cup \{0, i\})) - v(P(S \cup \{0\}))], \tag{9}$$

where $N = \{1, 2, 3, 4\}$ indexes the four content parts, and $v(P(S))$ is a performance metric (e.g., accuracy) obtained after retraining the model only on prompts in $P(S)$.

Sentence-level feature contributions at the inference stage

Unlike traditional machine learning models that primarily handle fixed-length feature vectors, LLMs process variable-length text sequences as input[52]. This characteristic makes commonly used Shapley value approximation methods, such as KernelSHAP[53] and DeepSHAP, less applicable to LLMs. Recent approaches like TokenSHAP[54] and TransSHAP[55] have been proposed to address this by decomposing input text into tokens and computing Shapley values at the token level. However, applying token-level Shapley value computation to *SafeTraffic LLM* introduces two primary challenges: (1) Computational limitations. The computational complexity of Shapley values is exponential in the number of players. In our *SafeTraffic LLM*, with an input size of approximately 500 tokens, the large-scale computation of token-level Shapley values for crash data becomes impractical. (2) Limited interpretability. Decomposing the prompt at the token level disregards inter-token dependencies, and the arbitrary masking or replacement of tokens can lead to semantic ambiguity and contextual shifts. These issues hinder a precise understanding of how individual features contribute to predictions. Moreover, paragraph-level analysis is too coarse for detailed attribution, since it can merge distinct features into a single category (e.g., driver and vehicle details under "unit information").

To overcome these limitations, we propose a sentence-level feature contributions calculation method for inputs of LLMs, which proceeds as follows:

- Sentence segmentation. The prompts are segmented using delimiters (e.g., commas "," or periods ".") to produce sentence-level units.
- Feature groups annotation. GPT-4o is used to group and label these sentences (see Fig. 5 for the groups' content). Each group is represented as c_k, where $k \in N' = \{1, 2, 3, \ldots n\}$. For the Washington dataset, $n = 14$, while for the Illinois dataset $n = 12$. Given index set $S' \subseteq N' \setminus \{i\}$, we can construct the prompt $T_j(S')$ similar to the process Equation (8).
- Feature contributions calculation based on the feature groups. Based on the constructed dataset, the feature contribution for the ith sentence-group $\varphi_{i,j}^{inf}$ for the jth item in the dataset can be calculated as:

$$\varphi_{i,j}^{\text{inf}} = \sum_{S' \subseteq N' \setminus \{i\}} \frac{|S'|!\,(n-|S'|-1)!}{n!} \cdot \Big[p_\theta\Big(y_j | T_j(S' \cup \{0, i\})\Big) - p_\theta\Big(y_j | T_j(S' \cup \{0\})\Big)\Big] \tag{10}$$

where p_θ represents the LLM, which returns the predicted probability of the targets y_j given the inputs. A higher $\varphi_{i,j}^{\text{inf}}$ indicates a greater contribution of the ith sentence group to the model's confidence for predicting y_j. To reduce computational overhead, we adopt a stratified sampling–based Shapley estimation method using complementary contributions[36].

Data availability

The examples of processed prompts generated in this study have been deposited in Zenodo (https://zenodo.org/records/16896765). The completed HSIS raw data[29] are available under restricted access, as HSIS is designed to provide data solely for research conducted in the public interest and intended for publication in a scientific journal or other national publication. Access can be obtained by submitting a formal request to HSIS staff via hsis@dot.gov, accompanied by a description of the proposed research and confirmation of compliance with HSIS usage guidelines. Detailed information can be found at https://highways.dot.gov/research/safety/hsis. The source data for Figs. 4, 6 and 8 and Supplementary Figs. 5, 7, 14 and 15 generated in this study are provided in the Source data file. Source data are provided with this paper.

Code availability

Code is publicly accessible at https://zenodo.org/records/16896765[56].

References

1. International Transport Forum (ITF). *Road Safety Annual Report 2023* (OECD Publishing, 2023).

2. Islam, M. R., Wang, D. & Abdel-Aty, M. Calibrated confidence learning for large-scale real-time crash and severity prediction. *npj Sustain. Mobil. Transp.* **1**, 1 (2024).
3. Bougna, T., Hundal, G. & Taniform, P. Quantitative analysis of the social costs of road traffic crashes literature. *Accid. Anal. Prev.* **165**, 106282 (2022).
4. Wen, X., Xie, Y., Jiang, L., Pu, Z. & Ge, T. Applications of machine learning methods in traffic crash severity modelling: current status and future directions. *Transp. Rev.* **41**, 855–879 (2021).
5. Yan, X. et al. Learning naturalistic driving environment with statistical realism. *Nat. Commun.* **14**, 2037 (2023).
6. Carrodano, C. Data-driven risk analysis of nonlinear factor interactions in road safety using Bayesian networks. *Sci. Rep.* **14**, 18948 (2024).
7. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
8. Mannering, F. L. & Bhat, C. R. Analytic methods in accident research: methodological frontier and future directions. *Anal. Methods Accid. Res.* **1**, 1–22 (2014).
9. Rahim, M. A. & Hassan, H. M. A deep learning based traffic crash severity prediction framework. *Accid. Anal. Prev.* **154**, 106090 (2021).
10. Sattar, K. et al. Transparent deep machine learning framework for predicting traffic crash severity. *Neural Comput. Appl.* **35**, 1535–1547 (2023).
11. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
12. Achiam, J. et al. Gpt-4 technical report. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.08774 (2023).
13. Grattafiori, A. et al. The llama 3 herd of models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2407.21783 (2024).
14. Gao, C. et al. S3: social-network simulation system with large language model-empowered agents. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2307.14984 (2023).
15. Schulze Buschoff, L. M., Akata, E., Bethge, M. & Schulz, E. Visual cognition in multimodal large language models. *Nat. Mach. Intell.* **7**, 96–106 (2025).
16. Du, H. et al. Advancing real-time infectious disease forecasting using large language models. *Nat. Comput. Sci.* **5**, 467–480 (2025).
17. Zhou, L. et al. Larger and more instructable language models become less reliable. *Nature* **634**, 61–68 (2024).
18. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
19. Singh, C., Askari, A., Caruana, R. & Gao, J. Augmenting interpretable models with large language models during training. *Nat. Commun.* **14**, 7913 (2023).
20. Dang, Y. et al. Explainable and interpretable multimodal large language models: a comprehensive survey. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2412.02104 (2024).
21. Fan, Z. et al. Learning traffic crashes as language: Datasets, benchmarks, and what-if causal analyses. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2406.10789 (2024).
22. de Zarzà, I., de Curtò, J., Roig, G. & Calafate, C. T. Llm multimodal traffic accident forecasting. *Sensors* **23**, 9225 (2023).
23. Zhen, H., Shi, Y., Huang, Y., Yang, J. J. & Liu, N. Leveraging large language models with chain-of-thought and prompt engineering for traffic crash severity analysis and inference. *Computers* **13**, 232 (2024).
24. Transportation Research Board and National Academies of Sciences, Engineering, and Medicine. *Leveraging Artificial Intelligence and Big Data to Enhance Safety Analysis: A Guide* (eds Wang, Y. et al) (The National Academies Press, Washington, DC, 2025).
25. Pei, X., Wong, S. & Sze, N.-N. A joint-probability approach to crash prediction models. *Accid. Anal. Prev.* **43**, 1160–1166 (2011).
26. Abdel-Aty, M., Hasan, T. & Anik, B. T. H. An advanced real-time crash prediction framework for combined hard shoulder running and variable speed limits system using transformer. *Sci. Rep.* **14**, 26403 (2024).
27. Abdel-Aty, M., Keller, J. & Brady, P. A. Analysis of types of crashes at signalized intersections by using complete crash data and tree-based regression. *Transp. Res. Rec.* **1908**, 37–45 (2005).
28. 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.* **43**, 1666–1676 (2011).
29. U.S. Department of Transportation, Federal Highway Administration. Highway Safety Information System (HSIS). https://highways.dot.gov (Accessed 13 January 2025) (2025).
30. Google for Developers. Google Maps Static API Documentation. https://developers.google.com/maps/documentation/maps-static (Accessed 13 January 2025) (2025).
31. Bordt, S. & von Luxburg, U. From shapley values to generalized additive models and back. In *Proc. International Conference on Artificial Intelligence and Statistics* 709–745 (PMLR, 2023).
32. Shapley, L. S. A value for n-person games. in *Contributions to the Theory of Games II* (eds Kuhn, H. W. & Tucker, A. W.) 307–317 (Princeton University Press, 1953).
33. Washington State Legislature. Revised code of washington: Driving under the influence. https://app.leg.wa.gov/rcw/default.aspx?cite=46.61.502 (Accessed 19 January 2025) (2025).
34. Illinois Secretary of State. Driving under the influence (dui). https://www.ilsos.gov/departments/drivers/traffic_safety/DUI/home.html (Accessed 19 January 2025) (2025).
35. Zhang, J., Huang, J., Jin, S. & Lu, S. Vision-language models for vision tasks: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**, 5625–5644 (2024).
36. Zhang, J. et al. Efficient sampling approaches to shapley value approximation. *Proc. ACM Manag. Data* **1**, https://doi.org/10.1145/3588728 (2023).
37. Hu, E. J. et al. LoRA: low-rank adaptation of large language models. In *Proc. International Conference on Learning Representations* (2022).
38. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *Proc. International Conference on Learning Representations* (2019).
39. Liu, A. et al. DeepSeek-V3 technical report. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2412.19437 (2024).
40. Ahmed, S., Hossain, M. A., Ray, S. K., Bhuiyan, M. M. I. & Sabuj, S. R. A study on road accident prediction and contributing factors using explainable machine learning models: ANAlysis and performance. *Transp. Res. Interdiscip. Perspect.* **19**, 100814 (2023).
41. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16* 785–794. https://doi.org/10.1145/2939672.2939785 (Association for Computing Machinery, 2016).
42. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
43. Quinlan, J. R. Induction of decision trees. *Mach. Learn.* **1**, 81–106 (1986).
44. Freund, Y. & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139 (1997).
45. Cox, D. R. The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **20**, 215–232 (1958).
46. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. In *Proc. Advances in Neural Information Processing Systems* Vol. 31 (Curran Associates, Inc., 2018).

47. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* 4171–4186 (Association for Computational Linguistics, 2019).
48. Arik, S. Ö. & Pfister, T. Tabnet: Attentive interpretable tabular learning. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 35, 6679–6687 (AAAI Press, 2021).
49. Transportation Research Board and National Academies of Sciences, Engineering, and Medicine, Lord, D., Geedipally, S., Pratt, M. P., Park, E. S., Khazraee, S. H. & Fitzpatrick, K. *Safety Prediction Models for Six-Lane and One-Way Urban and Suburban Arterials* (The National Academies Press, 2022).
50. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. Advances in Neural Information Processing Systems* Vol. 35, 24824–24837 (Curran Associates, Inc., 2022).
51. Chen, H., Lundberg, S. M. & Lee, S.-I. Explaining a series of models by propagating shapley values. *Nat. Commun.* **13**, 4512 (2022).
52. Chen, H., Covert, I. C., Lundberg, S. M. & Lee, S.-I. Algorithms to estimate shapley value feature attributions. *Nat. Mach. Intell.* **5**, 590–601 (2023).
53. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Proc. Advances in Neural Information Processing Systems* (eds Guyon, I. et al.) Vol. 30 (Curran Associates, Inc., 2017).
54. Goldshmidt, R. & Horovicz, M. TokenSHAP: interpreting large language models with monte carlo shapley value estimation. In *Proc. 1st Workshop on NLP for Science (NLP4Science)*, 1–8 (Association for Computational Linguistics, Miami, FL, USA, 2024).
55. Kokalj, E., Škrlj, B., Lavrač, N., Pollak, S. & Robnik-Šikonja, M. BERT meets shapley: extending SHAP explanations to transformer-based classifiers. In *Proc. EACL Hackashop on News Media Content Analysis and Automated Report Generation* (eds Toivonen, H. & Boggia, M.) 16–21 (Association for Computational Linguistics, 2021).
56. Zhao, Y., Wang, P., Zhao, Y., Du, H. & Yang, H. F. SafeTraffic Decision Copilot: adapting large language models for trustworthy safety assessments and policy interventions (this paper). Puw242/SafeTraffic: Release for paper version, Zenodo, https://doi.org/10.5281/zenodo.16896764 (2025).

Acknowledgements

The authors would like to thank Dr. Wei Zhang of the Federal Highway Administration (FHWA) and Jeffrey P. Michael, EdD, of the Bloomberg School of Public Health at Johns Hopkins University, for their valuable input on improving *SafeTraffic Copilot*. The authors are also grateful to the Office of Safety and Operations R&D, Dr. Carol Tan, and Jessica G. Rich of the FHWA for their generous support in providing the Highway Safety Information System (HSIS) data.

Author contributions

Y.Z., P.W. and H.F.Y. conceptualized and designed the study. P.W. and Yibo Z. collected data. P.W. and Yibo Z. processed the data and designed prompts. Y.Z. and P.W. performed experiments. Yibo Z. run the baseline models. Y.Z., P.W., Yibo Z. and H.F.Y. prepared the figures. Y.Z., P.W., Yibo Z. and H.F.Y. analyzed results. Y.Z., P.W., Yibo Z. and H.F.Y. wrote the initial draft. H.D. and H.F.Y. provided guidance and feedback for the study. H.D. and H.F.Y. revised the manuscript. H.F.Y. acquired the funding. H.F.Y. provided computational resources. All authors prepared the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-64574-w.

Correspondence and requests for materials should be addressed to Hao Frank Yang.

Peer review information *Nature Communications* thanks Dungar Singh and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.