Enhancing Bayesian Network with LLM Semantic Inference for Traffic Accident Severity Prediction

Zirong Feng
20717372
scxzf1@nottingham.edu.cn
School of Computer Science
University of Nottingham Ningbo China

***Abstract*—The Bayesian network has certain limitations in the real-world traffic accident dataset: its causal inference is heavily dependent on the conditional probability table (CPT). And CPT will become unreliable in the case of serious category imbalance and sparse data. This structural defect can lead to inaccurate prediction of the severity of the accident. The latest progress of the large language model (LLM) provides an opportunity to solve this limitation.**

This study proposes a selective fusion framework based on the number of samples. The framework integrates the semantic CPT values generated by LLM into the Bayesian Belief Network (BBN). This method only replaces CPT entries with insufficient sample size combinations, while retaining the original BBN estimates with sufficient sample size combinations. We have developed a fully automatic CPT extraction, semantic prompts, selective fusion and evaluation process using the British multi-year traffic accident dataset.

The experimental results show that the method has made significant improvements compared to the baseline BBN. The recall rate of the rarest severity category has increased by more than 80 times, and the Macro F1 Score and Balanced Accuracy have also improved significantly. These findings show that semantic CPT can effectively alleviate data sparsity without affecting causal explainability. The potential of combining probability reasoning with semantic inference based on LLM to improve the prediction of accident severity is highlighted.

***Index Terms*—Bayesian Belief Network, Conditional Probability Table, Large Language Model, Traffic Accident Severity, Data Sparsity**

I. Introduction

According to the World Health Organization (WHO), road traffic crashes result in the deaths of approximately 1.19 million people around the world each year [1]. These incidents pose serious threats not only to public safety and property but also to the stability and sustainable development of urban transport systems. The causes of traffic accidents are multi-faceted, involving external factors such as weather and road conditions, as well as human factors including driving behavior and compliance with traffic rules. To prevent accidents effectively and design targeted interventions, developing accurate and interpretable risk prediction models has become a critical task in modern traffic safety management.

Bayesian Belief Networks (BBNs) are attractive for this task because they provide a clear causal structure and allow probabilistic reasoning under uncertainty. A BBN represents variables as nodes in a directed acyclic graph, and conditional probability tables (CPTs) specify the strength of causal relationships between the parent and child nodes. In practice, CPTs are usually estimated from historical data. However, real-world traffic datasets contain strong class imbalance and many low-frequency parent–state combinations. As a result, some CPT entries are estimated from extremely limited samples. These unreliable parameters weaken the causal reasoning ability of the BBN.

Large Language Models (LLMs) offer a new opportunity to solve this limitation. Because they are trained on large text corpora, LLMs encode broad semantic knowledge and can produce reasonable probability estimates when prompted in natural language. These semantic estimates can supplement the BBN parameters in cases where data are sparse.

This study explores how to integrate the probability generated by LLM into the Bayesian Belief Network (BBN) to improve the accuracy of accident severity prediction. We proposed a selective fusion method based on sample size. For each CPT entry, we check the number of parent-state combinations. If the sample size is small, replace the BBN estimate with the probability generated by LLM; if the sample size is sufficient, the original BBN estimate is retained. This method enhances the weakest part of the BBN - the sparse area, while retaining the causal structure and parameters with good support.

We evaluated the performance of the fusion network at different percentiles and tested it with accident data from different years. The results show clear improvements in minority-class detection and overall predictive performance. These findings suggest that combining BBNs with LLM semantic knowledge can reduce the impact of data sparsity and enhance severity prediction under real-world conditions.

II. Related Work

In recent years, traffic safety research has widely used machine learning technology to analyze accident data. Behboudi et al. [2] reviewed 191 papers on the risk, frequency, severity and duration of accidents published in recent years. They pointed out that methods such as deep learning, integrated models and other advanced algorithms can achieve high prediction accuracy. But at the same time, many such models are difficult to explain and fail to clearly show how

road, environment and human factors work together to cause accidents.

Bayesian networks have been used in the field of traffic security to satisfy the need for interpretation. Yuan et al. [3] built an explainable Bayesian network of traffic events and used it to study how weather conditions affect the risk of accidents, so that users can understand the changes in the probability of collision in different situations. Carrodano et al. [4] uses data-driven Bayesian networks to analyze nonlinear interactions between geometric, traffic and environmental variables. Sulaie et al. [5] uses the Bayesian network with sensitivity analysis to study how different variables affect the consequences of collisions. Zahran et al. [6] combined causal findings and reasoning based on the Bayesian network to support road safety assessment, which helps identify meaningful causal relationships in collision data. These studies confirm that the Bayesian network can effectively model complex transportation systems.

At the same time, there is growing interest in combining large language models(LLM) with Bayesian methods. Nafar et al. [7] proved that the conditional probability can be directly asked in the large language model, and these answers can be used to fill or improve the conditional probability tree (CPT) of the existing Bayesian network, especially in the case of a small amount of data. Babakov et al. [8] use LLM to assist in designing the Bayesian network (BN) structure, allowing the model to propose candidate links between variables, and then comparing these links with expert knowledge or data. Feng et al. [9] proposed the BIRD framework, which processes the output of LLM within the Bayesian reasoning framework, so that the probability generated by the model is better calibrated and more consistent with the standard uncertainty reasoning.

In summary, existing Bayesian network applications in the field of road safety still mainly rely on data-driven conditional probability tables (CPT). And existing LLM-BBN fusion research has not been optimized for data sets with serious category imbalances and containing real accidents. Based on the above research, this study uses the probability generated by LLM to selectively replace the CPT value with low dataset support, while keeping the reliable value of the dataset unchanged. To improve the prediction accuracy of the severity of the accident under the condition that the data is sparse.

III. Methodology

This chapter discusses various aspects, including data collection, data processing, Bayesian Network construction, Prompt Engineering with LLM and dynamic fusion, starting with data collection.

A. Data Collection and Reprocessing

This study uses the 2024 UK traffic accident records obtained from the official Road Safety Data system [10], which contains three relational data sets: collision data, vehicle data and casualty data. Each document contains specific details, including the victim's basic information, vehicle characteristics, driving operations, road characteristics and environmental factors. These three files are linked by the public identifier "collision_index". After the merger, we removed variables that were not related to the prediction of the severity of the accident (for example, vehicle brand, driver's occupation) and eliminated invalid or missing values to ensure data quality.

TABLE I
Basic Information of Data Files

File Name	Sample Size	Feature Size
Safety Data - Collision 2024	100,928	44
Safety Data - Vehicles 2024	183,515	32
Safety Data - Casualties 2024	128,273	23

A prominent feature of this data set is its extremely unbalanced category distribution (for target node collision severity: Class 3: 72%, Class 2: 26%, Class 1: 2%). This imbalance means that many parent-state combinations appear very frequently in the sample data. These structural characteristics directly prompt us to adopt selective fusion strategies in subsequent experiments.

B. Feature Selection and Variable Identification

In the past, research using Bayesian networks for traffic safety analysis has provided valuable guidance for variable selection. For example, Mujalli et al. [11] emphasized that environmental and road factors, such as weather, lighting and road types, should be incorporated when predicting the severity of injury, and pointed out that these situational variables have a significant impact on collision results. Similarly, Joo et al. [12] further integrated driver behavior variables (such as vehicle operation, driver age, driver gender) to improve the accuracy of individualized risk prediction.

In order to ensure interpretability and causal correlation, we follow the previous security modeling research based on the Bayesian network and prune related features. Among the initial 97 variables, we retained 13 key variables that were agreed in the literature to have an impact on the severity of the injury:

- weather conditions, light conditions, road surface conditions, road type, speed limit, urban or rural area, sex of driver, age of driver, vehicle manoeuvre, casualty type, number of vehicles, number of casualties, and collision severity.

C. Bayesian Network Construction

The Bayesian Belief Network (BBN) is built in a two-stage process in GeNIe. First, use a greedy Bayesian search to find the conditional dependencies between variables and generate a directed acyclic graph. The diagram describes how environmental, road and behavioral factors together affect the severity of the collision. Secondly, use maximum likelihood estimation for parameter learning to generate a data-driven conditional probability table (CPT). Its entries directly reflect the frequency of experience observed in the data set in 2024.

The BBN is represented by breaking down the joint probability distribution $P(X_1, X_2, \ldots, X_n)$ on all variables into a local conditional model. Formally, for the variable Xi with the parent set of Pa(Xi), its factorization is as follows:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i)) \tag{1}$$

Each factor $P(X_1 \mid Pa(X_i))$ describes the changes in the probability of each state of Xi under different parent-state conditions. This representation method can achieve efficient probability reasoning. It allows the network to dynamically update the severity forecast of the accident when new evidence (such as weather, speed limit or road type) is introduced.

D. Prompt Engineering with LLMs

We use the Large Language Model (LLM), specifically Chatgpt-4o, to realize the parallel generation mechanism of semantic conditional probability tables. For each conditional entry in the conditional probability table (CPT), we programmatically convert the corresponding parent-child relationship into natural language queries. Each prompt follows a fixed template, which aims to lead to subjective probability estimates from LLM while minimizing ambiguity.

Prompt template:

- "You are a traffic safety expert. Respond only with a probability between 0 and 1. Under the condition of [Parent States], what is the probability of [Target_State]?"

The prompt information is automatically generated according to each line in the conditional probability table. Among them [Parent States] reflects a specific combination of observation characteristics (for example, weather, road type, driver information). Then, the LLM response is parsed and normalized into a value (p_LLM) to construct semantic CPT, which can be used as an alternative probability value for comparison or integration with CPT_BBN.

E. Dynamic CPT Fusion Based on Sample Percentile Thresholding

We introduced a selective fusion mechanism based on percentile. For each CPT row, calculate the sample count n. It indicates how much sample size matches the corresponding parent-state configuration. This empirical support can be used as a reliability indicator of CPT_BBN. In order to identify the data sparse area, we evaluated a set of percentiles $p \in \{0\%, 10\%, 20\%, \ldots, 90\%, 100\%\}$. For a given percentile p, the corresponding threshold determines the maximum n-sample value of the row. Any CPT entry with a sample value less than or equal to this percentile threshold will be replaced with CPT_LLM, and all remaining entries will retain their original CPT_BBN.

Formally, the fused CPT is obtained through a selective replacement rule:

$$\mathrm{CPT}_{\text{fused}}(i) = \begin{cases} \mathrm{CPT}_{\text{LLM}}(i), & \text{if } n(i) \le \tau_p, \\ \mathrm{CPT}_{\text{BBN}}(i), & \text{otherwise.} \end{cases}$$

where τ_p denotes the pth percentile of the empirical sample counts.

After replacement, each CPT row is renormalized in its parent state group to ensure the validity of the conditional distribution. This percentile selective fusion allows LLM to play a role mainly in the part where the sample number is sparse, while retaining the frequency-based estimation of BBN in the part where the data is sufficient.

F. Automated Python-GeNIe/PySMILE Pipeline to facilitate Multi-Year Validation

We have developed a fully automatic processing and reasoning process. The selective fusion Bayesian network can be systematically evaluated and its stability on multiple time datasets can be verified. The process uses Python and integrates with GeNIe through the PySMILE module. The baseline network uses the accident dataset in 2024 for training. The verification is carried out using an external dataset containing the 2023 collision record released by the same national agency. This multi-year verification framework can evaluate the predictive robustness and consistency of fusion CPT values.

The automated workflow first loads the Bayesian Belief Networks structure (e.g., bbn_base.xdsl) exported from GeNIe. And then replaces the CPT with the selective fusion probability table generated for each percentile setting through the program. Each fusion network variant is saved as an independent .xdsl model. Then, use Python's pandas library to process the external verification data set stored in CSV format. The standardized mapping process converts the original table attributes into the corresponding discrete node state required by the Bayesian reasoning engine.

For each record in the validation dataset, the observed characteristic values will be inserted into the network as evidence in turn. PySMILE's reasoning engine then performs a confidence update through the "update_beliefs( )" function to generate the posterior distribution of the target severity variable. The severity class of the prediction will be written back to the verification dataframe to generate a complete prediction log. This process will automatically repeat for all fusion models to ensure that each CPT configuration (corresponding to the percentile $p \in \{0\%, 10\%, 20\%, \ldots, 90\%, 100\%\}$ can be evaluated under the same conditions.

RESULTS AND DISCUSSION

We used the GeNIe software platform to conduct structure and parameter learning. Fig. 1 illustrates the final structure of the Bayesian Belief Network (BBN). It captures a set of directed dependencies among environmental, road, vehicle, and demographic variables that collectively influence collision severity.

Through maximum likelihood estimation, the Conditional Probability Tables (CPTs) associated with each node were estimated from the 2024 training dataset. These CPTs, denoted as CPT_BBN, form the empirical foundation of the baseline model. Fig.2 illustrates the CPT

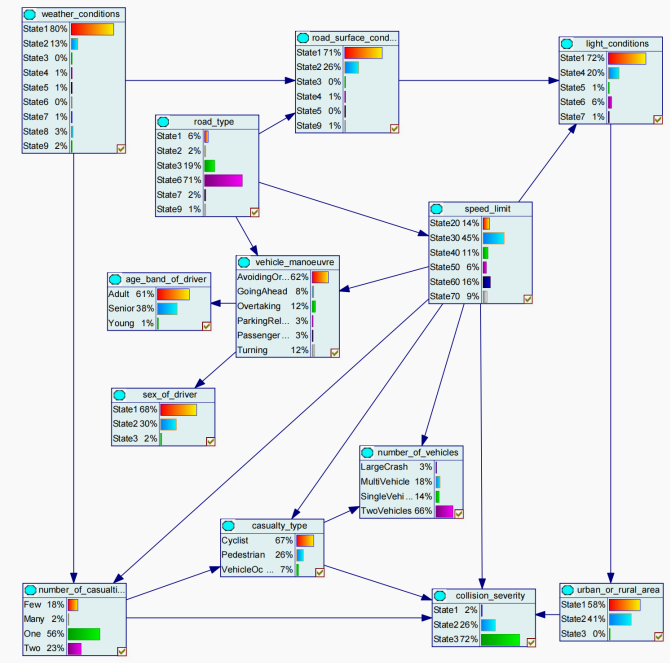

Fig. 1. Final structure of the Bayesian Belief Network

road_type	State1								
weather_conditions	State1	State2	State3	State4	State5	State6	State7	State8	State9
State1	0.858998...	0.013408...	0.083333...	0.553160...	0.001851...	0.166666...	0.064625...	0.689121...	0.495993...
State2	0.1300169	0.983488...	0.1547619	0.441091...	0.990740...	0.166666...	0.860544...	0.237025...	0.120993...
State3	0.000211...	0.000110...	0.583333...	0.001436...	0.001851...	0.166666...	0.003401...	0.003493...	0.000801...
State4	0.008555...	0.000110...	0.1547619	0.001436...	0.001851...	0.166666...	0.044217...	0.057385...	0.010416...
State5	0.000105...	0.000775...	0.011904...	0.001436...	0.001851...	0.166666...	0.003401...	0.000499...	0.000801...
State9	0.002112...	0.002105...	0.011904...	0.001436...	0.001851...	0.166666...	0.023809...	0.012475...	0.370993...

Fig. 2. Example of Conditional Probability Table of BBN

values of road_surface_conditions under road_type and weather_conditions.

The network serves as a fully functional inference engine. It can update the posterior beliefs about the severity of the collision when evidence is provided about the road environment, traffic conditions, and human factors. This baseline BBN represents the purely data-driven model against which all fused variants are compared.

Then we use a large language model (Chatgpt-4o) to generate semantic conditional probabilities for each unique combination of parent-state. Each prompt follows a structured natural-language prompt, describing the specific conditions (e.g., weather, lighting, road surface, speed limit). Then the LLM returns a single numerical probability representing its subjective assessment of the likelihood of each severity class under the specified conditions.

Structured CPT prompts are shown in table 2:

For every parent-state combination in the CPT, an n-sample value was computed to count how many records in the 2024 dataset matched that exact combination. The distribution of n-samples shows that the data density distribution in CPT exhibits a highly uneven characteristic.

In order to understand the root cause of the performance difference, Figure 3 shows the number of samples associated

TABLE II
EXAMPLE OF STRUCTURED CPT PROMPT ENTRY

Target Node	speed_limit
Target State	State50
Parent State	road_type_One way street
P_BBN	0.0039
P_LLM	0.1428
Prompt	You are a traffic safety expert. Respond only with a probability between 0 and 1. Given that road type is One way street, what is the probability that speed limit is 50?

with each CPT line. Specifically refers to the number of training samples in each unique parent-state. Overall, it shows an extremely skewed distribution. At the lower percentile (0-10%), the corresponding number of samples is almost zero, indicating that a considerable part of the parent state has almost never appeared in the actual data.

Even in the 30% and 40%, the sample size is still very low ($n \approx 26$–108), which confirms the existence of a large data sparse area. In contrast, the upper tail grew sharply: the 90% reached 12,797 samples, with a maximum value of more than 145,000.

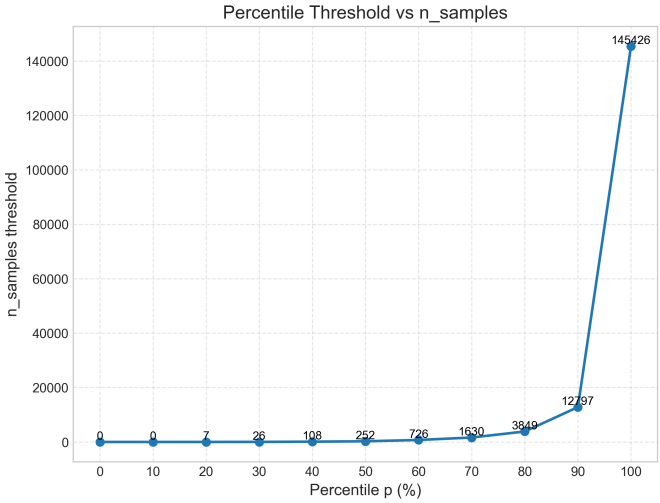

Fig. 3. Distribution of n_samples in CPT Rows at Different Percentile

In order to strictly evaluate the performance of the model, we used an external dataset containing 213718 accident records in 2023 to verify the prediction results of the hybrid Bayesian network based on the 2024 data set training. We have calculated multiple performance indicators, including macro F1 values, balance accuracy and class recall rates. These indicators evaluate both the overall predictive performance of the model and the ability of the model to handle highly unbalanced accident severity class.

Figure 4 shows the Macro F1 and Balanced Accuracy under the fusion ratio $p \in [0, 100]$. There are several important trends in the figure: when p exceeds 70%, the macro F1 value rises sharply to 0.3576. When p = 90%, the F1 value reaches

the highest value of 0.3683, which is 10.6% higher than the baseline value. The Balanced Accuracy rate also shows a moderate but stable improvement, reaching a peak of 0.3719 when p = 70%, and stable at about 0.364–0.366 in the range of $p \in [80, 100]$. Compared with the baseline value of 0.355, this increase reflects the improvement of class recall balance.

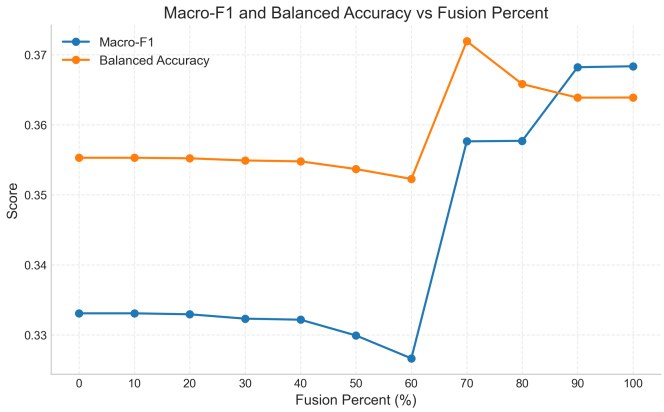

Fig. 4. Macro-F1 and Balanced Accuracy under Varying CPT Fusion Ratios

As a supplement to the overall evaluation indicators, Figure 5 shows the class recall rate under different fusion ratios. Under the baseline BBN model, the recall rate of class 1 is extremely low (0.0008), which reflects the lack of supporting information in the sample data. With the gradual integration of CPT_LLM, the recall rate has increased significantly, reaching 0.072 at the highest integration ratio. Although the absolute value is still not high, it is equivalent to an increase of about 86 times. This shows that when the sample data is insufficient, semantic knowledge can provide effective corrections.

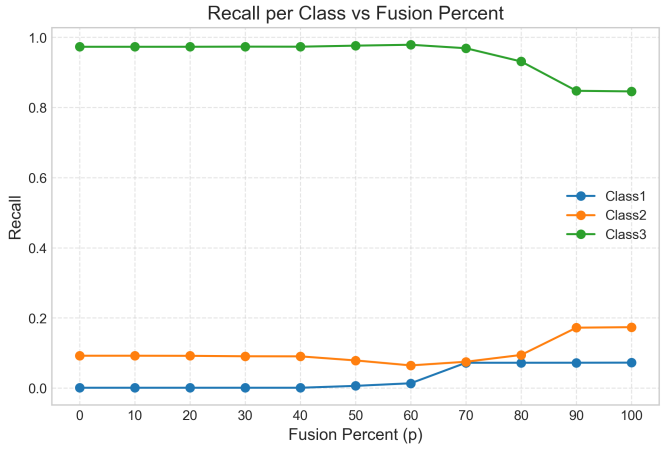

Fig. 5. Recall for Each Severity Class under Varying CPT Fusion Ratios

The increase in the Class 2 is relatively moderate, but it is still significant. The recall rate increased from 0.092 to 0.174 in the baseline model. This is a relative increase of nearly 90%. In contrast, the recall rate of the Class 3 gradually decreases with the increase of the fusion ratio, to about 0.85 when p = 100%. This decline reflects the expected redistribution of probability quality: as the model's sensitivity to a few categories of samples increases, the number of samples assigned to the majority categories of samples decreases.

Among all the evaluation indicators, the selective fusion strategy significantly improved the performance of the severity class with insufficient samples. When p = 90%, Recall_Class1 increased by 8659% and Recall_Class2 increased by 86.8%. F1 and Balanced Accuracy also increased by 10.6% and 2.4% respectively. This shows that the overall class balance has been improved without affecting the performance of most class. These results show that CPT_LLM can effectively make up for data sparsity and significantly enhance the ability of the Bayesian Belief Network to detect the results of high-risk accidents.

Conclusions

This study proposes a selective fusion framework that integrates the semantic probability generated by the large language model (LLM) into the Bayesian Belief Network (BBN) structure. It is used to predict the severity of traffic accidents. In view of the structural sparsity of real-world accident datasets, this method allows LLM to provide probability estimates with very few dataset samples. When the data are rich, the model still depends on the actual data. Specifically, it includes systematic processes, including prompt engineering, semantic CPT generation, dynamic integration, and external data verification. This study proves the feasibility and significant performance improvement of this method.

The results show that selective replacement of CPT entries with semantic estimates of LLM can significantly improve the ability of the severity class of insufficient data. The recall rate of Class 1 has increased by more than 80 times, and the recall rate of Class 2 has also nearly doubled. The Macro F1 value has been increased by 10.6%, and the Balance Accuracy has been improved by 2.4%, which reflects the enhancement of the prediction balance of all class. These improvements verify the core hypothesis of this article: CPT_LLM can effectively make up for the learning limitations of CPT_BBN in areas where training data is insufficient or unbalanced.

In addition to empirical performance, the research results also highlight a broader methodological contribution. By retaining the original Bayesian network (BBN) structure, we only intervene at the level of probability parameters. The fusion mechanism can be operated without structural modification. It provides a scalable solution to enrich the Bayesian network.

However, the study also admits that there are some practical limitations. The probability generated by LLM may contain model-specific deviations, and semantic estimates may vary depending on the prompt design. In addition, this work only focuses on a single LLM. Future research can explore the multi-LLM consensus mechanism and dynamic calibration strategy.

In short, this study shows that integrating semantic knowledge in LLM into the Bayesian network is a feasible and effective strategy to improve the prediction of traffic accident severity. By combining causal interpretation with semantic reasoning, this method opens up a new direction for the construction of a stable accident prediction model.

REFERENCES

[1] World Health Organization, "Road safety." [Online]. Available: https://www.who.int/health-topics/road-safety

[2] N. Behboudi, S. Moosavi, and R. Ramnath, "Recent advances in traffic accident analysis and prediction: A comprehensive review of machine learning techniques," 2024, arXiv:2406.13968. doi: 10.48550/arXiv.2406.13968. [Online]. Available: https://arxiv.org/abs/2406.13968

[3] T. Yuan, J. Yang, and Z. Wen, "Interpretable Traffic Event Analysis with Bayesian Networks," Oct. 10, 2023, arXiv: arXiv:2310.06713. doi: 10.48550/arXiv.2310.06713. [Online]. Available: https://arxiv.org/abs/2310.06713

[4] C. Carrodano, "Data-driven risk analysis of nonlinear factor interactions in road safety using Bayesian networks," *Scientific Reports*, vol. 14, no. 1, p. 18948, Aug. 2024. doi: 10.1038/s41598-024-69740-6.

[5] S. A. Sulaie, "Sensitivity analysis of factors affecting consequences due to traffic crashes: A Bayesian network modelling," *J. Road Safety*, vol. 36, no. 1, pp. 21–30, Feb. 2025. doi: 10.33492/JRS-D-25-1-2442769.

[6] O. E. Zahran, Y. Xin, E. M. M. Zahran, and W. P. Cheah, "Enhancing road safety evaluation with AI: Causal discovery and reasoning in road traffic accident analysis," in *Proc. 2025 6th Int. Conf. Computer Vision, Image and Deep Learning (CVIDL)*, May 2025, pp. 701–706. doi: 10.1109/CVIDL65390.2025.11085562.

[7] A. Nafar, K. B. Venable, Z. Cui, and P. Kordjamshidi, "Extracting probabilistic knowledge from large language models for Bayesian network parameterization," 2025, arXiv:2505.15918. doi: 10.48550/arXiv.2505.15918. [Online]. Available: https://arxiv.org/abs/2505.15918

[8] N. Babakov, E. Reiter, and A. Bugarin, "Scalability of Bayesian network structure elicitation with large language models: A novel methodology and comparative analysis," 2024, arXiv:2407.09311. doi: 10.48550/arXiv.2407.09311. [Online]. Available: https://arxiv.org/abs/2407.09311

[9] Y. Feng, B. Zhou, W. Lin, and D. Roth, "BIRD: A trustworthy Bayesian inference framework for large language models," 2025, arXiv:2404.12494. doi: 10.48550/arXiv.2404.12494. [Online]. Available: https://arxiv.org/abs/2404.12494

[10] Department for Transport. Road Safety Data. [Online]. Available: https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-accidents-safety-data

[11] R. Mujalli, G. López, and L. Garach, "Bayes classifiers for imbalanced traffic accidents datasets," *Accident Analysis & Prevention*, vol. 88, pp. 37–51, Mar. 2016. doi: 10.1016/j.aap.2015.12.003.

[12] Y.-J. Joo, S.-Y. Kho, D.-K. Kim, and H.-C. Park, "A data-driven Bayesian network for probabilistic crash risk assessment of individual driver with traffic violation and crash records," *Accident Analysis & Prevention*, vol. 176, p. 106790, Oct. 2022. doi: 10.1016/j.aap.2022.106790.