Enhancing Bayesian Networks 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

1 Introduction

Road traffic accidents remain a major global public health challenge. According to the World Health Organization (WHO), approximately 1.19 million people lose their lives in road crashes every year, and tens of millions suffer non-fatal injuries [1]. These incidents pose substantial threats to public safety, economic stability, and the sustainable development of urban transport systems. Understanding and predicting accident severity is therefore a critical task that supports risk assessment, resource allocation, and targeted policy interventions. However, the underlying mechanisms of traffic accidents are complex, influenced by intertwined environmental, infrastructural, and human-behavioral factors. This complexity requires predictive models that are both accurate and interpretable.

Bayesian Belief Networks (BBNs) have emerged as a promising solution for this problem due to their ability to represent explicit causal structures and perform probabilistic reasoning under uncertainty. A BBN encodes dependencies through directed edges, and Conditional Probability Tables (CPTs) quantify how parent variables influence child nodes. In practice, these CPTs are learned from historical datasets. Yet real-world accident data exhibit severe class imbalance and a highly skewed distribution of parent-state combinations. Many CPT entries are estimated from extremely small sample sizes, resulting in unreliable parameters that degrade the inference capability of the network, especially for rare but high-risk outcomes.

Recent advances in Large Language Models (LLMs) provide a new opportunity to address this fundamental limitation. Because LLMs are trained on large and diverse text corpora, they implicitly encode broad semantic knowledge about road environments, human behavior, and safety-related phenomena. When queried using natural-language prompts, LLMs can produce reasonable probability estimates even for conditions that are sparsely represented—or entirely absent—in the empirical dataset. These semantic estimates offer a potential remedy for the weakly supported regions of BBN CPTs.

This study investigates how LLM-generated semantic probabilities can be integrated into a Bayesian Belief Network to improve accident severity prediction under real-world data sparsity. We propose a sample-size-aware selective fusion framework that replaces only those CPT entries with insufficient empirical support while preserving the original BBN values where data are reliable. Specifically, the method computes the sample count associated with each parent-state configuration and uses percentile thresholds to determine which CPT rows should be substituted with LLM-derived probabilities. This strategy strengthens the weakest part of the BBN without altering its causal structure or compromising interpretability.

We evaluate the fused networks across multiple percentile thresholds and validate their performance using accident data from different years. Experimental results demonstrate substantial improvements in minority-class recall, Macro-F1 score, and balanced accuracy, confirming that LLM semantic knowledge can effectively mitigate data sparsity and enhance predictive performance under realistic conditions. These findings highlight the potential of combining probabilistic reasoning with LLM-based semantic inference to develop more robust and interpretable traffic accident prediction models.

2 Related Research

Research on traffic accident analysis has increasingly adopted machine learning techniques to model the complex relationships among environmental, infrastructural, and human-behavioral factors. Behboudi et al. [2] conducted a comprehensive review of 191 studies on accident risk, frequency, severity, and duration. Their findings highlight that advanced algorithms—including deep learning and ensemble models—achieve high predictive accuracy.

Bayesian Networks (BNs) have emerged as a strong candidate for interpretable traffic safety modeling. Yuan et al. [3] constructed an explainable BN to quantify the influence of weather conditions on collision risk, enabling users to visualize how probability distributions shift under different scenarios. Carrodano et al. [4] explored nonlinear interactions among geometric, traffic, and environmental factors using data-driven BNs, demonstrating their suitability for capturing complex multi-variable dependencies. Sulaie et al. [5] employed BN-based sensitivity analysis to determine the factors most strongly associated with collision outcomes. Zahran et al. [6] further incorporated causal discovery techniques to refine BN structures for road safety assessment, offering deeper causal interpretations of accident data. Collectively, these studies confirm the value of BNs as interpretable and flexible tools for modeling transportation systems.

Parallel to these developments, recent progress in Large Language Models (LLMs) has motivated research into integrating semantic reasoning with Bayesian methods. Nafar et al. [7] demonstrated that LLMs can provide direct estimates of conditional probabilities, which can be used to populate or refine BN CPTs, particularly in small-data scenarios. Babakov et al. [8] showed that LLMs can assist in BN structure elicitation by proposing candidate dependencies that can be validated against data or expert judgment. Feng et al. [9] introduced the BIRD framework, which calibrates LLM-generated probabilities within a Bayesian inference pipeline to produce more reliable and uncertainty-aware estimates. These works illustrate the potential of LLMs to enhance BN-based reasoning through their embedded semantic knowledge.

Despite these advances, two major limitations remain. First, existing BN applications in road safety overwhelmingly rely on data-driven CPTs, which become unreliable when parent-state combinations appear infrequently—as is common in highly imbalanced accident datasets. Second, current LLM–BN integration methods typically replace or adjust CPT values without differentiating between well-supported and sparsely supported regions of the data. As a result, they lack mechanisms for selective, sample-aware fusion, which is critical for real-world accident datasets characterized by extreme skewness.

To address these gaps, this study proposes a selective CPT replacement framework that uses LLM-generated semantic probabilities only for CPT entries with insufficient empirical support, while preserving the original data-driven estimates where sample counts are adequate. This approach leverages the complementary strengths of BNs and LLMs to improve accident severity prediction under severe data sparsity.

3 Aims and Objectives

The primary aim of this project is to develop and evaluate a sample-size-aware semantic fusion framework that integrates Large Language Model (LLM)–generated conditional probabilities into Bayesian Belief Networks (BBNs) in order to enhance accident severity prediction under real-world data sparsity.

To achieve this aim, the project sets out the following specific objectives:

Objective 1: Data Preparation and Integration

Collect, preprocess, and integrate multi-year UK Road Safety Data, ensuring consistency across collision, vehicle, and casualty datasets. This includes cleaning invalid entries, resolving missing values, and mapping categorical fields to discrete states suitable for Bayesian inference.

Objective 2: Construction of Baseline Bayesian Network

Develop a baseline BBN using structure learning (e.g., greedy Bayesian search) and maximum likelihood estimation within the GeNIe modeling environment. Extract the full CPTs programmatically to serve as the empirical foundation for later comparison and fusion.

Objective 3: Semantic CPT Generation via LLMs

Design structured natural-language prompt templates and automatically query an LLM (such as ChatGPT-4o) to obtain semantic probability estimates for each parent–state configuration. Normalize and store these semantic CPT values for downstream fusion.

Objective 4: Selective CPT Fusion Based on Sample Percentiles

Implement a percentile-based thresholding mechanism to identify CPT entries with insufficient empirical support. Replace these sparse-data CPT values using LLM-generated probabilities while retaining original data-driven values when sample sizes are adequate. Renormalize the fused CPTs to maintain valid conditional distributions.

Objective 5: Multi-Year Predictive Evaluation

Evaluate the fused BBN models using accident records from a different year (e.g., 2023) to measure cross-year generalization. Compare performance across fusion thresholds ($p \in \{0, 10, \ldots, 100\}$) using Macro-F1, Balanced Accuracy, and class-wise recall metrics, with special attention to minority severity classes.

Objective 6: Analysis of Model Impact, Robustness, and Interpretability

Assess how the selective semantic fusion influences predictive behavior, minority-class sensitivity, and causal interpretability. Analyze whether LLM-enhanced CPTs stabilize or alter inference patterns across varying conditions and discuss implications for real-world traffic safety modeling.

4 Project Plan

This project will be carried out across 11 structured tasks that follow the complete lifecycle of the research workflow—from conceptualization to final reporting. Each task produces a concrete output that contributes to the final dissertation.

Task 1 — Conduct Literature Review and Define Research Problem

Review recent studies on traffic accident modelling, Bayesian Belief Networks, sparse CPT problems, and LLM-generated probabilistic knowledge. Identify methodological limitations and clearly formulate the research gap addressed in this project.

Task 2 — Acquire and Consolidate Multi-year Accident Datasets

Collect the UK Road Safety Data for 2023 and 2024, including collision, vehicle, and casualty tables. Align schemas, merge datasets using collision index, and prepare them for preprocessing.

Task 3 — Perform Data Cleaning and Variable Selection

Remove invalid entries, discretize categorical features, and prune irrelevant variables based on domain knowledge and prior studies. Retain the 13 critical variables for Bayesian modeling.

Task 4 — Learn and Construct the Bayesian Network Structure in GeNIe

Apply greedy Bayesian search to learn the Directed Acyclic Graph (DAG) and verify its interpretability. Finalize and export the BBN structure for parameter learning.

Task 5 — Train and Export Conditional Probability Tables (CPTs)

Perform Maximum Likelihood Estimation (MLE) on all nodes to generate CPTs. Export the full CPT set from the .xdsl model into a structured CSV for further analysis and fusion.

Task 6 — Design Prompt Templates and Generate LLM Semantic Probabilities

Create automated natural-language prompts based on parent–state combinations and query ChatGPT-4o to obtain semantic probability estimates for each CPT entry.

Task 7 — Compute Sample Support n for Each CPT Entry

For every row of the baseline CPT, calculate the empirical sample count n from the 2024 dataset. Use these values to identify sparsely supported CPT entries.

Task 8 — Perform Selective CPT Fusion Using Percentile Thresholds

Define percentile thresholds τ_p and selectively replace weak CPT entries with LLM-generated probabilities while retaining well-supported empirical values. Reconstruct the fused CPTs and generate updated BBN models.

Task 9 — Build Automated PySMILE Inference and Evaluation Pipeline

Implement a batch inference script that loads each fused BBN, inserts 2023 evidence, performs reasoning, and outputs predictions for severity classification.

Task 10 — Evaluate Performance and Produce Visual Analysis

Compute Macro-F1, Balanced Accuracy, class-wise Recall, LogLoss, and generate visual performance curves across fusion percentiles. Analyze model behaviour and interpret improvements.

Task 11 — Write, Revise, and Submit the Final Research Report

Produce the full dissertation manuscript, including methodology, results, limitations, and future work. Prepare supplementary materials, figures, and appendices.

5 Gantt Chart Plan

The project is scheduled from 22 September 2025 to 11 December 2025, following an 11-task workflow. Each task spans one or more weeks as illustrated by the corresponding Gantt Chart.

1.Conduct literature review and define research problem
2.Acquire and consolidate multi-year datasets
3.Data cleaning and variable selection
4.Learn and construct Bayesian Network structure
5.Train and export CPTs
6.Design LLM prompts and generate semantic probabilities
7.Compute sample support n for CPT entries
8.Selective CPT fusion using percentile thresholds
9.Build PySMILE inference pipeline
10.Performance evaluation and visualisation
11.Report writing and submission

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