Incorporating Large Language Models (LLMs) into Transportation Safety Analytics

December 2024

Region 1:
New England University Transportation Center

Caiwen Ding, Ph.D. (PI)

Kai Wang, Ph.D. (Co-PI)

Jinwei Tang

Yuebo Luo

Tianxin Li, Ph.D.

University of Connecticut

In cooperation with U.S. Department of Transportation,
Office of the Assistant Secretary for Research and Technology (OST-R)

Grant #: 69A3552348301

TECHNICAL DOCUMENTATION

1. Project No.
161159

2. Government Accession No.
01904458

3. Recipient's Catalog No.

4. Title and Subtitle
Incorporating Large Language Models (LLMs) into Transportation Safety Analytics

5. Report Date
December 2024

6. Performing Organization Code
N/A

7. Author(s)
Jinwei Tang - ORCiD-0009-0004-3310-9456; Yuebo Luo - ORCiD- 0009-0005-1738-886X; Tianxin Li, Ph.D. ORCiD - 0000-0002-3061-8077; Kai Wang, Ph.D. ORCiD - 0000-0003-1452-4000; Caiwen Ding, Ph.D. ORCiD - 0000-0003-0891-1231

8. Performing Organization Report No.
N/A

9. Performing Organization Name and Address
New England University Transportation Center
181 Presidents Drive
University of Massachusetts - Amherst
Amherst, MA 01003

10. Work Unit No. (TRAIS)

11. Contract or Grant No.
69A3552348301

12. Sponsoring Agency Name and Address
United States Department of Transportation
Research and Innovative Technology Administration
1200 New Jersey Avenue, SE
Washington, DC 20590

13. Type of Report and Period Covered
Final Research Report

14. Sponsoring Agency Code
USDOT OST-R

15. Supplementary Notes
Report uploaded and accessible at the NEUTC Website (www.umass.edu/neutc)

16. Abstract

This project focuses on enhancing transportation safety through LLMs. Its goal is to improve crash data analysis by highlighting disparities in traffic safety, particularly for vulnerable road users. Leveraging the Connecticut Crash Data Repository, the project develops an LLM algorithm to automatically analyze crash patterns and generate visual reports. This innovative approach by our interdisciplinary team combines expertise in machine learning and transportation safety, ensuring a comprehensive approach to these challenges. The project's outcomes are expected to influence policy and planning decisions, fostering safer transportation systems.

<table>
<tr><td colspan="2"><b>17. Key Words</b>
LLMs, Traffic Safety, Vulnerable Road Users, Crash Visualization and Analysis</td><td colspan="2"><b>18. Distribution Statement</b>
No restrictions.</td></tr>
<tr><td><b>19. Security Classif. (of this report)</b>
Unclassified</td><td><b>20. Security Classif. (of this page)</b>
Unclassified</td><td><b>21. No. of Pages</b>
8</td><td><b>22. Price</b></td></tr>
</table>

About NEUTC

The New England Regional University Transportation Center (NEUTC) is a diverse, multidisciplinary consortium committed to addressing the pressing issue of traffic safety. In line with the Infrastructure Investment and Jobs Act (IIJA), our objective is to drive transformative research, education, and technology transfer to address critical traffic safety needs at a time when roadway fatalities are distressingly high.

Our research and educational activities at NEUTC are guided by four principal safety themes, each addressing a critical challenge in transportation safety. These themes capture the various integral components of the transportation system, focusing on technology, infrastructure, vehicles, and users with a commitment to public engagement. Our overarching theme is promoting safety, with the common underlying science being the study of behavioral, systemic, environmental, and mobility-driven factors on safety.

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, under 69A3552348301 from the U.S. Department of Transportation's University Transportation Centers Program. The U.S. Government assumes no liability for the contents or use thereof.

All matching funds were graduate tuition waivers.

Motivation

Improving access to safe and reliable transportation for all users has been identified as a critical priority for future transportation. This includes addressing challenges faced by groups who may have limited access to transportation facilities due to geographic or situational factors. Incorporating broad accessibility considerations into safety analytics helps ensure that transportation systems serve a wide range of users and supports sound decision-making and investments for building a safer and more sustainable system.

Traditional methods of traffic safety analysis and transportation decision-making often focus primarily on facilities used by motor vehicles. This can overlook the needs of active transportation users such as pedestrians and bicyclists, as well as other groups with differing mobility needs.

Crash data generated from police reports has long been used by State Departments of Transportation (DOTs) and related agencies to conduct safety analysis that informs project planning and prioritization. A deeper investigation of these data can uncover complex crash patterns and contributing factors, helping DOTs identify transportation safety concerns for all road users. This process requires extensive data mining across crash, person, and vehicle files as well as crash narratives, which is typically labor-intensive.

To reduce this burden, an algorithm and tool that can automatically analyze crash data and generate reports on active transportation safety are urgently needed by state DOTs. With the rapid advancement of Artificial Intelligence (AI), applying Large Language Models (LLMs) in crash data analysis can improve DOTs' capabilities to conduct traditional safety assessments, while also revealing patterns and disparities across transportation modes and crash types.

This project seeks to develop and implement an LLM-based tool within the Connecticut Crash Data Repository (CTCDR) to automatically query crashes, identify patterns, and generate visualization reports based on different user-defined inputs, particularly for crashes involving pedestrians, bicyclists, and older drivers.

Executive Summary

Crash data generated from police reports has been widely used by the State Departments of Transportation (DOTs) and other agencies in conducting transportation safety analyses to support project planning and prioritization. An in-depth analysis and investigation of crash data can help uncover the complex crash paradigms and contributing factors and help DOTs identify transportation safety issues related to VRUs to meet the requirements of the Bipartisan Infrastructure Law.

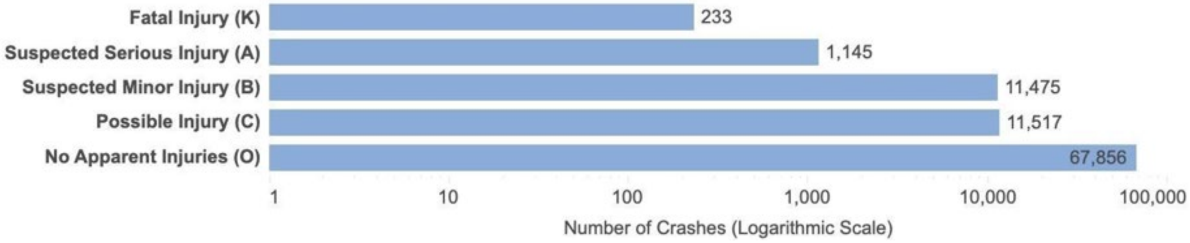


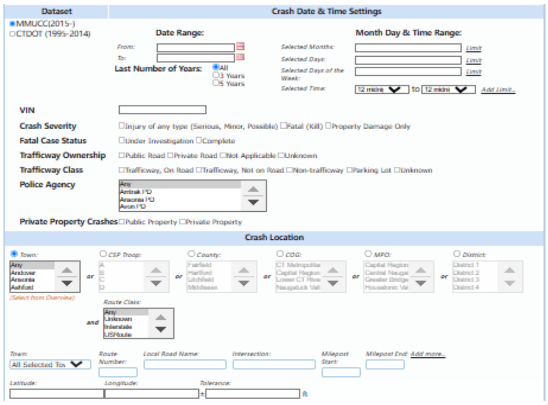
Figure 1: Crash Count by KABCO Injury Rating System. This figure is a demonstration of a graph that can be generated by a police man for in depth analysis on a crash scene.

During this process, a multi-angled view of the data is necessary to further enhance the report's quality. A better report on the traffic data will help DOTs make better decisions in changing protocols and further improve traffic safety.

Table 1 Crash Count by KABCO Injury Rating System. This table is linked with Figure 1.

Fatal Injury (K)	Suspected Serious Injury (A)	Suspected Minor Injury (B)	Possible Injury (C)	No Apparent Injuries (O)
233	1145	11475	11517	67856

However, this involves thorough mining and exploration of crash data from not just the crash files but also person and vehicle files. Therefore, an algorithm and corresponding tool that can automatically analyze crash data and generate VRU safety reports without manually going through the tedious filtering process in the existing tool are urgently needed to assist state DOTs. [2] Currently, there's a CRCDR website, which has an online query tool open to the public for whoever is interested in downloading and study the data in their own way. [1] However, Figure 2 shows this website has three main parts: basic information, categories, and options. Each



Dataset
● MMUCC(2015-)
○ CTDOT (1995-2014)

Crash Date & Time Settings

Date Range:
From:
To:
Last Number of Years: ● All ○ 3 Years ○ 5 Years

Month Day & Time Range:
Selected Months: Limit
Selected Days: Limit
Selected Days of the Week: Limit
Selected Time: 12 midni to 12 midni Add Limit...

VIN
Crash Severity □ Injury of any type (Serious, Minor, Possible) □ Fatal (Kill) □ Property Damage Only
Fatal Case Status □ Under Investigation □ Complete
Trafficway Ownership □ Public Road □ Private Road □ Not Applicable □ Unknown
Trafficway Class □ Trafficway, On Road □ Trafficway, Not on Road □ Non-trafficway □ Parking Lot □ Unknown
Police Agency: Any, Amtrak PD, Ansonia PD, Avon PD
Private Property Crashes □ Public Property □ Private Property

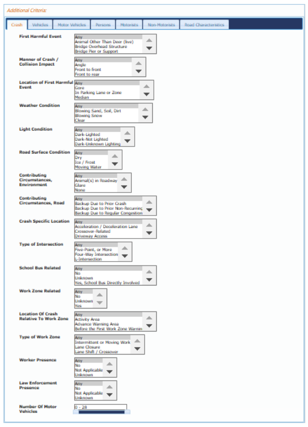
Crash Location

● Town: Any, Andover, Ansonia, Ashford (Select from Overview) or ○ CSP Troop: A, B, C, D or ○ County: Fairfield, Hartford, Litchfield, Middlesex or ○ COG: CT Metropolitan, Capital Region, Lower CT Rive, Naugatuck Vall or ○ MPO: Capital Region, Central Naugat, Greater Bridge, Housatonic Val or ○ District: District 1, District 2, District 3, District 4

and Route Class: Any, Unknown, Interstate, USRoute

Town: All Selected To
Route Number:
Local Road Name:
Intersection:
Milepost Start:
Milepost End: Add more...
Latitude:
Longitude: ±
Tolerance: ft

Figure 2: Basic Info. User must fill in this section to proceed, such as time, crash severity, fatal case status, etc. The page is composed with more than 30 boxes to be filled in for each query



Additional Criteria:

Crash | Vehicles | Motor Vehicles | Persons | Motorists | Non-Motorists | Road Characteristics

First Harmful Event: Any, Animal Other Than Deer (live), Bridge Overhead Structure, Bridge Pier or Support
Manner of Crash / Collision Impact: Any, Angle, Front to front, Front to rear
Location of First Harmful Event: Any, Gore, In Parking Lane or Zone, Median
Weather Condition: Any, Blowing Sand, Soil, Dirt, Blowing Snow, Clear
Light Condition: Any, Dark-Lighted, Dark-Not Lighted, Dark-Unknown Lighting
Road Surface Condition: Any, Dry, Ice / Frost, Moving Water
Contributing Circumstances, Environment: Any, Animal(s) in Roadway, Glare, None
Contributing Circumstances, Road: Any, Backup Due to Prior Crash, Backup Due to Prior Non-Recurring, Backup Due to Regular Congestion
Crash Specific Location: Any, Acceleration / Deceleration Lane, Crossover-Related, Driveway Access
Type of Intersection: Any, Five-Point, or More, Four-Way Intersection, L-Intersection
School Bus Related: Any, No, Unknown, Yes, School Bus Directly Involved
Work Zone Related: Any, No, Unknown, Yes
Location Of Crash Relative To Work Zone: Any, Activity Area, Advance Warning Area, Before the First Work Zone Warnin
Type of Work Zone: Any, Intermittent or Moving Work, Lane Closure, Lane Shift / Crossover
Worker Presence: Any, No, Not Applicable, Unknown
Law Enforcement Presence: Any, No, Not Applicable, Unknown
Number Of Motor Vehicles: 0 - 28

Figure 3: Additional Info. Don't have to fill all in this section since it defaults to Any. Note that there are multiple option tabs that can be selected.

category contains its options. If users want to analyze the crash data, they must select from these options and fill in the basic information to proceed. The basic information section contains essential filters such as the date and type of vehicle involved. And options contain some specific values from each column. Users must select from all categories to complete the query. However, they do have the option to leave an option "Any." This eventually leaves an unnecessary amount of data for download.

The data query tool by categories is classic but complicated for a beginner who has just

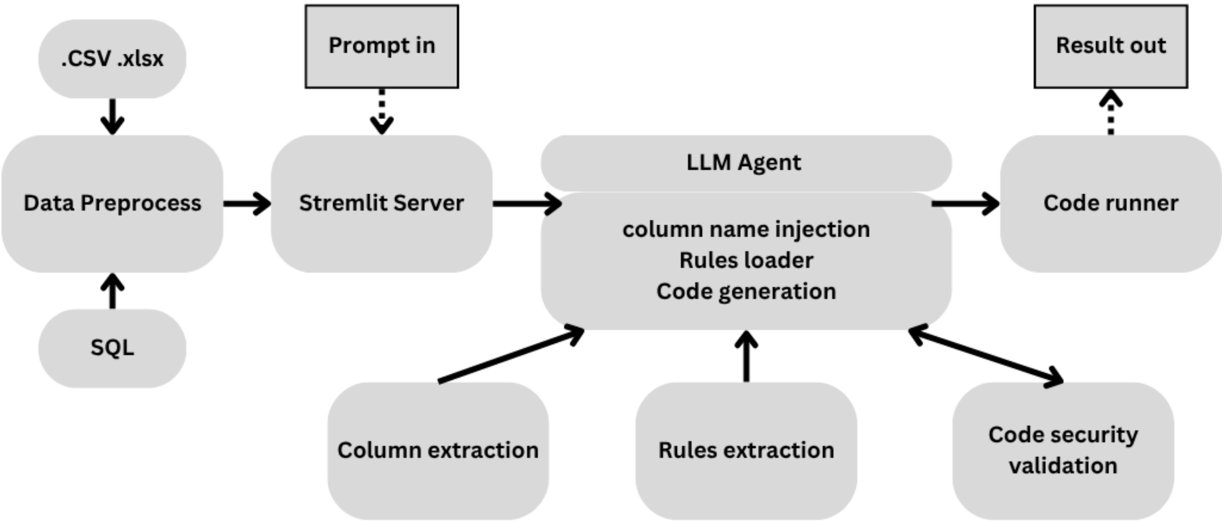


Figure 4: LLM-based Data Query Structure (Table Version)

Step No.	Component	Description / Action	Data Flow Direction
1	.CSV / .XLSX / SQL	Input data source	Input to Data Preprocess
2	Data Preprocess	Prepares data for query	To Streamlit Server
3	Prompt In	User prompt / query	To Streamlit Server
4	Streamlit Server	Interface layer between user and LLM	To LLM Agent

Figure 4: LLM-based Data Query Structure (Table Version)

Step No.	Component	Description / Action	Data Flow Direction
5	LLM Agent	Column name injection, rules loader, code generation	To Code Runner, Column Extraction, Rules Extraction, Code Security Validation
6	Column Extraction	Extract columns info	To LLM Agent (potential data loop)
7	Rules Extraction	Extract rules for safe query generation	To LLM Agent (potential data loop)
8	Code Security Validation	Validate generated code for security	To Code Runner
9	Code Runner	Execute generated code	Output to Result Out
10	Result Out	Display result to user	Final output

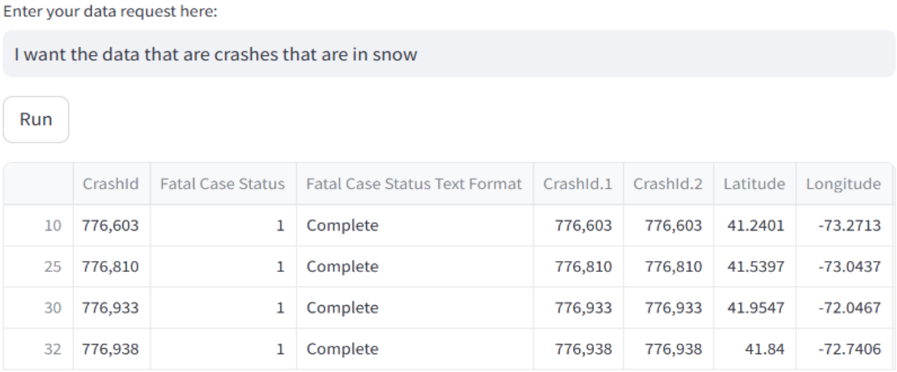
developed a passion for data analysis; this can be challenging. The purpose of the CTCDR data query tool is to allow everyone to discover patterns in traffic data, provide safer traffic, and help meet the Bipartisan Infrastructure Law. To further accommodate these goals, we now provide a new method: a hierarchical structure that allows users to enter prompts about the questions they are interested in and let LLM behave as the data analysis tool for each individual.

By involving the LLMs in this data analysis process, we can prompt a wave of new attention and engagement from advocates, policymakers, and scholars from many fields. [2] This will lower the boundaries for data analysis for everyone, not just professionals trained to do data analysis. Considering that the current web query tool is relatively complicated to operate, we produced a version using LLM that allows data retrieval and analysis without programming skills.

In recent years, LLMs trained from a text corpus have been successfully applied to various natural language processing tasks. However, the practicality of LLM in transportation safety has not been extensively investigated. With the implementation of the LLMs, we have developed an algorithm based on the existing crash data from the CTCDR website; we have also provided this as a public service to be used by CTDOT and public actors to query crashes and generate crash analysis automatically reports to investigate transportation safety issues [2,3]. We began by leveraging some common LLM tools in Python [4,5,7] and noticed their weaknesses and limitations. They usually include several instructions in the pre-prompt to teach the LLM how to generate graphs in the environment. However, because those packages only allow operations on the data frame, the capability to create interactive graphs, such as heatmaps, is limited, which led us to develop our own solution. Our implementation adds an extra step before the prompt is sent to the LLM. This step categorizes the intent of the prompt into three categories: data retrieval, graph generation, and interactive graph generation. For each category, we use different pre-prompting strategies, reducing the amount of confusing pre-prompting for the LLM and thereby increasing accuracy.

Our structure consists of four parts, as shown in Figure 4: the data preprocessing unit, Streamlit Server, LLM Agent, and Code Runner. The data preprocessing unit standardizes the data as much as possible, for example, formatting zip codes and changing data types in each column. Streamlit Server acts as a bridge, connecting the data to the LLM Agent. It takes the prompt from the user and sends the request to the LLM Agent. The agent processes the data to extract features and apply casting rules. Then, this additional information is appended to the user's prompt and sent to the LLM for processing. The LLM returns a code, which is supervised by security validation. If unauthorized behavior is detected, the request is sent back for regeneration, which is why the arrow in the graph shows in both directions. If the procedure passes all validations, the code is sent to the code execution environment, and the result is returned to the user.

In figure 5, it demonstrates the data query functionality, which allows users to input their queries in natural language. For example, "I want the data that are crashes in snow." Notice that the grammar of the query doesn't matter since the LLM will correct it in the backend, and it doesn't even need to be in English; depending on the user's choice of model, we will support all languages that selected LLM would.



Enter your data request here:

I want the data that are crashes that are in snow

	CrashId	Fatal Case Status	Fatal Case Status Text Format	CrashId.1	CrashId.2	Latitude	Longitude
10	776,603	1	Complete	776,603	776,603	41.2401	-73.2713
25	776,810	1	Complete	776,810	776,810	41.5397	-73.0437
30	776,933	1	Complete	776,933	776,933	41.9547	-72.0467
32	776,938	1	Complete	776,938	776,938	41.84	-72.7406

Figure 5: Data Query Demo Screenshot. User can visualize a download button when cursor moved on

Table 2 This table is linked with Figure 5. Data Query Demo, User can visualize a download button when cursor moved on. Using prompt: I want the data that are crashes taht are in snow.

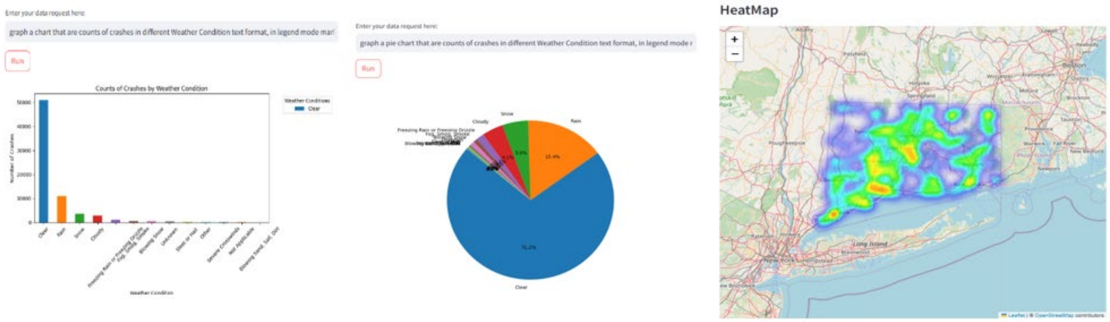
CrashId	Fatal Case Status	Fatal Case Status Text Format	CrashID.1	CrashId.2	Latitude	Longitude
776603	1	Complete	776603	776603	41.2401	-73.2713
776810	1	Complete	776810	776810	41.5397	-73.0437

776933	1	Complete	776933	776933	41.9547	-72.0467
776938	1	Complete	776938	776938	41.84	-72.7406

The system leverages an advanced LLM to handle the corresponding code generation and execution seamlessly. This approach significantly enhances user convenience, eliminating the need for technical expertise or manual intervention. Additionally, since the backend operations are fully automated, the system does not require dedicated personnel for management, resulting in substantial cost savings for the service.

Moreover, the platform features a versatile graphing system that supports the creation of various visualizations. This capability accelerates data analysis, especially for users without a technical background, by providing intuitive and ready-to-use graphical insights. For professionals with more requirements, the system accommodates advanced visualizations, including intricate graphs and a comprehensive series of charts. This ensures the tool remains robust and adaptable to diverse analytical needs.

As in Figure 6, the server received a request for a graph that counts occurrences under different weather conditions and colors the columns accordingly. Similarly, users can control the specific style of the graph directly in their query. The platform supports bar graphs, pie charts, and other types of visualizations, such as scatter plots.



Enter your data request here:

graph a chart that are counts of crashes in different Weather Condition text format, in legend mode mar

Enter your data request here:

graph a pie chart that are counts of crashes in different Weather Condition text format, in legend mode r

Figure 6: 3 Data query demo screenshot. Two graphing categories and with two types of graphing in static graphing mode. Showing one bar graph, one pie chart, one heatmap.

On the right side of Figure 6, it demonstrates a heat map feature. Unlike traditional graphs, the heat map is an interactive window similar to Google Maps, allowing users to zoom in, zoom out, and point the cursor to specific locations for more detailed information. It will be defaulted to centralize at Connecticut central coordinates, with read being more accidents and blue being fewer accidents. User can also configure these in their prompt.

We have experimented and believe that there is an expected use of LLMs that can be used to replace the work of the data analysis. Giving more opportunities to those people who want to find things in CTCDR and not have professional data analysis skills. The new data query system also allows faster and cleaner data download for professionals who want to do their own data analysis.

Nevertheless, we seek to exceed mere satisfaction with our current offerings. We are advancing this website by developing a dynamic dashboard that will be self-organizing and showcase the results of recent trending searches conducted by users. This initiative aims to reduce duplication of effort among individuals while encouraging exploration of new directions using the dataset. As a result, a wider range of users may contribute to the report through their analyses. In the final version of this product, automated report generation will occur at designated intervals, lowering program costs and providing a standardized report for residents of Connecticut.

Outcomes

Several supplementary modules were developed during this research, and each can be adopted independently.

First, our new LLM-based data retrieval system offers a more convenient method for performing data queries on the website. This system benefits both users and maintainers. For users, it provides a more intuitive and natural way to gather the data they need. For maintainers, the service significantly reduces costs compared to the traditional SQL-to-CSV approach, as it consumes less local CPU power and memory.

Second, the data cleaning module now functions as a fully automated process compatible with both SQL and CSV formats. This enhances the program’s stability and ensures a clean dataset for the database. Additionally, the cleaning rules implemented in the module can be adapted to build a more consistent and reliable database.

Finally, the capability for on-site data analysis adds significant value. This feature streamlines workflows for transportation agencies and professionals, providing a simpler and more efficient process. Furthermore, the implementation reduces power and equipment consumption, ultimately lowering costs for transportation agencies compared to employing a large team of data analysts.

Impacts

This project delivers multifaceted impacts, addressing both immediate and long-term benefits for traffic safety research and public policy in Connecticut. By integrating an LLM into the CTCDR database, we have transformed data access and analysis, providing a more intuitive and efficient interface compared to traditional query methods.

Innovation and Uniqueness

This project is distinguished by its innovative application of LLMs for database interaction, a pioneering approach in traffic safety research. Unlike conventional methods requiring complex filtering, our interface enables users to input plain language queries, dramatically improving accessibility and usability.

Impacted Groups

The project's benefits extend to various stakeholders:

- Scholars and Researchers: Easier data access and analysis enable more comprehensive studies on traffic safety.

- Citizens: Increased awareness and understanding of traffic risks empower individuals to make safer choices.
- Policymakers: Enhanced data-driven decision-making leads to more effective traffic policies and infrastructure improvements.

Quantified Impact

While exact metrics will vary, the potential impact is significant:

- Time Savings: Streamlined data retrieval reduces the time needed for data analysis.
- Error Reduction: Minimizing manual steps lowers the likelihood of data handling errors.
- Increased Utilization: Easier access encourages broader and more frequent use of data for research and policy-making.

Policy and Decision Support

The streamlined data retrieval process supports policymakers in creating informed, evidence-based policies. This can lead to faster responses to traffic issues and more targeted infrastructure improvements, ultimately enhancing road safety.

Long-term Effects

The project's contributions are expected to catalyze continuous improvements in traffic systems, resulting in safer roads and lower accident rates, especially for those involving VRUs. These advancements contribute to a more secure and efficient transportation network in Connecticut.

Social and Economic Benefits

Reducing traffic accidents can yield substantial economic benefits, including decreased medical costs and minimized productivity losses. Improved traffic safety also enhances the overall quality of life for residents.

Technology Diffusion

The natural language retrieval technology developed through this project has broad potential applications in other domains, such as medical data management and financial data analysis. This versatility positions it as a transformative solution for multiple industries.

Conclusion

Overall, the LLM-based approach would perform well with our innovative splitter and finely-tuned pre-prompting. It will enhance the data query system's user experience and reduce the need for extensive data analysis when exploring interests in the CTCDR database. While this will lower the costs associated with hiring data analysts to address people's questions, it will also encourage more individuals to participate in the exploration and verification of their areas of interest. These explorations can further enrich the report, making it more comprehensive and accessible. Furthermore, we will continue experimenting with and refining this service to ensure its reliability and robustness.

References

[1]“Connecticut Crash Data Repository.” Accessed: Dec. 26, 2024. [Online]. Available: https://www.ctcrash.uconn.edu/

[2]“Incorporating Large Language Models (LLMs) into Transportation Safety Analytics: NEUTC : UMass Amherst.” Accessed: Dec. 27, 2024. [Online]. Available: https://www.umass.edu/neutc/projects/incorporating-large-language-models-llms-transportation-safety-analytics

[3]“Introducing ChatGPT.” Accessed: Dec. 26, 2024. [Online]. Available: https://openai.com/index/chatgpt/

[4]“LangChain.” Accessed: Dec. 26, 2024. [Online]. Available: https://www.langchain.com/

[5]“langroid.” Accessed: Dec. 26, 2024. [Online]. Available: https://langroid.github.io/langroid/

[6]V. Singh et al., “Panda: Performance debugging for databases using LLM agents,” Amazon Science. Accessed: Dec. 26, 2024. [Online]. Available: https://www.amazon.science/publications/panda-performance-debugging-for-databases-using-llm-agents

[7]“PandasAI - Conversational Data Analysis.” Accessed: Dec. 26, 2024. [Online]. Available: https://pandas-ai.com/