Beyond Machine Learning: The Power of Large Language Models in Traffic Accident Management

Artur Grigorev (PhD student)

Artur.Grigorev@student.uts.edu.au

University of Technology Sydney

Background

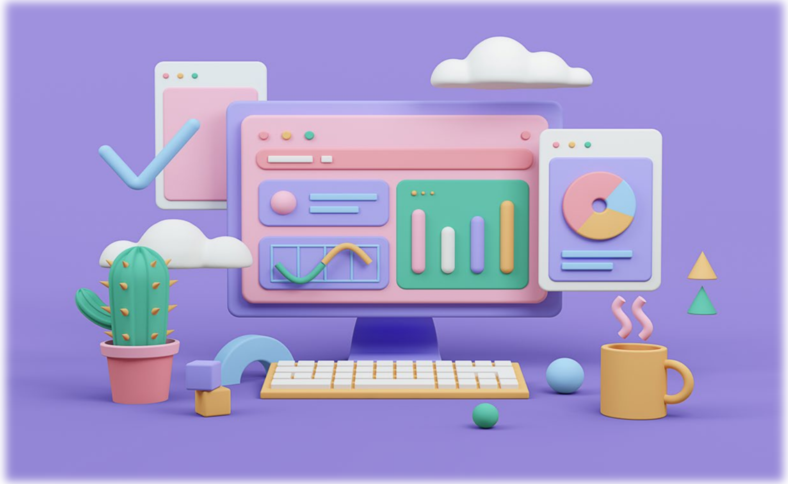
What do we know about traffic accidents?

Statistics: The annual economic cost of road crashes in Australia was estimated at $27 billion in 2017 [1]. Over 5 million accidents happen annually in the United States [2]. Also, accidents result in 1.35 million fatalities worldwide in 2016.

Congestion: Traffic accidents pose significant challenges to modern transportation systems, affecting traffic flow and public safety.

Prediction: Accurate modelling of traffic accidents is crucial for intelligent transportation systems, for reducing traffic congestion and economic cost associated with accidents.

Large Language Models: These models hold considerable promise for addressing the complexities associated with processing unstructured datasets and enhancing the efficiency of accident modelling.



[1] https://infrastructure.gov.au/roads/safety/,

[2] National Highway Traffic Safety Administration. Traffic safety facts 2013. U.S. department of transportation, 2013.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Current Limitations & Potential of LLMs

Limitations of Traditional Models:

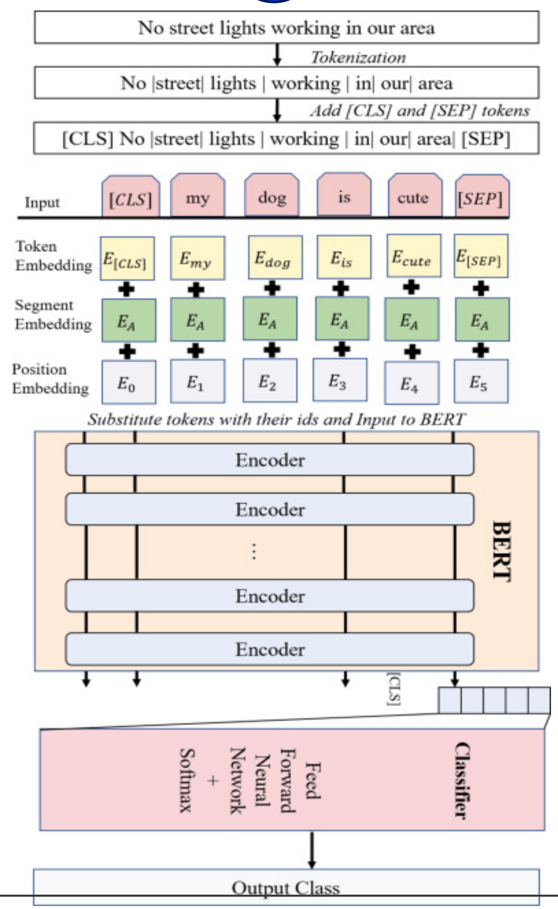
- **Accident Report Format**: Models built on structured/tabular data often can't transfer between systems due to using different accident report formats.
- **Linguistic Features**: Inability to capture complex linguistic features in textual accident reports.

Potential of Language Models:

- **Leveraging Unstructured Accident Report representation:** Traffic incident reports and other related text data represent a rich source of information that is often underutilized in traditional predictive models.
- **Model Transferability (e.g. between countries)**: Aim to develop a universally applicable model (cross-dataset) by leveraging language models.

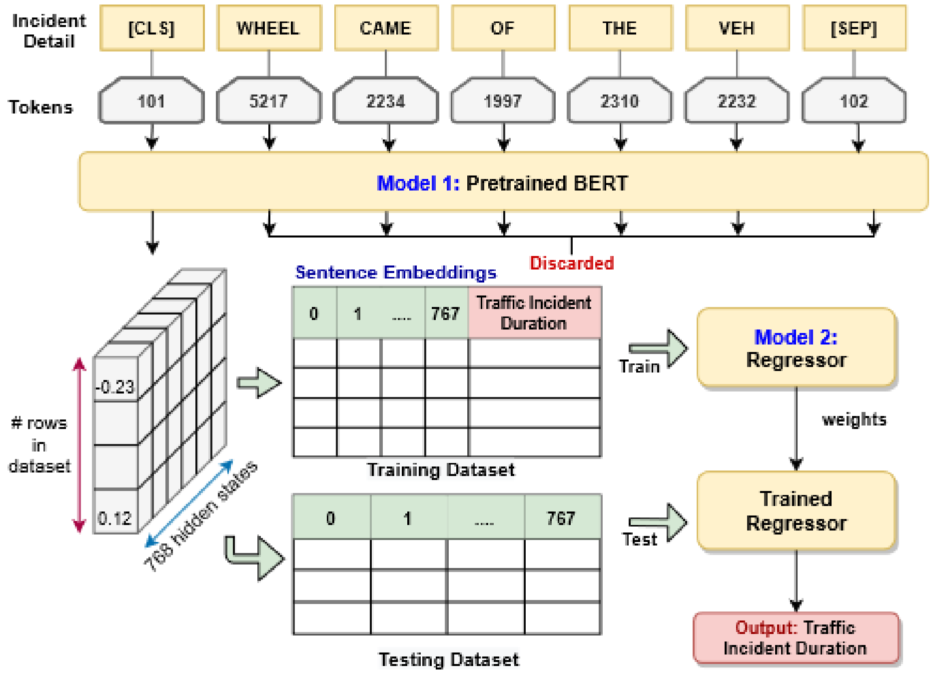
Existing research: incident severity classification / duration prediction

1



Oliaee, A. H., Das, S., Liu, J., & Rahman, M. A. (2023).
Using Bidirectional Encoder Representations from Transformers (BERT) to classify traffic crash severity types.
Natural Language Processing Journal, 3, 100007.

2



Agrawal, P., Franklin, A., Pawar, D., & Srijith, P. K. (2021, September).
Traffic Incident Duration Prediction using BERT Representation of Text.
In *2021 IEEE 94th Vehicular Technology Conference.* IEEE.

3

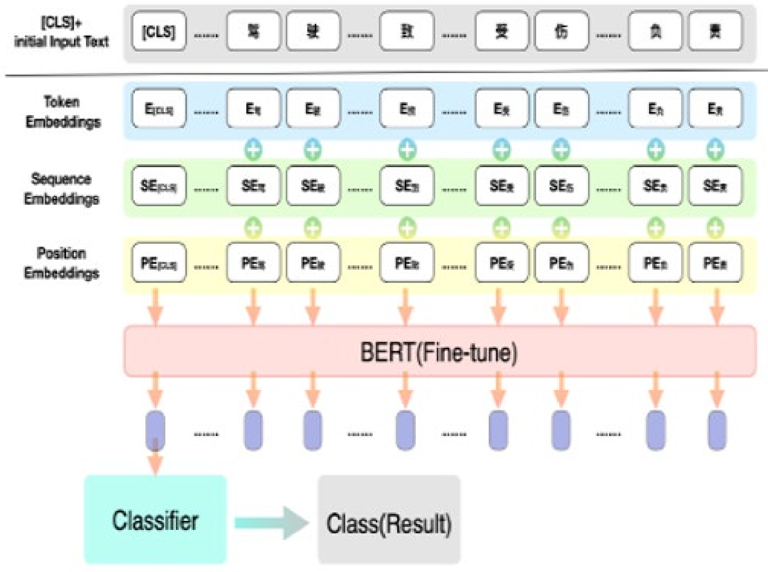


Figure 1. The input and output of the BERT model when doing classification tasks.

Yuan, S., & Wang, Q. (2022, February).
Imbalanced traffic accident text classification based on Bert-RCNN.
In *Journal of Physics: Conference Series* (Vol. 2170, No. 1, p. 012003). IOP Publishing.

LLM for text-to-dataframe processing

Zheng, O., Abdel-Aty, M., Wang, D., Wang, Z., & Ding, S. (2023). ChatGPT is on the horizon: Could a large language model be all we need for Intelligent Transportation?. arXiv preprint arXiv:2303.05382.

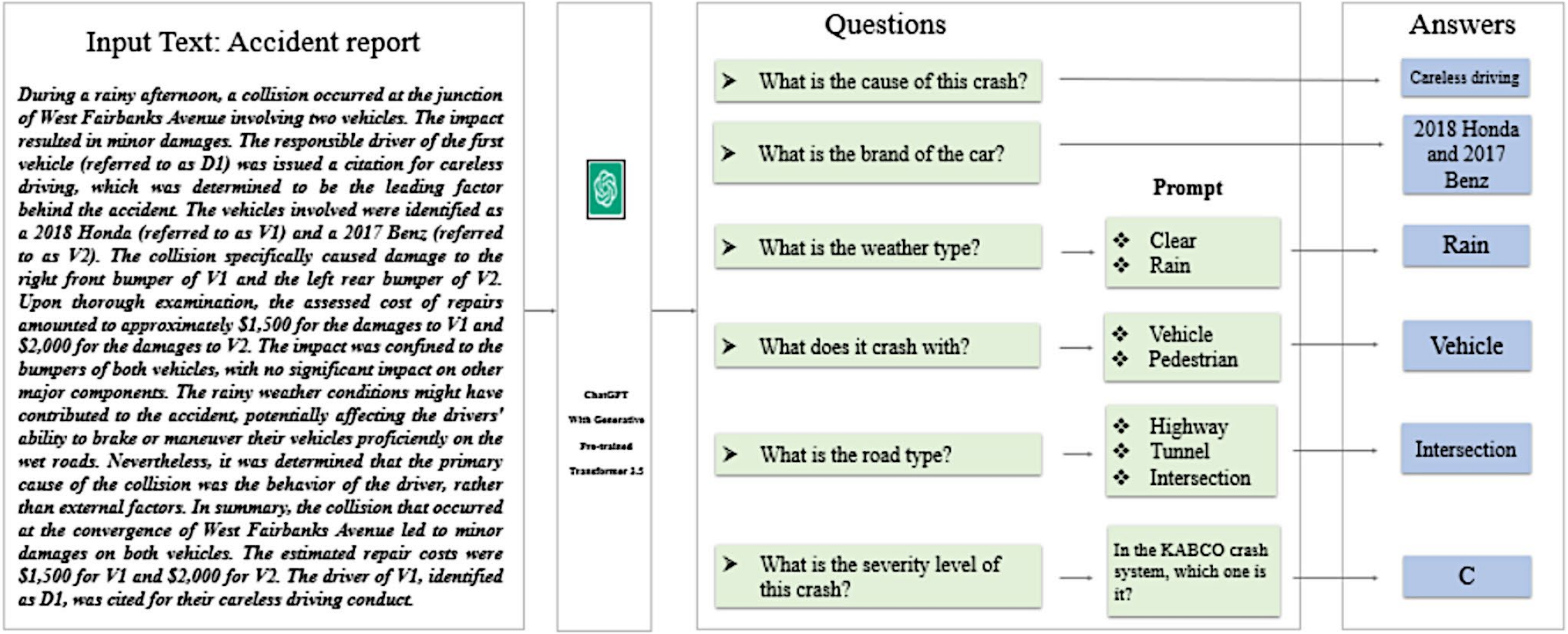


Figure 1 Example of accident information extraction through ChatGPT.

Novel Approach: Application of LLMs in traffic accident modelling

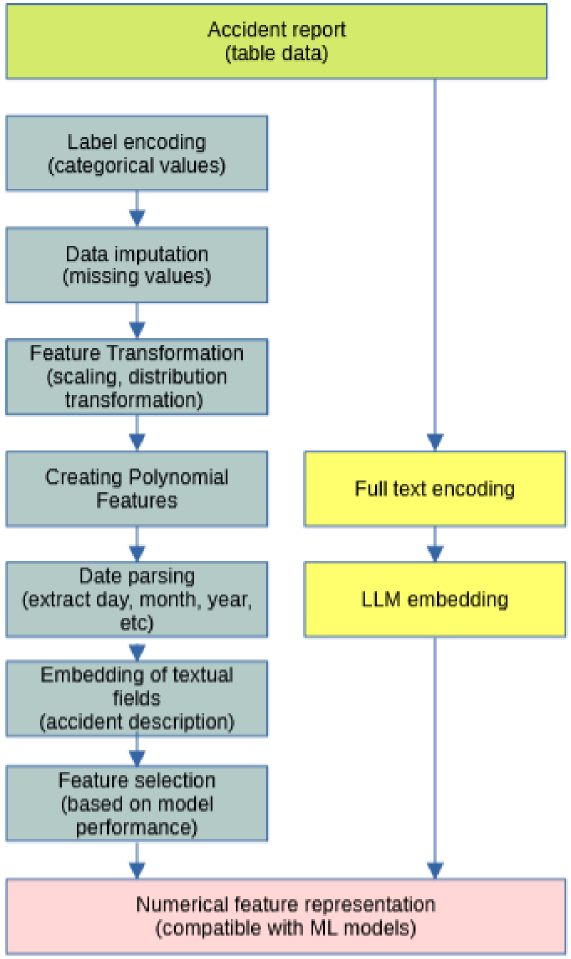
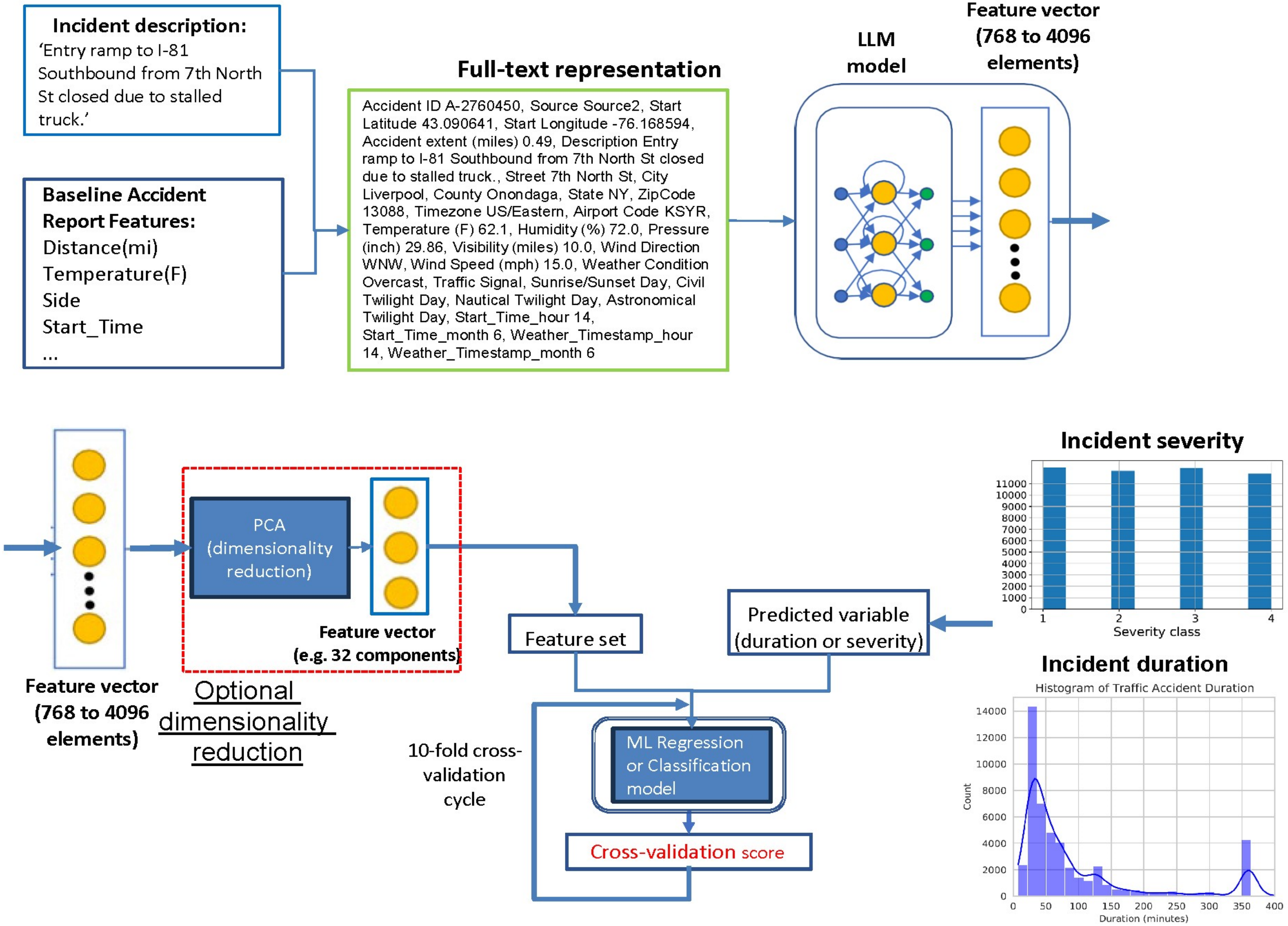


FIGURE 1 The benefit of using LLM models

Novel Approach: Diagram



UTS

Datasets

1. **Countrywise Traffic Accident Dataset (USA) – 25,000 cases (even severity class sampling)**

   https://smoosavi.org/datasets/us_accidents

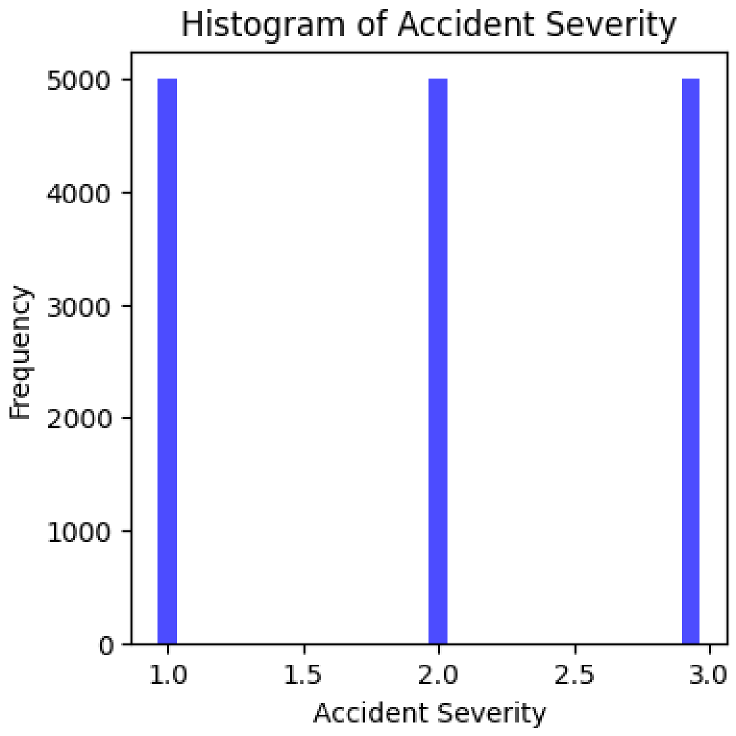
2. **Road Safety Data (UK) 2018,2019,2020,2021 – 20,000 cases (even severity class sampling)**

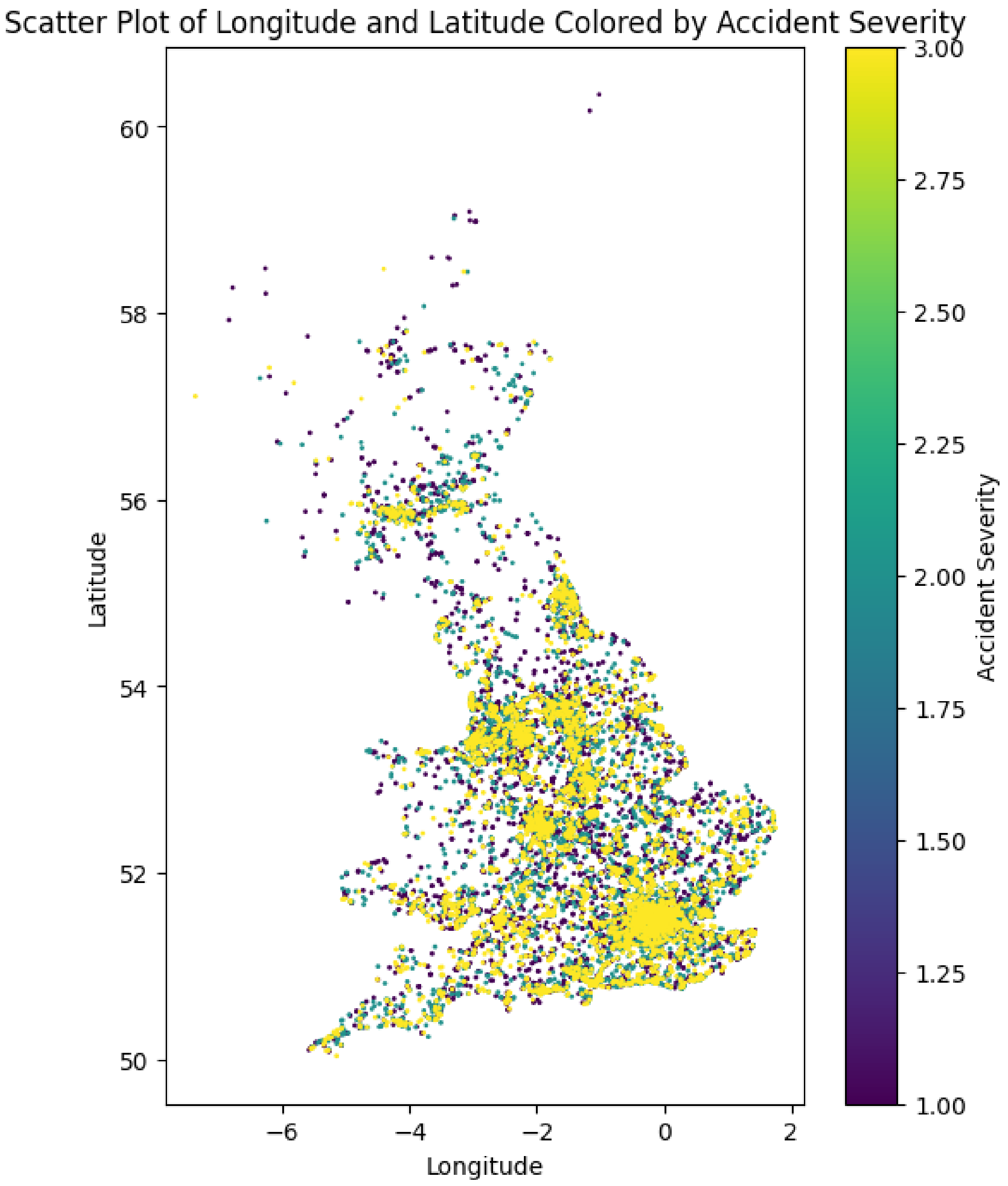
   https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

3. **Queensland Road crash data (Q) – 25,000 cases (even severity class sampling)**

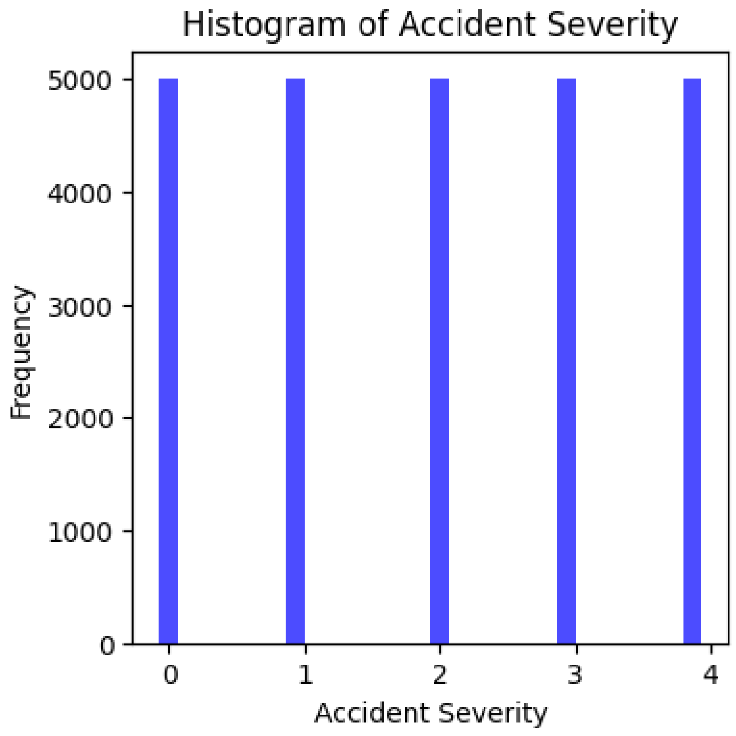
   https://www.data.qld.gov.au/dataset/crash-data-from-queensland-roads/resource/e88943c0-5968-4972-a15f-38e120d72ec0

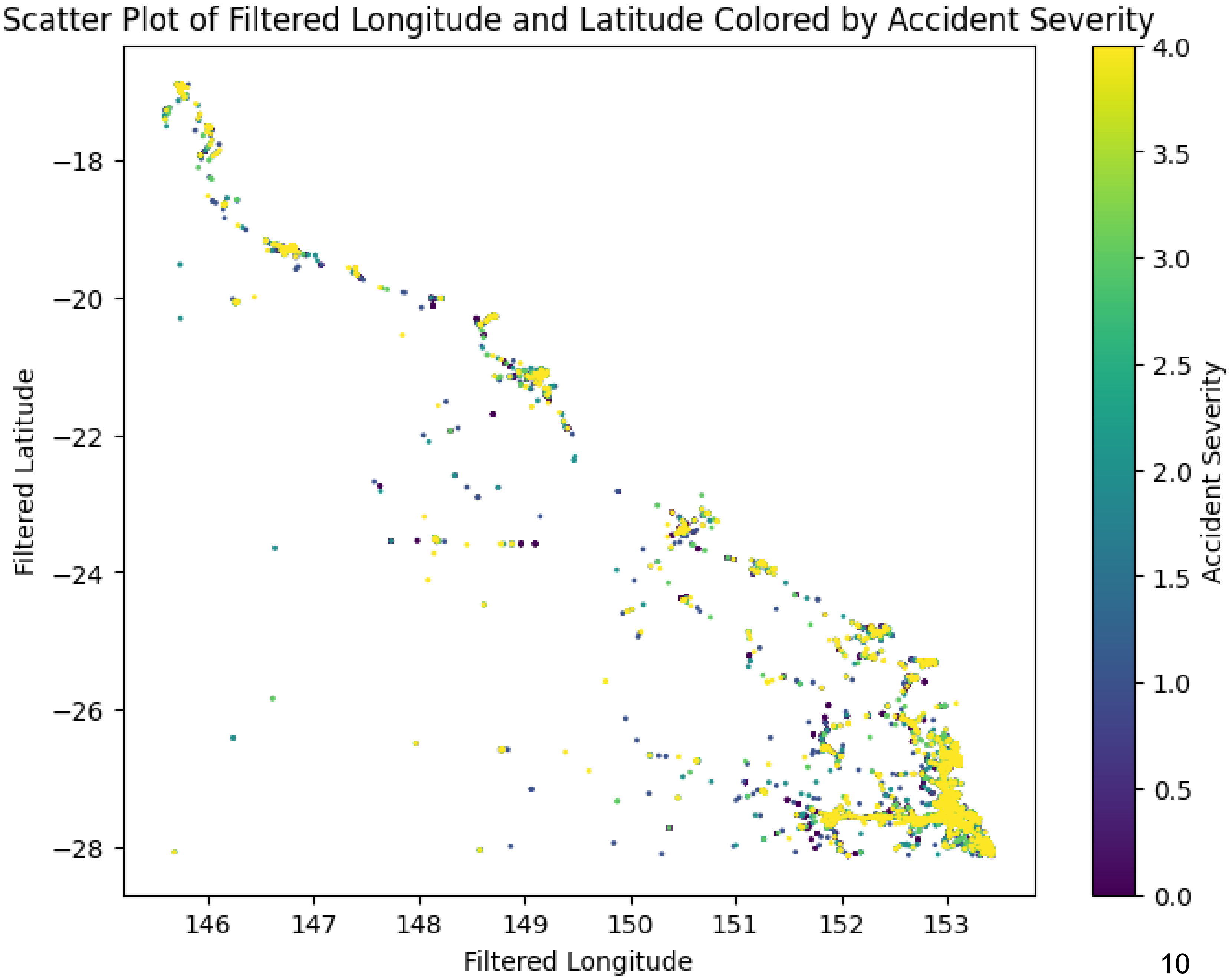
Road Safety Data (UK)





Queensland Road crash data (Q)





Full text representation

Example of full text representation for USA data set:

Accident ID A-7463401, Source Source1, Start Latitude 32.68116, Start Longitude -97.02426, End Latitude 32.67618, End Longitude -97.03483, Accident extent (miles) 0.7040000000000001, Description Ramp to I-20 Westbound - Accident., Street President George Bush Tpke S, City Grand Prairie, County Dallas, State TX, ZipCode 75052, Timezone US/Central, Airport Code KGPM, Temperature (F) 48.2, Humidity (%) 75.0, Pressure (inch) 30.26, Visibility (miles) 10.0, Wind Direction South, Wind Speed (mph) 5.8, Weather Condition Mostly Cloudy, Junction, Sunrise/Sunset Night, Civil Twilight Night, Nautical Twilight Night, Astronomical Twilight Night, Start_Time_hour 22, Start_Time_month 1, Weather_Timestamp_hour 22, Weather_Timestamp_month 1

Example of full text representation for UK data set:

accident_index: 2018460317259, accident_year: 2018, accident_reference: 460317259, location_easting_osgr: 556147.0, location_northing_osgr: 165830.0, longitude: 0.241871, latitude: 51.370065, police_force: 46, number_of_vehicles: 1, number_of_casualties: 1, date: 08/08/2018, day_of_week: 4, time: 11:35, local_authority_district: 538, local_authority_ons_district: E07000111, local_authority_highway: E10000016, first_road_class: 3, first_road_number: 20, road_type: 6, speed_limit: 60, junction_detail: 3, junction_control: 4, second_road_class: 6, second_road_number: 0, pedestrian_crossing_human_control: 0, pedestrian_crossing_physical_facilities: 0, light_conditions: 1, weather_conditions: 1, road_surface_conditions: 1, special_conditions_at_site: 0, carriageway_hazards: 0, urban_or_rural_area: 2, did_police_officer_attend_scene_of_accident: 1, trunk_road_flag: 2, lsoa_of_accident_location: E01024433

Example of full text representation for Queensland (Australia) data set:

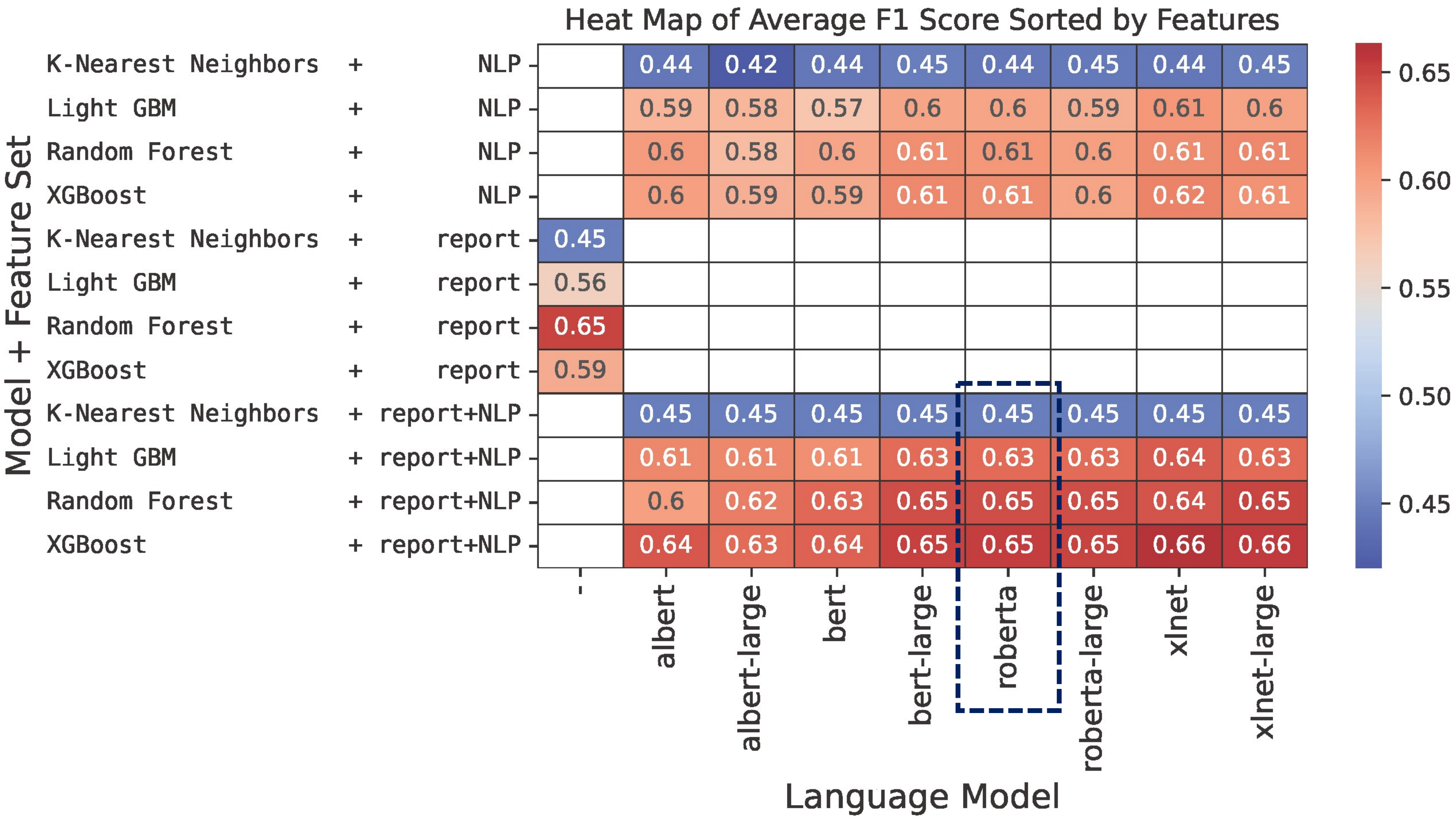
Crash_Ref_Number: 28863.0, Crash_Year: 2004.0, Crash_Month: September, Crash_Day_Of_Week: Wednesday, Crash_Hour: 6.0, Crash_Nature: Angle, Crash_Type: Multi-Vehicle, Crash_Longitude: 152.872284325108, Crash_Latitude: -27.5455985592659, Crash_Street: Kangaroo Gully Rd, Crash_Street_Intersecting: Mount Crosby Rd, State_Road_Name: Mount Crosby Road, Loc_Suburb: Anstead, Loc_Local_Government_Area: Brisbane City, Loc_Post_Code: 4070, Loc_Police_Division: Indooroopilly, Loc_Police_District: North Brisbane, Loc_Police_Region: Brisbane, Loc_Queensland_Transport_Region: SEQ North, Loc_Main_Roads_Region: Metropolitan, Loc_ABS_Statistical_Area_2: Pinjarra Hills - Pullenvale, Loc_ABS_Statistical_Area_3: Kenmore - Brookfield - Moggill, Loc_ABS_Statistical_Area_4: Brisbane - West, Loc_ABS_Remoteness: Major Cities, Loc_State_Electorate: Moggill, Loc_Federal_Electorate: Ryan, Crash_Controlling_Authority: State-controlled, Crash_Roadway_Feature: Intersection - T-Junction, Crash_Traffic_Control: No traffic control, Crash_Speed_Limit: 70 km/h, Crash_Road_Surface_Condition: Sealed - Dry, Crash_Atmospheric_Condition: Clear, Crash_Lighting_Condition: Daylight, Crash_Road_Horiz_Align: Curved - view open, Crash_Road_Vert_Align: Level, Crash_DCA_Code: 202.0, Crash_DCA_Description: Veh'S Opposite Approach: Thru-Right, Crash_DCA_Group_Description: Opposing vehicles turning, DCA_Key_Approach_Dir: E, Count_Unit_Car: 1.0, Count_Unit_Motorcycle_Moped: 1.0, Count_Unit_Truck: 0.0, Count_Unit_Bus: 0.0, Count_Unit_Bicycle: 0.0, Count_Unit_Pedestrian: 0.0, Count_Unit_Other: 0.0

LLM models

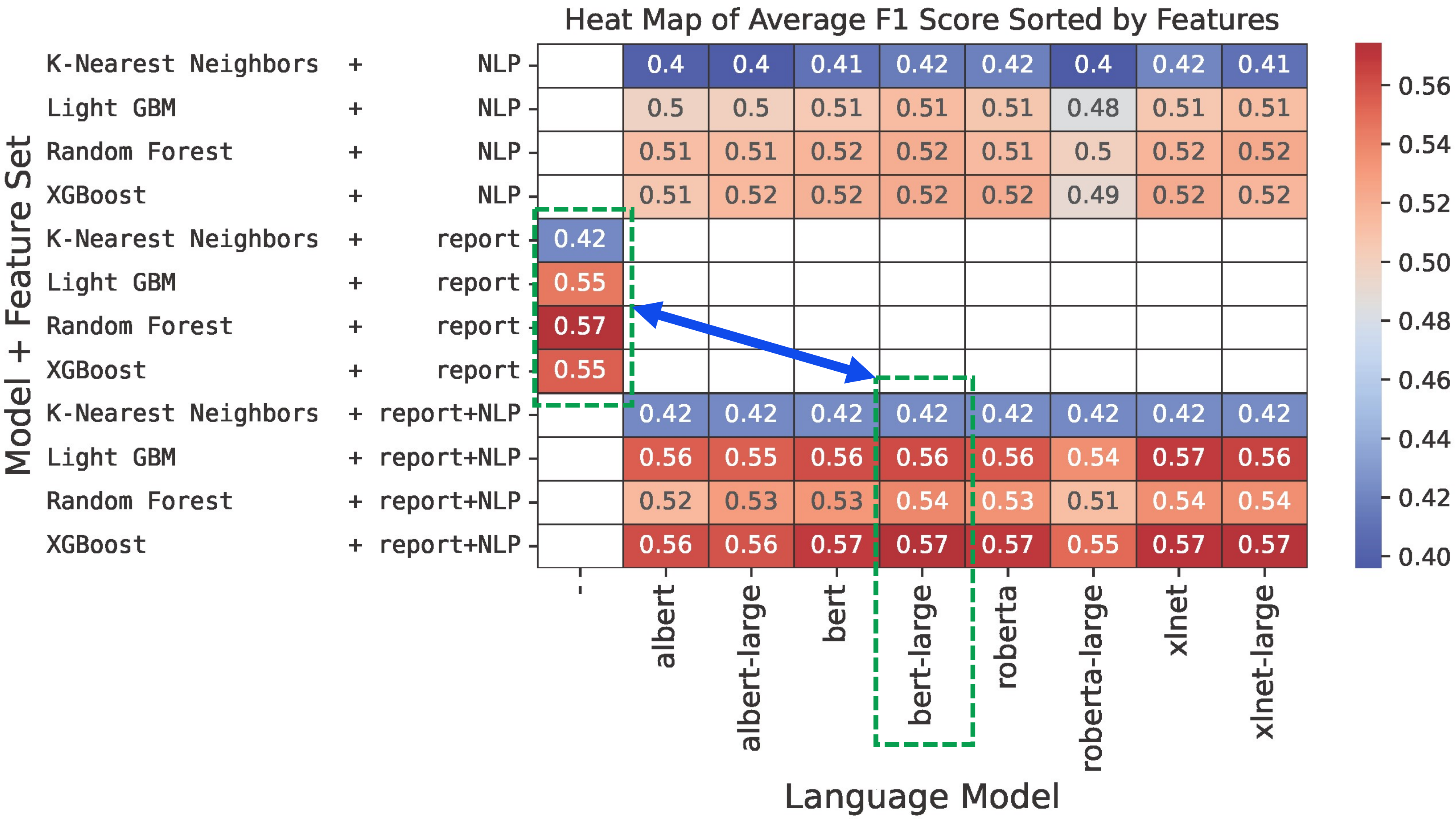
Model	Number of parameters	Training Method	Notable Features
BERT [9]	110 mil	Masked Language Modeling (MLM)	Bidirectional context, Pretrain-finetune discrepancy
BERT-large [9]	345 mil	Masked Language Modeling (MLM)	Bidirectional context, Pretrain-finetune discrepancy
XLNet [10]	110 mil	Generalized Autoregressive Pretraining	Overcomes BERT limitations, Transformer-XL integration
XLNet-large [10]	340 mil	Generalized Autoregressive Pretraining	Overcomes BERT limitations, Transformer-XL integration
GPT-2 [11]	1.5 billion	Autoregressive Language Modeling	Large-scale unsupervised, Zero-shot learning
RoBERTa [13]	125 mil	Optimized BERT (MLM with changes)	Longer training, Removed next sentence prediction, Dynamic masking
RoBERTa-large [13]	355 mil	Optimized BERT (MLM with changes)	Longer training, Removed next sentence prediction, Dynamic masking
ALBERT [14]	18.2 mil	Optimized BERT (MLM with changes)	Sentence Ordering Prediction, Layer-Sharing Architecture, Reduced Memory Footprint
ALBERT-large [14]	223 mil	Optimized BERT (MLM with changes)	Sentence Ordering Prediction, Layer-Sharing Architecture, Reduced Memory Footprint

TABLE I
SUMMARY OF NLP MODELS

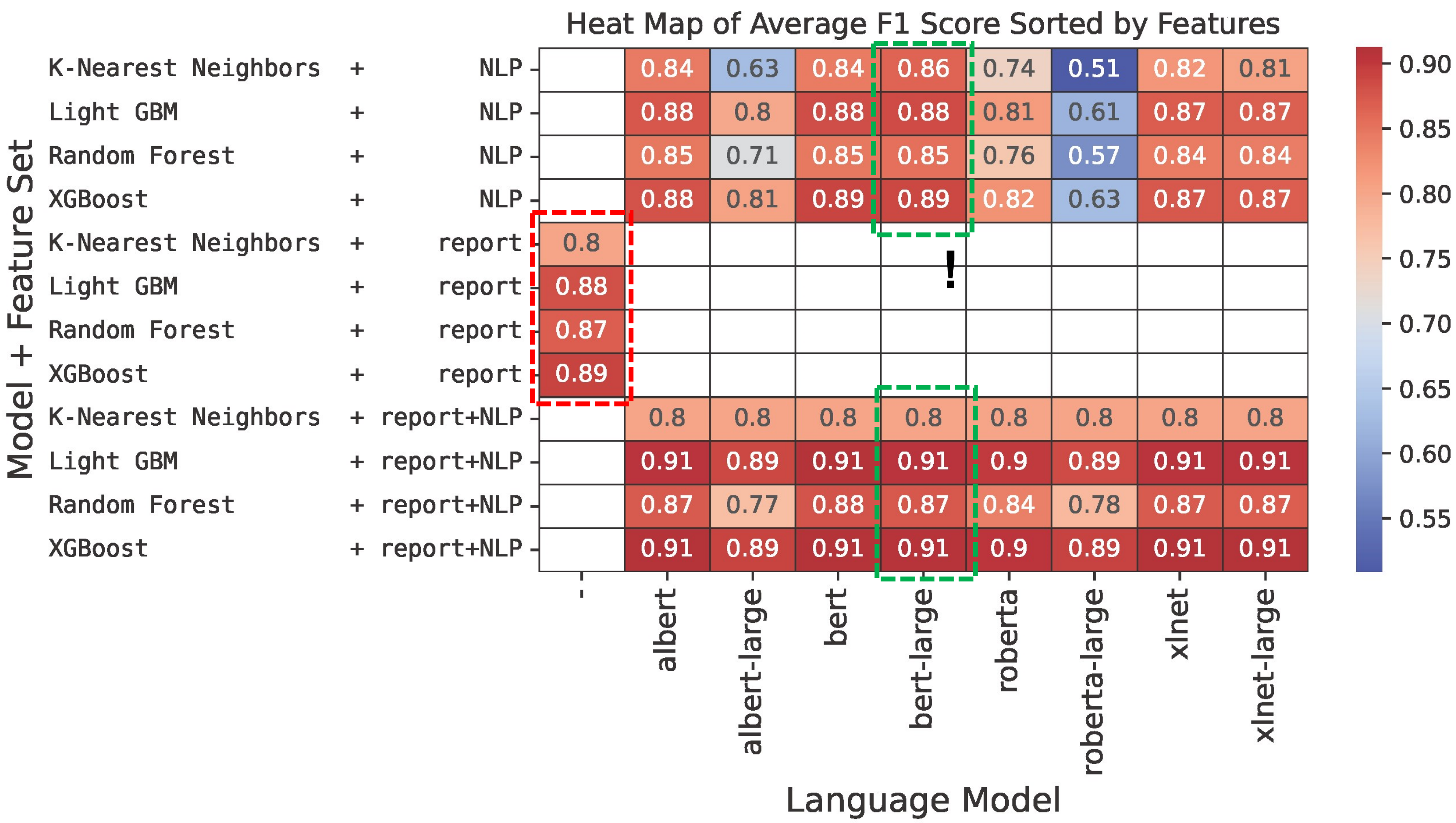
Queensland: Performance of LLM models



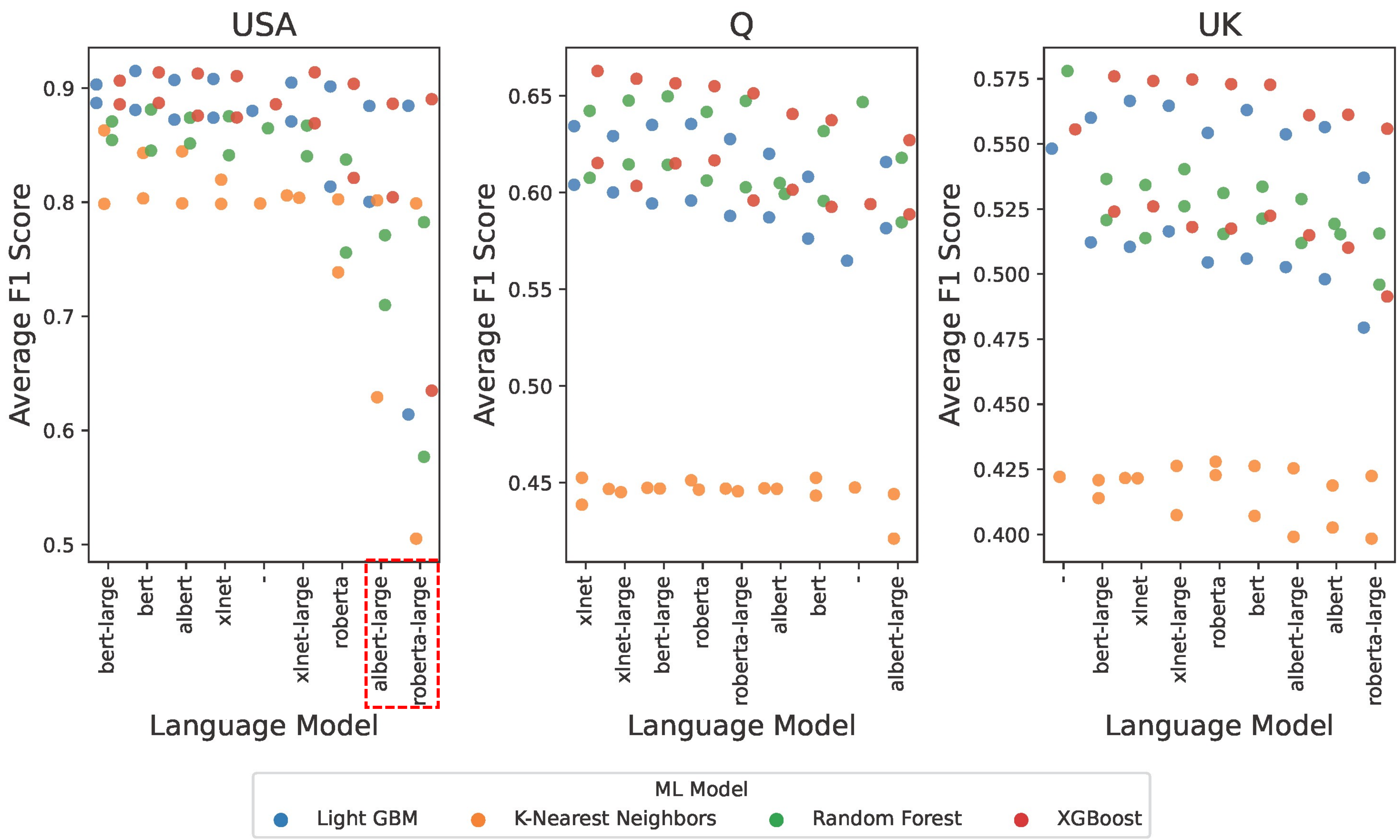
UK: Performance of LLM models



USA: Performance of LLM models



Overall performance of LLM models (NLP features only)



Research Summary

Goal: Improve traffic management and emergency response through more accurate severity classification and accident duration prediction.

Motivation: Traditional Machine Learning Approaches show reasonable accuracy but have the limitation of structured report representation.

Current Study

- **Model Variety**: We use 8 large language models (BERT, XLNet, RoBERTa, etc.).
- **Datasets**: We apply models to 3 diverse accident data from USA, UK, and Australia.

Implications

- **Higher performance**: Language models can outperform traditional machine learning in some scenarios.
- **Global Transferability**: LLM promise more accurate and universally applicable traffic management solutions, unconstrained to reporting format (which can vary across countries/cities).

Conclusion

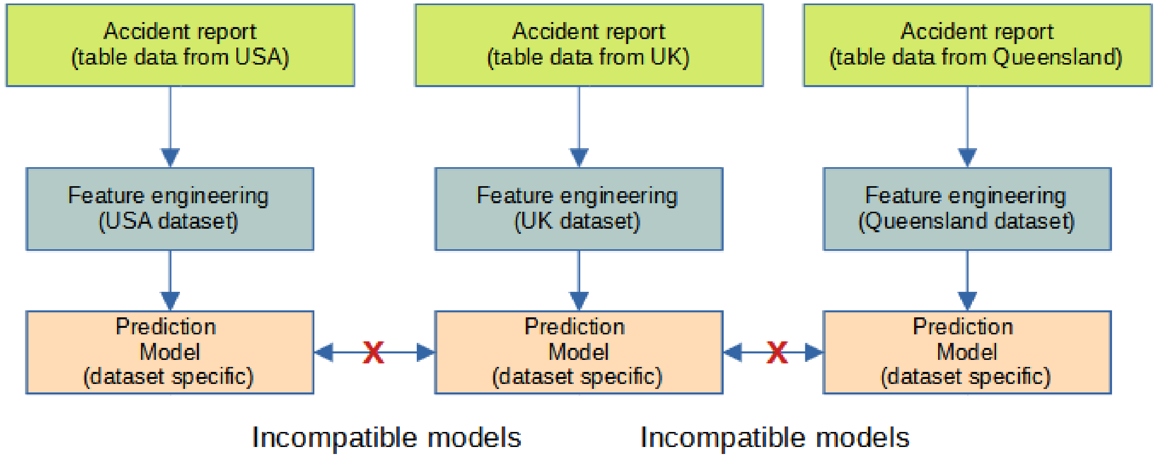
- **Leveraging Unstructured Accident Report representation:** Traffic incident reports and other related text data represent a rich source of information that is often underutilized in traditional predictive models.

- **The use of LLMs for accident severity classification:** This study presents a comprehensive comparison of various machine learning (including Random Forest and XGBoost) and large language models (BERT, RoBERTa, and Albert, etc) for feature extraction from textual accident report representation for the task of classification of traffic accident severity.

- **Insights:** The findings of our study offer valuable insights into the performance of different ML-LLM model combinations, which can support the development of future Traffic Incident Management Systems (TIMS).

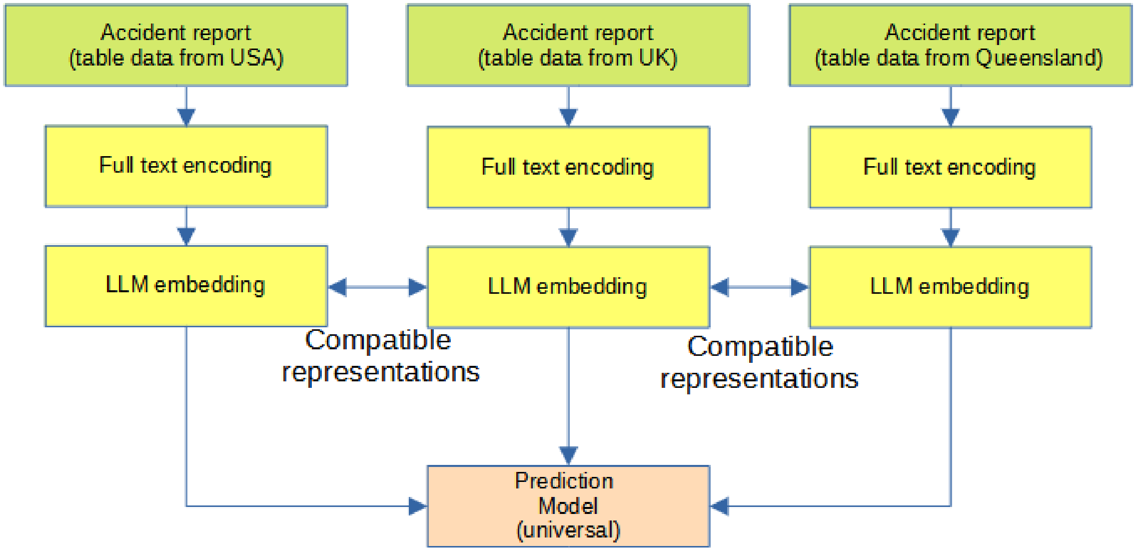
Future research on Applications of LLMs in traffic accident modelling

Before:



ML Models fine-tuned to data sets

After:



Model transferability:
Models for cross-dataset prediction

UTS

Beyond Machine Learning: The Power of Large Language Models in Traffic Accident Management

Thank You!

Artur Grigorev (PhD student)
Artur.Grigorev@student.uts.edu.au

University of Technology Sydney