Towards Building a Foundation AI Model for Railway Safety

Investigator Name: Evangelos (Vagelis) Papalexakis
Title: Professor
Department: Computer Science and Engineering
University: University of California Riverside

Investigator Name: Jia Chen
Title: Associate Professor of Teaching
Department: Electrical and Computer Engineering
University: University of California Riverside

Investigator Name: Yue Dong
Title: Assistant Professor
Department: Computer Science and Engineering
University: University of California Riverside

A Report on Research Sponsored by

University Transportation Center for Railway Safety (UTCRS)

University of California Riverside (UCR)

September 2025

Technical Report Documentation Page

1. Report No. UTCRS-UCR-O6CY24	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Towards Building a Foundation AI Model for Railway Safety		5. Report Date September 30, 2025
		6. Performing Organization Code UTCRS-UCR
7. Author(s) Evangelos Papalexakis, Jia Chen, and Yue Dong		8. Performing Organization Report No. UTCRS-UCR-O6CY24
9. Performing Organization Name and Address University Transportation Center for Railway Safety (UTCRS) University of California Riverside (UCR) 900 University Ave Riverside, CA 92521, USA		10. Work Unit No. (TRAIS)
		11. Contract or Grant No. 69A3552348340
12. Sponsoring Agency Name and Address U.S. Department of Transportation (USDOT) University Transportation Centers Program 1200 New Jersey Ave. SE Washington, DC, 20590		13. Type of Report and Period Covered Project Report June 1, 2024 – August 31, 2025
		14. Sponsoring Agency Code USDOT UTC Program
15. Supplementary Notes		

16. Abstract

Large language models (often also termed foundation models) are revolutionizing multiple aspects of every-day life and work, with the use of artificial intelligence (AI) agents like ChatGPT being now commonplace and transforming life, work, and scientific discovery as we know it. In this project, we harness the power of such foundation models in order to transform railway safety. Specifically, we build proof-of-concept prototypes that demonstrate the viability of foundation models for railway safety question-answer generation and accident report-news article alignment. Furthermore, we develop algorithms for the improvement of performance of large language models, including hallucination detection foundation and model denoising.

17. Key Words Safety Analysis, Data Mining, Data Science	18. Distribution Statement This report is available for download from https://www.utrgv.edu/railwaysafety/research/operations/index.htm		
19. Security Classification (of this report) None	20. Security Classification (of this page) None	21. No. of Pages 17	22. Price

Table of Contents

List of Abbreviations

AI	Artificial Intelligence
FRA	Federal Railroad Administration
LLMs	Large Language Models
MCCA	Multiview Canonical Correlation Analysis
MLP	Multi-Layer Perceptron
NeAT	NeuRal Additive Tensor Decomposition
PARAFAC	Parallel Factor Analysis
QA	Question-Answer
RAG	Retrieval Augmented Generation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
TRAWL	Tensor Reduced and Approximated Weights for Large Language Models
USDoT	U.S. Department of Transportation
UTC	University Transportation Center
UTCRS	University Transportation Center for Railway Safety
VLM	Vision Language Model

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 under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Acknowledgements

The authors wish to acknowledge the University Transportation Center for Railway Safety (UTCRS) for funding this project under the USDOT UTC Program Grant No. 69A3552348340.

Introduction

During the last couple of years, we have witnessed a revolution in generative models and large language models in particular, with examples like ChatGPT by OpenAI and Gemini by Google, that are able to digest vast amounts of written knowledge and generate outputs that appear very realistic and human-like. Even though the chatbot application of the models behind ChatGPT is the most widely known to the public, in fact, the models that power this functionality, often termed "foundation models" are of much more general interest, since they can act as substrate for additional AI functionality, which leverages the knowledge they have digested, in order to tackle different hard problems in a data-driven manner.

In particular, a very exciting research direction is to harness the power of those foundation models in order to amplify and ultimately transform research and practice in a variety of domains that span scientific discovery, engineering, and all aspects of everyday life. For example, a recent such push has happened in the field of computational biology, where a large consortium of researchers created bioCLIP [1], the first foundation model for computational biology, which can help answer problems that relate phenotypes and genotypes of different species, and can serve as the substrate for addressing hard computational biology problems and lead to discovery.

The major goal of this project is to develop proof-of-concept foundation large language models for railway safety. During this year, we have developed two prototype systems that demonstrate the viability and utility of large language models (LLMs) in railway safety. Below we first outline short summaries of the outcomes and we elaborate on more details in the rest of the document. Regarding our proof-of-concept systems, we have the following:

- **System 1**: We develop a prototype system that can use publicly available sources, such as online news articles and other public domain information, to populate the FRA Highway–Rail Grade Crossing Accident/Incident Report (Form 57) for a specific accident in a specific location.
- **System 2**: We develop a prototype system that has been fine-tune appropriately using FRA railway safety policy related documents and is capable of accurately answering questions pertaining to railway safety policy.

In addition to the two prototype systems, we have developed a suite of algorithms which are meant to improve the performance of large language models or improve the quality of the accident data, both tasks essential to powering the two systems that we have developed and ensuring their accurate performance. In summary, we developed the following algorithms:

- **Algorithm 1**: We develop a multi-view learning based algorithm that seeks to identify the most important information in an FRA Form 57 accident report.
- **Algorithm 2**: We develop an algorithm that can detect hallucinations in large language model outputs.
- **Algorithm 3**: We develop an algorithm that can remove noise from pre-trained large language models in a way that it can improve their performance accuracy without the need for additional expensive training or new data.

In the remainder of the report, we provide more details on each of the system and the algorithms developed.

System 1: A Real-Time System to Populate FRA Form 57 from News

Introduction: Local railway committees need timely situational awareness after highway–rail grade crossing incidents, yet official FRA investigations can take days to weeks. We received motivation for building this system through our interactions with stakeholders Sheldon Peterson and Paul Mim Mack from the Riverside County Transportation Commission. Our system populates Form 57 from news in near real time. Even though the key application we present in this prototype is the population of Form 57, our system can be also used to link a certain accident from the FRA database to articles online, and vice versa.

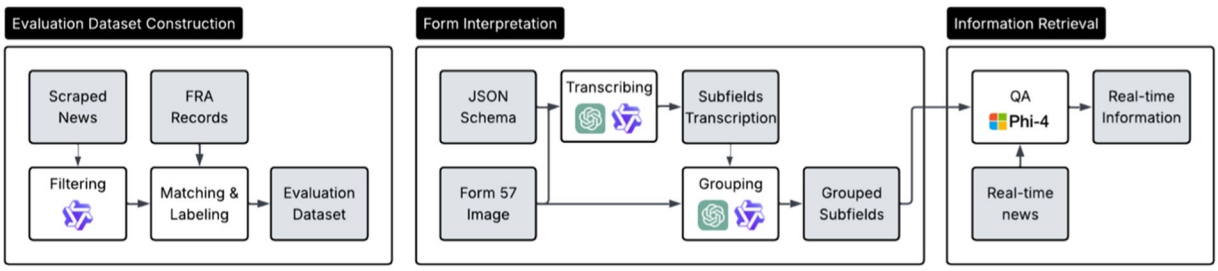

Figure 1: Overview of the pipeline of our system (figure adapted from accepted paper [8])

Technical details: Our approach addresses two core challenges: the form is visually irregular and semantically dense, and news is noisy. As an example, Figure 2 shows how challenging it is for off-the-shelf tools to work on this problem.

First, we build an evaluation dataset by scraping news and linking articles to official FRA records via fuzzy matching over several factors. Second, we design an information retrieval (IR) pipeline that (i) converts Form 57 into a JSON schema using a vision language model (VLM) with self-consistency merging, and (ii) performs grouped question answering (QA) following the intent of the form layout to reduce ambiguity. We evaluate with IR accuracy and coverage, comparing against various alternatives. Figure 1 shows the overview of the entire system pipeline.

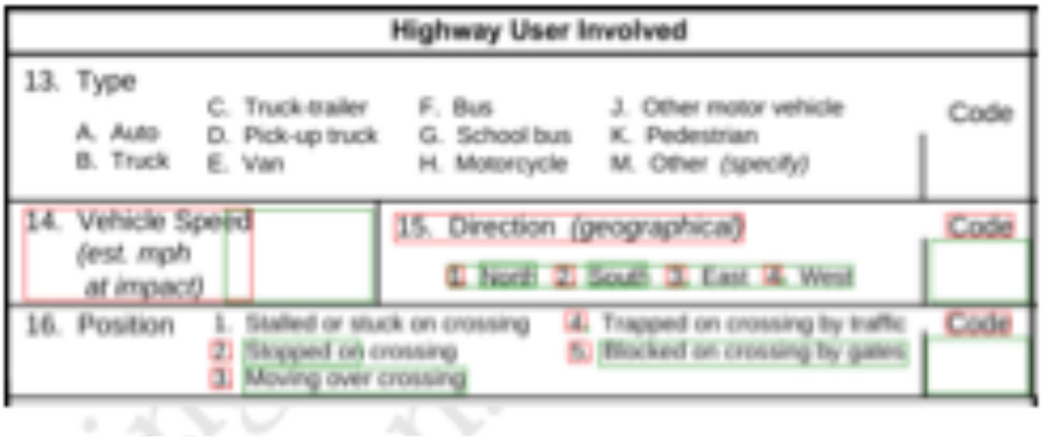

Figure 2: One of the technical challenges we faced was that existing tools were inadequate in properly reading Form 57. As shown, the parser fails to extract all fields properly. As part of building the system we had to adapt a Vision Language Model to do so.

Description of the system's functionality: Figure 3 shows a screenshot that captures the functionality of our prototype implementation for System 1. There are three key elements to the **front-end** of the system:

1. **Sidebar**: allows for filtering news articles by date
2. **Left Panel**: Select a news article and display its original website
3. **Right Panel**: Demonstrates the retrieved information that can be used to populate a version of Form 57 for the given accident

As this is a prototype implementation, we are planning to add more visual features before presenting the demo at the WSDM 2026, such as showing a completed version of Form 57 in its

original format, as well as introducing other criteria with which one can filter news articles, besides date.

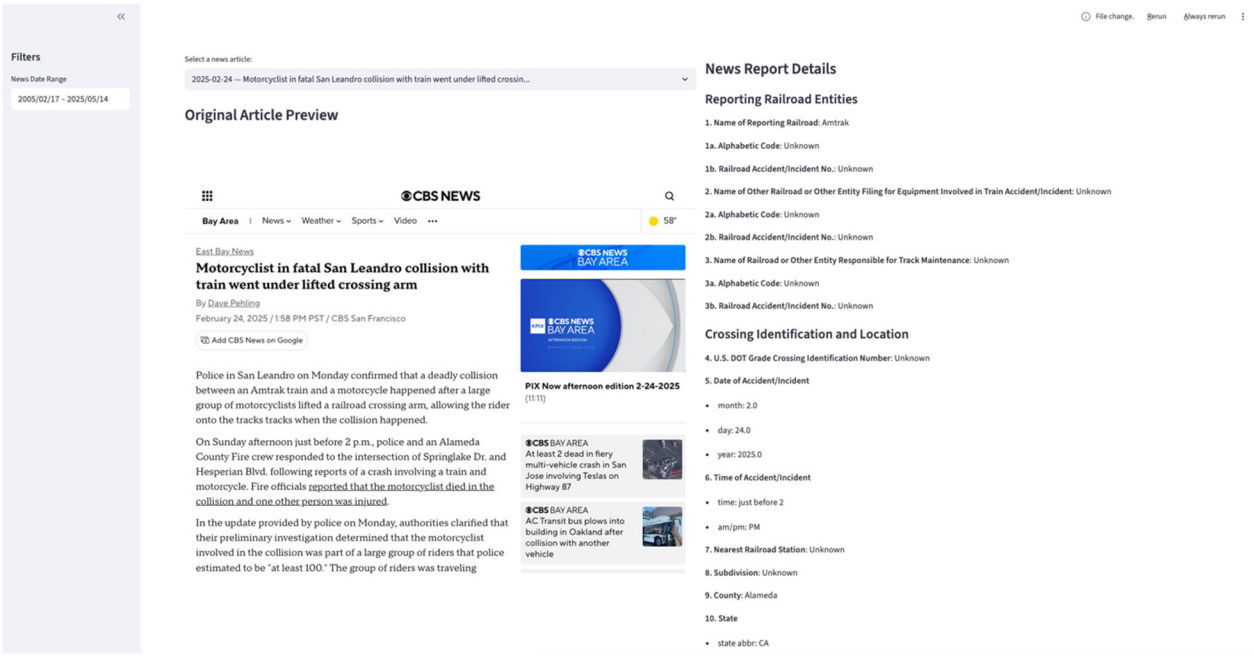

Figure 3: Example front-end functionality for System 1

The **back-end** of the system is making sure that the information displayed on the front-end. More specifically, a persistent news-monitoring bot continuously crawls news articles reporting train accidents. The scraped articles undergo a filtering process to retain only recent train–vehicle or train–pedestrian incidents. For each article, an LLM extracts the relevant details required for every field of Form 57.

Finally, we are planning to introduce the following updates to the system before its presentation at the WSDM 2025 conference, including:

- Multiple articles referring to the same accident will be aggregated and presented as a single group.
- Collected articles will be periodically re-checked to detect updates or newly published information.

Products:

1. Proof-of-concept system implementation: https://github.com/dlacksthd94/Railway-Safety

2. Paper accepted at the ACM International Conference on Web Search and Data Mining (WSDM) 2025, Demo Track [8]

System 2: Open-ended question-answer generation in railway safety

Introduction: Large Language Models (LLMs) have been extremely useful in powering chatbots that can converse with humans in a variety of topics with impressive performance. A typical use-case for such chatbots is question-answering, where the user poses a question and the LLM-powered chatbot is tasked with providing a factual and accurate answer to that question. Existing LLMs are trained on public domain data, and as a result can perform very well in question-answering tasks on general knowledge. However, when we attempt to do so for a very specialized domain, such as railway safety, which has its own nuances and intricacies, off-the-shelf methods fail to deliver the desired accuracy. In this work, we develop a prototype LLM-powered question-answering system that is specifically aligned with the railway safety domain by ensuring that it has been trained on appropriate FRA policy documents, and is able to answer questions that pertain to railway safety policy much more accurate than off-the-shelf or naïve approaches.

Technical details: Generating high-quality question-answer (QA) pairs for railway safety domain plays a critical role in both evaluating a person's transportation knowledge and document retrieval. More diverse, information-rich questions expose AI models to broader semantic patterns, enhancing their ability to generalize to unseen human-written queries. However, QA generation remains challenging, with existing approaches facing a tradeoff between leveraging expert examples and achieving topical diversity, and it has not been studied in literature. We present ExpertGenQA, a protocol that combines few-shot learning with structured topic and style categorization to generate comprehensive domain-specific QA pairs. Using U.S. Federal Railroad Administration documents as a test bed, we demonstrate that ExpertGenQA achieves twice the efficiency of baseline few-shot approaches while maintaining 94.4% topic coverage. Through systematic evaluation, we show that current LLM-based judges and reward models exhibit strong bias toward superficial writing styles rather than content quality. Our analysis using Bloom's Taxonomy reveals that ExpertGenQA better preserves the cognitive complexity distribution of expert-written questions compared to template-based approaches. When used to train retrieval

models, our generated queries improve top-1 accuracy by 13.02% over baseline performance, demonstrating their effectiveness for downstream applications in railway safety domain.

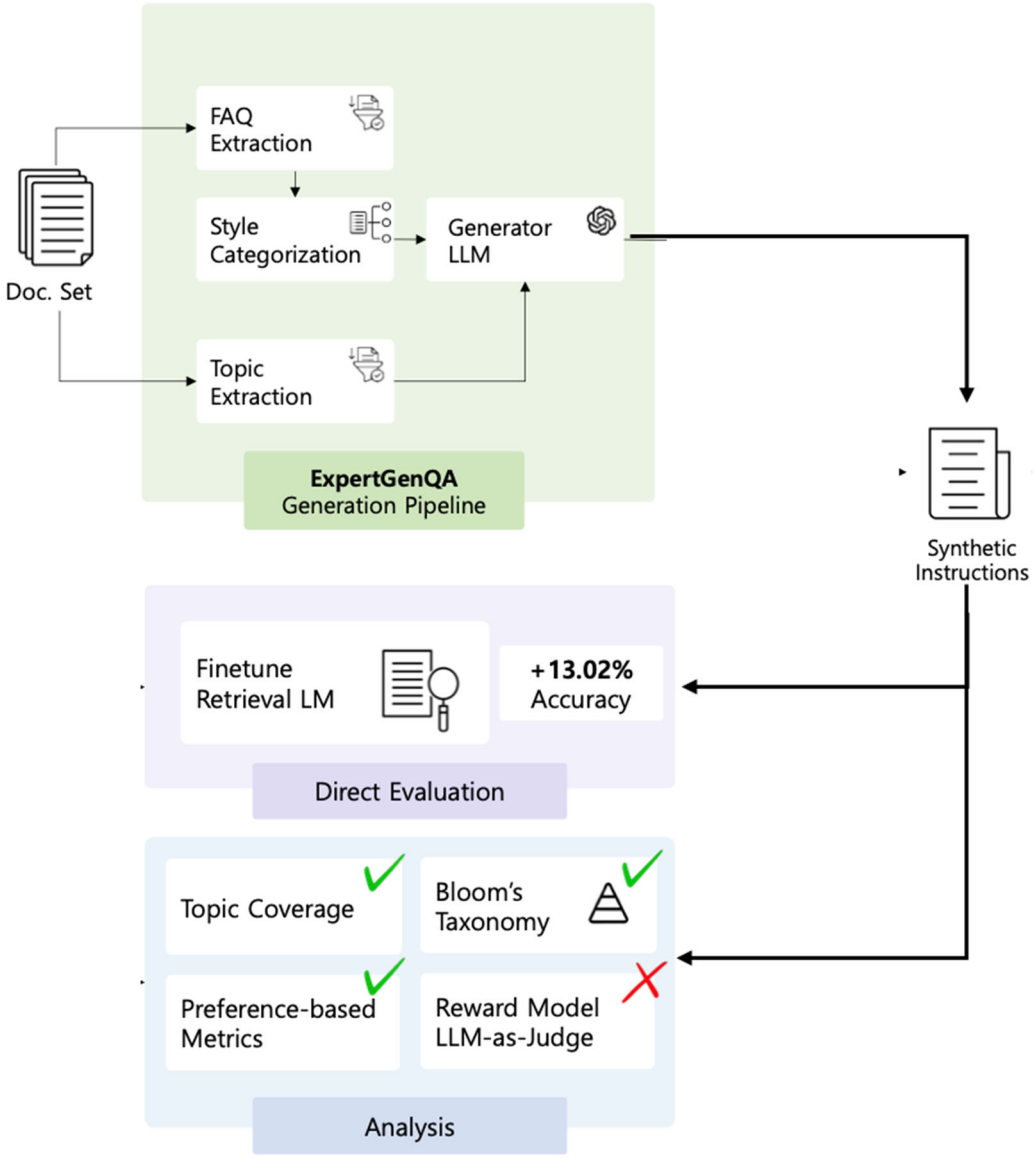

Figure 4: Overview of the system architecture (left) and the evaluation measures used (right). Figure adapted from accepted paper [2].

In Figure 4 we show an overview of the system and the evaluation measures used. Green checkmarks indicate interpretable metrics that correlate with improved retrieval accuracy, our primary evaluation metric. The red cross indicates our finding that both Reward Models and LLM-as-Judge show bias toward superfluous writing style and lack correlation with retrieval accuracy.

Products:

1. Proof-of-concept system implementation: https://github.com/Patchwork53/ExpertGenQA
2. Paper accepted at the Conference on Empirical Methods in Natural Language Processing (EMNLP 2025), Findings Track [2]

Algorithm 1: Multiview Machine Learning for Highway Railroad Crossing Accident Data Clustering

Introduction: The amount of information recorded per railway accident in the FRA Highway–Rail Grade Crossing Accident/Incident Report (Form 57) is very large, and documents everything about the accident, including environmental conditions, driver condition all the way to the road and crossing conditions. Given this abundance of information, a natural question is what is the best way to combine it in order to provide more meaningful analysis and processing of accident data. In previous work [4], which showed promising results in analyzing accident data, we have used this information in a "flat manner", where we do not distinguish between different "types" of information, such as "environmental" vs. "train-specific" information. In this work, we explicitly acknowledge that different groups of information are of different nature and offer different "views", as it is often called in the machine learning literature, to the data. As such, we cast the problem of clustering accident data as a multi-view learning problem and we demonstrate that doing so provides better quality of accident clusters and analysis.

Technical details: We randomly select 500 accidents that occurred with death and 500 accidents without death involved from the US DoT Accident report dataset which is available at Kaggle [3]. We generate two views where one contains environmental features, such as temperature and train speed and other one contains train features, such as the number locomotive units and the number of train cars in the train that is involved in the accident. There are two labels (with and without death) used. Even though there are numerous variables recorded during an accident report, many of which can be predictive of an accident [4], we opted for simplicity at first by including a small subset of numerical variables, in order to be compatible with the standard formulations of multiview canonical correlation analysis (MCCA) and PARAFAC2, a tensor decomposition method, however, we reserve a more complete investigation of accident features in

the context of multiview learning for future work. We report the average K -means (with the true K) clustering accuracy as well as the standard derivation of 10 Monte Carlo tests of MCCA, PARAFAC2, MCCA with an auxiliary view from PARAFAC2 with the best rank among a few candidates, PARAFAC2 with an auxiliary view from MCCA with the best d among a few candidates on the four datasets with respect to the dimension/rank. Clearly, adding the shared representation of multiple views extracted from PARAFAC2 to MCCA as an auxiliary view increases the clustering accuracy of MCCA compared against the performance of MCCA on the raw views. Similar statement holds true for MCCA to PARAFAC2. More details can be found in our conference publication [5].

Products:

1. Paper accepted at the Asilomar Conference on Signals, Systems, and Computers 2024 [5]

Algorithm 2: Hallucination Detection in Large Language Models

Introduction: Large Language Models (LLMs) have demonstrated effectiveness across a wide variety of tasks involving natural language. However, a fundamental problem of hallucinations still plagues these models, limiting their trustworthiness in generating consistent, truthful information. Detecting hallucinations has quickly become an important topic, with various methods such as uncertainty estimation, LLM Judges, retrieval augmented generation (RAG), and consistency checks showing promise.

Technical details: Many of these methods build upon foundational metrics, such as ROUGE or Perplexity, which often lack the semantic depth necessary to detect hallucinations effectively. In this work, we propose a novel approach inspired by ROUGE that constructs an N-Gram frequency tensor from LLM- generated text. This tensor captures richer semantic structure by encoding co-occurrence patterns, enabling better differentiation between factual and hallucinated content. We demonstrate this by applying tensor decomposition methods to extract singular values from each mode and use these as input features to train a multi-layer perceptron (MLP) binary classifier for hallucinations. Figure 5 shows an overview of the proposed algorithm.

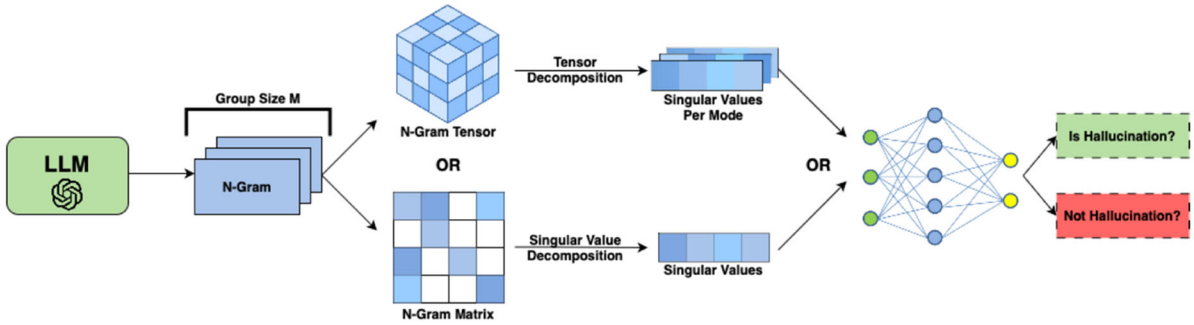

Figure 5: Overview of our proposed algorithm for hallucination detection. Figure adapted from paper submission [6].

Our method is evaluated on real dataset and demonstrates significant improvements over traditional baselines, as well as competitive performance against state-of-the-art LLM judges. This work is currently under submission [6].

Products:

1. Proof-of-concept algorithm implementation: https://github.com/Jeli04/NGram-Subspace-Evaluation
2. Paper currently under submission [6].

Algorithm 3: Tensor Reduced and Approximated Weights for Large Language Models

Introduction: The power of Generative Pre-trained Transformer (GPT) style Large Language Models (LLMs) rests on the fact that they have a very large number of parameters which has been trained over vast amounts of data, and as such, it has captured robust high-level concepts about the data so that it can be easily specialized to work on different "downstream" tasks with high success. Due to the sheer number of parameters that such LLMs have, there has been an increasing trend in developing methods for reducing the number of parameters in a way that is not harmful to the model's performance in various downstream tasks. In our work, we have identified that certain ways that have traditionally been used for efficiency and reducing the number of parameters, the so-called "low-rank approximation", can, in fact, be shown to have a denoising effect on the model: when approximating the parameters of the model in this way, we show that in many cases we observe an increase in performance, without the need for additional training or new data, which indicates that in addition to the useful high-level concepts that the model has learned, it has also learned noisy information which is eliminated by our approximation.

Technical details: Recent research has shown that factorizing large-scale language models for inference is an effective approach to improving model efficiency, significantly reducing model weights with minimal impact on performance. Interestingly, factorization can sometimes even improve accuracy by removing the noise that accumulates during training, particularly through matrix decompositions. However, recent work has primarily focused on single-matrix decompositions or lower precision techniques, which may fail to fully capture structural patterns. To address these limitations, we introduce TRAWL (Tensor Reduced and Approximated Weights for Large Language Models), a technique that applies tensor decomposition across multiple weight matrices to effectively denoise LLMs by capturing both global and local structural patterns.

Our experiments show that TRAWL improves the model performance by up to 16% over baseline models on benchmark datasets, without requiring additional data, training, or fine-tuning. We show an indicative experiment on Figure 6: We approximate the Fully-Connected layer weights for the RoBERTa LLM and test its performance on two standard benchmark datasets, "BiosProfession" (left) and "BigBench WikiQA" (right). The blue dashed line is the baseline existing method and the red dashed line is the baseline of no approximation. We show that for a number of different low-rank approximations, the test accuracy on those benchmark datasets improves. More details can be found in our conference publication [7].

Products:

1. Proof-of-concept algorithm implementation: https://github.com/HettyPatel/TRAWL
2. Paper accepted at the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) Special Sessions: Data Science: Foundations and Applications (DSFA) 2025 [6].

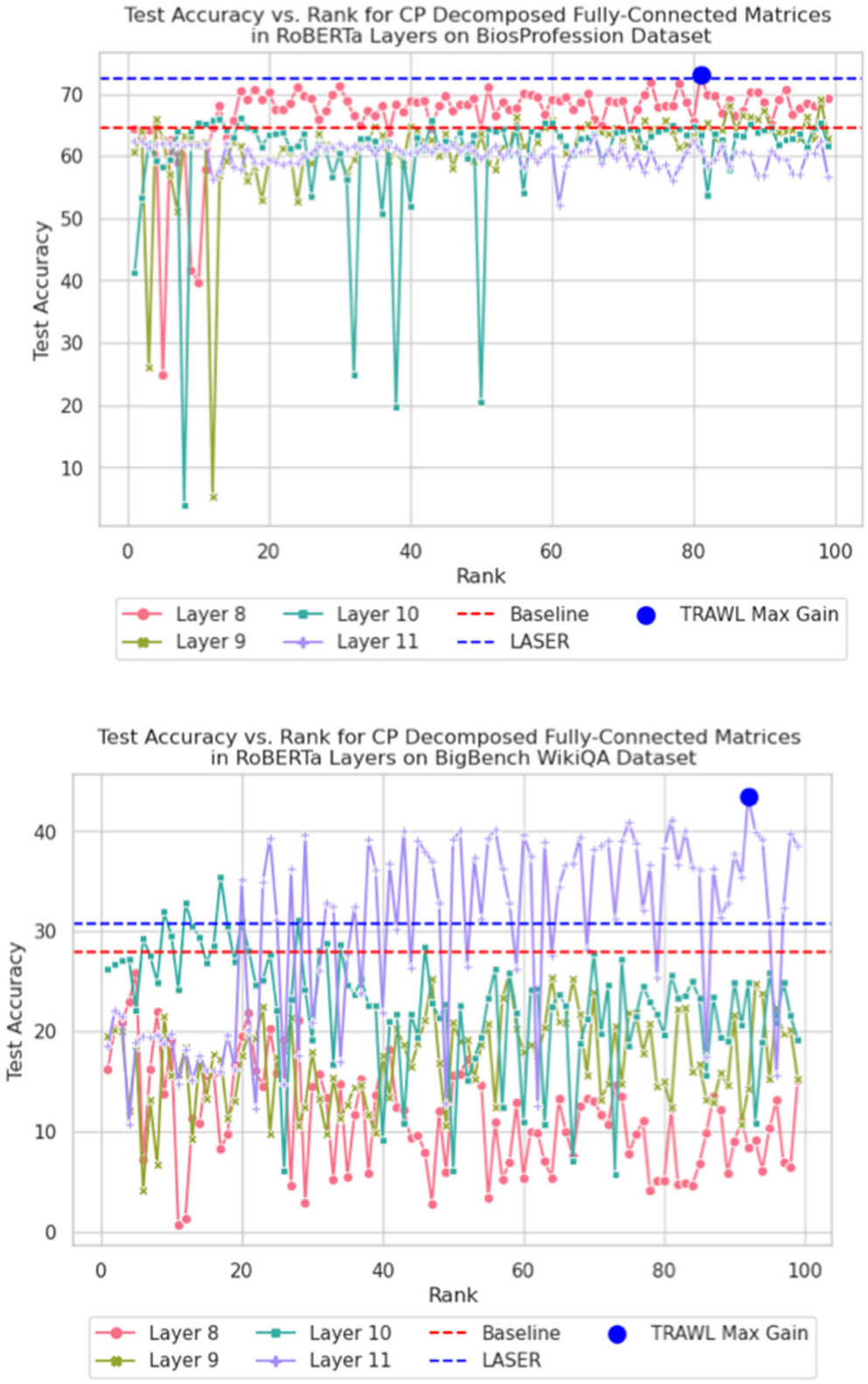

Figure 6: Low-rank approximation results on the RoBERTa LLM. Figure adapted from accepted paper [6].

Conclusions & Future Work

In this project we have started tapping on the potential of Large Language Models (LLMs) for railway safety, with two major proof-of-concept systems and three auxiliary algorithms that help with improving performance. We have demonstrated the viability of leveraging the power of LLMs in empowering railway safety professionals with tools that they can use in their day-to-day operations. Through our work, we have identified important research challenges that need to be addressed in order to develop and deliver workable and accurate systems. In the near future we are planning to address those challenges while at the same time introducing the visual aspect of railway safety into the proposed systems, as part of our Year 3 project within UTCRS.

References

[1] Stevens, S., Wu, J., Thompson, M.J., Campolongo, E.G., Song, C.H., Carlyn, D.E., Dong, L., Dahdul, W.M., Stewart, C., Berger-Wolf, T. and Chao, W.L., 2023. Bioclip: A vision foundation model for the tree of life. arXiv preprint arXiv:2311.18803.

[2] Haz Sameen Shahgir, Chansong Lim, Jia Chen, Evangelos E. Papalexakis, Yue Dong ``ExpertGenQA: Open-ended QA generation in Specialized Domains," Empirical Methods in Natural Language Processing (EMNLP), Findings, Suzhou, China, Nov. 2025.

[3] Us highway railroad crossing accident dataset. https://www.kaggle.com/ datasets/yogidsba/us-highway-railgrade-crossing-accident.

[4] Ethan Villalobos, Constantine Tarawneh, Jia Chen, Evangelos E. Papalexakis, and Ping Xu. Kernel ridge regression in predicting railway crossing accidents. In ASME/IEEE Joint Rail Conference, volume 87776, page V001T05A013. American Society of Mechanical Engineers, 2024.

[5] Jia Chen and Evangelos Papelexakis ``Project or Factorize? A case study of Multiview CCA and PARAFAC2 tensor factorization," Proc. of Asilomar Conf. on Signals, Systems, and Computers, Pacific Grove, CA, October 2024.

[6] Jerry Li and Evangelos Papalexakis. "Beyond ROUGE: N-Gram Subspace Features for LLM Hallucination Detection." arXiv preprint arXiv:2509.05360 (2025).

[7] Yiran Luo, Het Patel (co-1st author), Yu Fu, Dawon Ahn, Jia Chen, Yue Dong, Evangelos E. Papalexakis “TRAWL: Tensor Reduced and Approximated Weights for Large Language Models," Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) Special Sessions: Data Science: Foundations and Applications (DSFA), Sydney Australia, June 2025.

[8] Chansong Lim, Haz Sameen Shahgir, Yue Dong, Jia Chen, and Evangelos E. Papalexakis, “A Real-Time System to Populate FRA Form 57 from News”, ACM International Conference on Web Search and Data Mining (WSDM) 2025, Demo Track