Towards Harmonised Rail Safety Knowledge: LLM Techniques for EU Accident Report Processing

Shahrom Sohi[1], Dragomir Balan[1], Amin Anjomshoaa[1] and Axel Polleres[1]

[1]Vienna University of Economics and Business, Institute of Data, Process and Knowledge Management, Welthandelsplatz 1, 1020 Vienna

Abstract

This paper investigates the application of large language models (LLMs) to extract structured information from unstructured European railway accident reports. A corpus of 354 multilingual reports from Austria, Belgium, Ireland, and Poland was processed, with model performance evaluated through accuracy and completeness metrics. To enhance extraction quality, it has developed a Retrieval-Augmented Generation (RAG) pipeline and integrated it with a Neo4j graph database for knowledge representation and interactive visualisation. The EU Rail Accident Vocabulary was incorporated to standardise terminology and improve LLM outputs, using two leading models—GPT-4o-mini and Gemini 2.5 Pro. Results demonstrate that RAG-enabled LLMs significantly improve information extraction efficiency and accuracy, while supporting harmonised rail safety knowledge across European contexts. These findings underscore the potential of combining LLMs, domain ontologies, and graph-based representations to advance cross-border safety data integration and decision-making in the European railway sector.

Keywords

Large Language Models, Railway Accident Reports, Information Extraction, Name Entity Recognitions, Accidents Knowledge Graphs, Accidents Analysis

1. Introduction

Drawing lessons from past railway accidents is essential to improve operational safety [1], understand the dynamics of accident development, and build knowledge to support prevention, resilience, and awareness across railway systems[2]. Accident reports offer valuable information, but much of the knowledge they contain remains locked in unstructured text [1, 2, 3].

The extraction of knowledge from railway accident reports is a broad research area. Techniques range from topological and causal analysis to natural language processing (NLP)[1, 4], ontology-based knowledge graphs[2], and text mining[5]. Each approach seeks to convert raw textual data into structured knowledge that can support engineering, operational, and policy-level improvements in rail safety.

This paper, to the best of the authors' knowledge, constitutes the first attempt to investigate the use LLM technologies for extracting unstructured information from EU-level railway accident reports. The study showcases how entity extraction enables the creation of accident graphs derived from LLM-generated data, with outcomes benchmarked against manually entered metadata. The results support improved knowledge representation of European railway accidents and offer an assessment of the opportunities and constraints of LLMs in the harmonisation of EU railway safety knowledge.

2. Background

The 2023 Eurostat annual safety report provides a comprehensive overview of railway safety trends in the European Union. Despite a slight increase in 2022, the overall number of significant railway

RAGE-KG 2025: The Second International Workshop on Retrieval-Augmented Generation Enabled by Knowledge Graphs, co-located with ISWC 2025, November 2–6, 2025, Nara, Japan

shahrom.sohi@wu.ac.at (S. Sohi); dragomir.balan@s.wu.ac.at (D. Balan); amin.anjomshoaa@wu.ac.at (A. Anjomshoaa); axel.polleres@wu.ac.at (A. Polleres)

0009-0000-0735-1645 (S. Sohi); /0000-0001-6277-742X (A. Anjomshoaa); https://orcid.org/0000-0001-5670-1146 (A. Polleres)

accidents has decreased by 29.7% since 2010 — from 2,227 in 2010 to 1,565 in 2023. In particular, more than half of the fatalities in 2023 involved unauthorized persons on the tracks (58.4%), while a further 26.6% occurred at level crossings[6].

Within the railway sector, the European Union Agency for Railways (ERA) works closely with National Investigation Bodies (NIBs) across member states[7]. NIBs are mandated to investigate serious accidents and, at their own discretion, other accidents and incidents[8] in order to determine the immediate causes of an accident and uncover root causes.

Traditionally, accident investigation methods have varied between NIBs. However, with the growth of cross-border rail traffic, the need for a harmonised approach has become increasingly important. ERA supports NIBs by developing guidance on common investigation methodologies, facilitating platforms for sharing best practices, and offering tailor-made services such as investigator training and assessments of investigation practices [7].

In this complex realm of several NIBs it is relevant to maintain a harmonised view of incidents and to approach safety holistically within the Single European Transport Area. A consistent, cross-border perspective allows stakeholders to identify systemic risks, monitor emerging safety trends, and coordinate preventive measures. Semantic technologies are central to this process, enabling the integration of diverse data sources ranging from structured databases to unstructured accident narratives into a coherent, machine-readable format [9]. By applying standardised ontologies and entity linking, these technologies ensure a uniform representation of safety information, enhance interoperability between national systems, and support advanced analytics for informed decision-making.

2.1. Knowledge extraction from accident reports

Natural Language Processing and Text Mining techniques are used for extracting structured data from unstructured accident reports. These technologies enable automated coding of reports, classification, and analysis of narrative content, thereby enhancing the understanding of complex accident scenarios.

The work of Christophe Pimm demonstrates the use of NLP to systematically code the Air France accident report. The study showcases how automated information extraction can support human experts by facilitating the labelling and analysis of report content, reducing manual workload, and improving consistency [1].

A notable advancement in multilingual NLP applications is found in the work on the Swiss Railway, which addresses the challenges of processing safety reports written in different languages [10]. Their research focuses on railway accident data in a multilingual context, developing an information extraction technique to construct a trilingual knowledge base centred on passenger-related incidents. This system supports querying and analysis across English, French, and German, thereby significantly broadening the accessibility and utility of the extracted insights. Importantly, working in a multilingual setting introduces complexities far beyond those encountered in English-only processing, including differences in syntax, semantics, idiomatic expressions, and domain-specific terminology across languages, which make accurate extraction and integration more challenging.

With the advent of LLM, knowledge extraction has entered a transformative phase. Early experiments with LLMs, even without task-specific fine-tuning, show remarkable performance in identifying causation patterns, contributing to the faster and more accurate construction of knowledge graphs [11].

Moreover, LLMs facilitate the acceleration and automation of report processing pipelines. Tests using single-type LLMs demonstrate that general-purpose models, when applied appropriately, can outperform many specialized approaches. This not only highlights the technological maturity of LLMs but also points to their potential in revolutionizing safety data analysis by enabling scalable, multilingual, and intelligent systems [12].

3. ERAIL Extraction Method

The methodology used for extracting structured knowledge from the ERAIL database involves automatic acquisition, semantic preprocessing, and iterative retrieval-refinement using two LLMs. Accident

investigation reports and their associated metadata were retrieved from the ERA website. The full-text reports were segmented into smaller chunks. These chunks were then transformed into vector embeddings using "all-mpnet-base-v2" model, enabling efficient similarity retrieval within a vector database.

Three primary document types were used in the knowledge extraction framework:

- Investigation Reports: full text descriptions of accident cases, including summaries as PDF files.
- EU Rail Vocabulary: a structured glossary of contributing and systemic factors used to classify events, functioning as a Retrieval-Augmented Generation (RAG).
- Metadata Files: structured fields detailing variables such as accident type, date, location, and operational context.

The extraction pipeline is designed to simulate domain-informed reasoning using an LLM within a retrieval-augmented framework. The pipeline includes Vector store initialization, all chunked report embeddings are stored in a vector database for similarity search. Initial retrieval for given query entities, the system retrieves the top three semantically relevant documents (e.g. the top 3 accident cause). Contextual refinement, the results are passed through a second stage LLM prompt, which incorporates structured inputs from the accident vocabulary. This guides the model toward accurate classification using predefined categories and "Top-1" prediction. This iterative approach enables the system to refine its predictions and improve classification accuracy by learning from structured feedback. After these steps it has implemented the results into a Neo4j data store to have visual representation of the accidents. The complete process from raw report ingestion to structured knowledge extraction is illustrated in 1. The use of embeddings, contextual prompts, and feedback loops ensures adaptability and scalability for broader transport safety datasets.

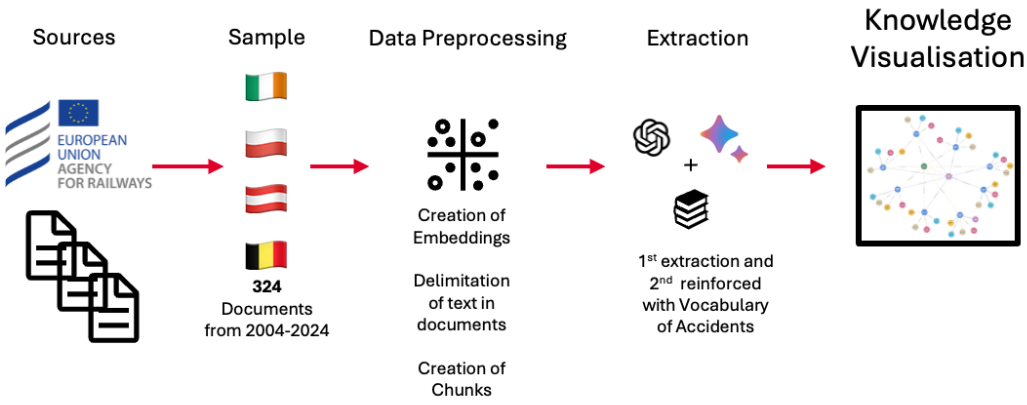

Figure 1: Method of Railway Accident Entity Extraction using GPT 4o mini and Gemini 2.5 pro models.

4. Preliminary Results

In the initial iteration, the task focused on extracting 353 documents reporting serious railway accidents from the years 2002 to 2024, across four European countries: Austria, Belgium, Ireland, and Poland. These countries were selected intentionally to create a multilingual environment, enabling evaluation of LLMs under diverse linguistic conditions.

The key elements extracted from each accident report are:

- Time of the event
- Date of the event
- Country of occurrence

- Accident type
- Regulatory body
- Contributing factors
- Systemic factors

Following extraction, a web application was developed to process the PDFs and visualize results and the results are automatically stored and compared across models using the following evaluation metrics:

- **Accuracy**: Defined as the degree to which the extracted entities match the official ERAIL database entries. Only exact matches are considered valid, which creates a strict benchmark. For instance, ”Derailment of a train” and ”Derailment of one or more rail vehicles” are treated as distinct.
- **Completeness**: Evaluates whether all target elements are successfully extracted from each report, e.g., 7 out of 7 elements.

The processed data were visualized using Neo4j graph, allowing cross-country visual comparisons of extracted content. Common contributing factors included communication failures, lack of awareness, and environmental conditions. Common systemic factors included weaknesses in operational planning, inadequate safety policies, and poor quality control. The results revealed an average accuracy of 34% across both GPT and Gemini models. The model scores low due to the strict exact-match requirement in the evaluation protocol, this number should be interpreted cautiously. For instance, textual variations or synonymous phrasing lead to mismatches even when the semantic meaning is identical.

The completeness metric showed an average of 95.5% across models, indicating that nearly all expected fields were successfully extracted. Date extraction had an accuracy of 90%, while time extraction reached only 43%. When relaxed to a tolerance of two minutes, time accuracy increased to 62.6%, and with five minutes tolerance, it rose to 68.7%. Lower time accuracy is likely due to the narrative style of reports, which often include multiple timestamps, making precise identification difficult. Moreover another reason why the time is accuracy is low is due to lower quality of Manual Metadata. After manual verification 50 reports contain incomplete metadata information on time of occurrence.

Contributing factors, and systemic factors yielded lower accuracy scores. This was largely because these entities are either absent from the ERAIL metadata or described using vocabulary that differs from the controlled terminology used in the ISS vocabulary. There results show a need to improve the metadata at EU level in accordance to EU Rail Vocabulary.

Furthermore, English language reports are more consistent having less mismatch information extracted. Below the details of mismatch for each country.

- Austria: 44.5%
- Poland: 23.6%
- Belgium: 19%
- Ireland: 12.9%

Further analysis indicated that the accident types most prone to mismatches were level crossing accidents, derailments, and train collisions. These categories are more complex in description and causality, leading to higher misclassification rates (e.g. the derailment can be a consequence of several events which are included in the report).This also suggests future adjustments in the EU Rail Vocabulary.

5. Conclusion and Future Work

In summary, this preliminary study demonstrates both the promise and the limitations of applying LLMs for structured information extraction in safety-critical domains such as railway transport. While the models exhibit strong performance in extracting basic metadata fields (e.g., date, location, country), they continue to face challenges when identifying more complex causal and systemic structures—especially across documents with varying formats, languages, and terminologies.

RAGs enhance information extraction by retrieving the top three most relevant chunks per query from a vector store, thereby narrowing the scope of outputs. This process reduces both processing costs and time requirements.

Limitation are present on certain documents that are PDF formats generated through scanning. While modern LLMs offer capabilities for interpreting scanned text, the current pipeline depends on retrieving text chunks, this process makes image-based PDFs incompatible. Overcoming this constraint will be a focus of future work.

Future enhancements will also focus on several areas: first, improving evaluation metrics to better reflect semantic similarity and partial matches will help provide a more nuanced understanding of model performance. Second, additional concepts related to accident causation, safety interventions, and incident resolution will be noted to enrich the metadata. Third focus on consistency of results evaluating multiple instances with same models. Finally, once these concepts are formalized, a comprehensive knowledge graph of European railway safety will be developed.

Acknowledgments

The contribution of this paper was supported by the Austrian Federal Ministry for Innovation, Mobility and Infrastructure (BMIMI) under the endowed professorship for "Data-Driven Knowledge Generation: Climate Action".

Declaration on Generative AI

Beside the core of the paper which is knowledge extraction with LLMs, during the preparation of this work, the author(s) used also GPT-4 in order to: Grammar and spelling check, and bug corrections. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] C. Pimm, C. Raynal, N. Tulechki, E. Hermann, G. Caudy, L. Tanguy, Natural Language Processing (NLP) tools for the analysis of incident and accident reports, in: International Conference on Human-Computer Interaction in Aerospace (HCI-Aero), 2012, Brussels, Belgium., 2012. URL: https://www.researchgate.net/publication/280751863_Natural_Language_Processing_NLP_tools_for_the_analysis_of_incident_and_accident_reports.

[2] C. Liu, S. Yang, Using text mining to establish knowledge graph from accident/incident reports in risk assessment, Expert Systems with Applications 207 (2022) 117991. URL: https://www.sciencedirect.com/science/article/pii/S0957417422012179. doi:10.1016/j.eswa.2022.117991.

[3] C. Y. Lam, K. Tai, Network topological approach to modeling accident causations and characteristics: Analysis of railway incidents in Japan, Reliability Engineering & System Safety 193 (2020) 106626. URL: https://www.sciencedirect.com/science/article/pii/S0951832019304247. doi:10.1016/j.ress.2019.106626.

[4] W.-T. Hong, G. Clifton, J. D. Nelson, Railway accident causation analysis: Current approaches, challenges and potential solutions, Accident Analysis & Prevention 186 (2023) 107049. URL: https://www.sciencedirect.com/science/article/pii/S0001457523000969. doi:10.1016/j.aap.2023.107049.

[5] F. Zhang, H. Fleyeh, X. Wang, M. Lu, Construction site accident analysis using text mining and natural language processing techniques, Automation in Construction (2019). doi:10.1016/j.autcon.2018.12.016.

[6] EUROSTAT, Railway safety statistics in the EU, 2023. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Railway_safety_statistics_in_the_EU.

[7] Directive (EU) 2016/798 of the European Parliament and of the Council of 11 May 2016 on railway safety (recast) (Text with EEA relevance), 2016. URL: http://data.europa.eu/eli/dir/2016/798/oj/eng.

[8] B. Accou, P. Guido, Elements of in-depth investigation - ERA expectations, Technical Report, 2023. URL: https://www.era.europa.eu/system/files/2023-11/Elements%20of%20in-depth%20investigation%20-%20ERA%20expectations.pdf?t=1754666783.

[9] T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web: A New Form of Web Content that is Meaningful to Computers will Unleash a Revolution of New Possibilities, in: O. Seneviratne, J. Hendler (Eds.), Linking the World's Information, 1 ed., ACM, New York, NY, USA, 2023, pp. 91–103. URL: https://dl.acm.org/doi/10.1145/3591366.3591376. doi:10.1145/3591366.3591376.

[10] P. Hughes, R. Robinson, M. Figueres-Esteban, C. van Gulijk, Extracting safety information from multi-lingual accident reports using an ontology-based approach, Safety Science 118 (2019) 288–297. URL: https://www.sciencedirect.com/science/article/pii/S0925753518307586. doi:10.1016/j.ssci.2019.05.029.

[11] L. Chen, J. Xu, T. Wu, J. Liu, Information Extraction of Aviation Accident Causation Knowledge Graph: An LLM-Based Approach, Electronics 13 (2024) 3936. URL: https://www.mdpi.com/2079-9292/13/19/3936. doi:10.3390/electronics13193936, number: 19 Publisher: Multidisciplinary Digital Publishing Institute.

[12] V. Toğan, F. Mostofi, O. Behzat Tokdemir, F. Kadioğlu, Efficient Management of Safety Documents Using Text-Based Analytics to Extract Safety Attributes From Construction Accident Reports, IEEE Access 13 (2025) 99758–99777. URL: https://ieeexplore.ieee.org/document/11023522. doi:10.1109/ACCESS.2025.3576442.