Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p27682-cd

A Data-Driven Framework for Systemic Risk Identification in Aviation Systems Using Large Language Models and Graph Learning

Changdong Zheng
China Aero-Polytechnology Establishment, China. E-mail: changdong_zheng@163.com

Wensheng Peng
China Aero-Polytechnology Establishment, China. E-mail: wshpeng@126.com

Shumao Qiu
China Aero-Polytechnology Establishment, China. E-mail: qiushumao777@163.com

Yu Hao
China Aero-Polytechnology Establishment, China. E-mail: adoly9@163.com

With the increasing complexity of modern aviation systems, high non-linearity and dynamic interactions introduce uncertainties that are difficult to enumerate using conventional expert-driven approaches, limiting their ability to capture emergent systemic risks. To address this challenge, this study proposes a data-driven framework that unifies Large Language Models (LLMs), graph learning, and system-theoretic perspectives to enable end-to-end risk identification from unstructured evidence. DeepSeek-R1-671B(Liu et al., 2024) are employed to extract structured event sequences from 22,800 aviation accident reports in the NTSB database, which are aggregated into a multi-layer knowledge graph representing fault evolution. A Graph Convolutional Network (GCN) is then applied to learn topological dependencies and identify high-risk propagation paths. By bridging semantic extraction and structural inference, the proposed framework enables traceable discovery of latent interaction patterns and provides a scalable, data-driven complement to traditional safety analysis methods for systemic risk identification.

Keywords: Large Language Models, Risk Identification, Graph Learning, Aviation Systems.

1. Introduction

With the digital transformation of modern aviation—from NextGen air traffic management to automated cockpits—safety-critical ecosystems have evolved into deeply integrated sociotechnical systems (Leveson, 2011). Characterized by tight coupling and non-linearity, safety risks in these systems increasingly stem from dysfunctional interactions across system boundaries—such as between crews and automation—rather than isolated component failures (Johnson et al., 2021). Identifying these implicit risks within intricate interaction networks has become a critical challenge that traditional expert-based analysis struggles to address (Scott, 2019; Mohaghegh, 2012).

Simultaneously, the accumulation of massive unstructured data, such as maintenance logs and near-miss records, provides a rich repository of real-world operational knowledge. These data often contain latent patterns that exceed the scope of predefined failure models. Recent breakthroughs in Large Language Models (LLMs) offer new opportunities to extract high-value information from this fragmented data (Zhao et al., 2023). By leveraging data-driven techniques to mine these logs, it is possible to compensate for the blind spots of expert knowledge and provide empirical evidence for discovering unknown emergent risks.

Despite abundant data, current aviation risk identification paradigms face two critical hurdles. First, the ”combinatorial explosion” of risk interactions limits traditional

models; static methods like FTA and FMEA rely on expert priors, creating "cognitive blind spots" regarding unknown, non-linear coupling(Ericson, 2015). Second, a "semantic gap" in unstructured data hampers extraction; noisy O&M logs exceed the capabilities of rule-based or shallow NLP methods, making it difficult to accurately reconstruct risk factors and temporal logic (Bertero et al., 2017).

To address these challenges, this paper proposes a data-driven framework bridging the process of "unstructured evidence acquisition—structural modeling—risk path mining." The framework leverages LLMs to extract event sequences from logs, aggregating them into a knowledge graph to characterize temporal and causal dependencies. Based on this, a GNN is trained to learn risk propagation patterns and mine high-risk paths as objective evidence. By synergizing semantic extraction with probabilistic discovery, this approach provides a scalable, traceable evidence chain for identifying latent hazards in complex aviation systems.

2. Related Work

This section provides an in-depth survey of natural language processing for log parsing and graph neural networks for risk modeling, identifying critical limitations that motivate our integrative framework.

2.1. LLM-Based Log Analysis & Event Extraction

Recent advances in large language models (LLMs), particularly transformer-based architectures like LogGPT (Qi et al., 2023) and LogBERT (Guo et al., 2021) have revolutionized log parsing and anomaly detection by treating logs as temporal sequences. Recent methods further utilize in-context learning (Xue and Salim, 2023) and event-level extraction (Stein Dani et al., 2024) for zero-shot understanding. However, these models typically treat logs as flat, linear sequences, ignoring the structural dependencies and architectural topology of the underlying system. This lack of awareness regarding hierarchical design and distributed control semantics makes them insufficient for reasoning about systemic risks or cascading failures.

2.2. Graph Learning (GNN) for Fault Diagnosis & Risk Assessment

GNNs have shown strong potential in modeling complex dependencies via techniques like GDN (Chen et al., 2021) for anomaly detection and fault localization. CausalGNN (Wang et al., 2022) further introduces causal inference and continuous-time dynamics to capture evolving fault chains. Despite their predictive power, most GNN methods remain "black-box" models that lack alignment with interpretable safety logic, offering limited insight into the mechanisms of risk propagation. Furthermore, few explicitly address epistemic uncertainty or out-of-distribution detection, which are essential for safety-critical environments where reconfigurable architectures and evolving failures are prevalent.

2.3. Summary and Research Gap

Existing research remains fragmented: NLP-based methods resolve semantic ambiguity but lack structural awareness, while graph learning captures topological non-linearity but often operates as an uninterpretable "black box." This disjoint prevents a seamless transition from unstructured narratives to structured risk analysis, leaving a synergized LLM-GNN framework yet to be realized for aviation safety. By establishing a mathematical mapping from unstructured evidence to graph-based representations, this study bridges the semantic-topological gap. Integrating the linguistic depth of LLMs with the structural inference of GNNs constructs a unified framework that is both data-driven in its scalability and traceable in its evidentiary reasoning.

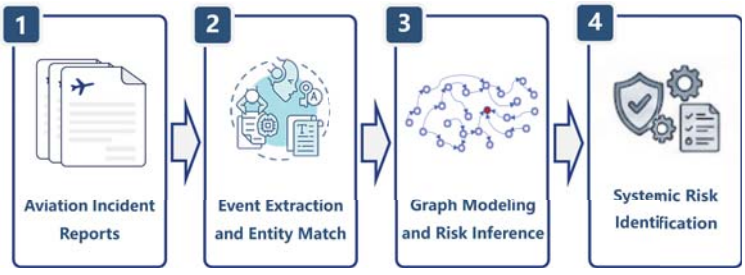

Fig. 1. A Data-Driven Framework for Systemic Risk Identification.

3. Methodology

3.1. Preliminaries and Problem Definition

This section formalizes the aviation risk identification task as a sequence of transformations across heterogeneous spaces. The methodology maps unstructured textual data into a structured topological space, then into a continuous feature space, and finally into an augmented evidence space for risk discovery. The overall pipeline is summarized by the following mapping:

$$\ell_k \xrightarrow{f_\theta} (V_k, E_k) \xrightarrow{\uplus} \mathcal{G} = (V, E),\ \mathbf{X} = \phi(V) \xrightarrow{g_\psi} \hat{\mathbf{Y}},\ \text{Unc} \xrightarrow{\text{Augment}} \mathcal{G}^+, \tag{1}$$

where ℓ_k denotes the raw report, f_θ is the extraction operator, $\mathcal{G}$ is the consolidated evolution graph, g_ψ is the risk inference model, and $\mathcal{G}^+$ represents the augmented graph containing identified risk paths and evidentiary links.

Let the input collection of unstructured aviation logs be $\mathcal{L} = \{\ell_k\}_{k=1}^{K}$. Our objective is to solve two interconnected sub-problems:

This framework follows a streamlined two-stage pipeline for aviation risk identification, as illustrated in Figure 1. Our objective is to solve the following interconnected sub-problems:

(1) Event Extraction and Entity Match: Extract multi-layered event elements (V_k, E_k) from raw logs via LLMs and perform global matching to unify them into a consolidated fault-evolution graph $\mathcal{G} = (V, E)$. This stage transforms discrete, unstructured semantics into a unified topological space.

(2) Graph Modeling and Risk Inference: Apply a GNN-based architecture to learn structural dependencies and estimate multi-label risk probabilities $\hat{\mathbf{Y}}$. By quantifying epistemic uncertainty Unc, the framework triggers graph augmentation to bridge informational gaps, outputting an augmented graph $\mathcal{G}^+$ with prioritized high-risk failure paths as a supplement to FHA/PSSA processes.

3.2. Event Extraction and Entity Match from Logs via LLMs

This study utilizes Large Language Models (LLMs) as extraction operators to convert unstructured logs ℓ_k into structured local graphs (V_k, E_k). Here, V_k contains event nodes classified into physical, information, and cognitive layers, while E_k represents temporal or causal dependencies. To analyze emergent risks across the dataset, local graphs are unified into a global fault-evolution graph $\mathcal{G} = (V, E)$ using a global indexing map.

To enable quantitative analysis, each node $v \in V$ is projected into a continuous embedding space $\mathbb{R}^d$ via a semantic mapping $\phi(v)$. This step transforms qualitative fault descriptions into feature tensors $\mathbf{X}$, serving as the foundational input for downstream graph learning.

3.3. Graph Modeling and Risk Inference via GNN

To capture nonlinear interactions within the system topology, a Graph Convolutional Network (GCN) is employed. The GCN aggregates neighborhood information to emulate fault propagation behaviors(Kipf and Welling, 2016):

$$\mathbf{H}^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right) \quad (2)$$

The final layer outputs a node-level risk probability matrix $\hat{\mathbf{Y}}$ via a sigmoid function. The model is trained using a multi-label binary cross-entropy objective to identify high-risk interaction patterns.

To address data sparsity and unseen failure modes, we quantify epistemic uncertainty using Monte Carlo (MC) Dropout. By performing T stochastic forward passes, the predictive mean μ and variance σ^2 are used to define a confidence indicator:

$$\text{Conf} = \mu \odot \left(1 - \lambda\sigma^2\right) \quad (3)$$

When confidence is low, the framework triggers graph augmentation. By searching for low-cost explanatory paths in a global knowledge base using shortest-path algorithms, the framework generates an augmented graph $\mathcal{G}^+$. These paths bridge missing links in raw logs, providing a traceable and prioritized evidence chain for identifying latent systemic risks.

4. Results and Discussion

This section shows the results, and the accident named in-flight engine shutdown and fuel leak during cruise are selected from Aviation Safety Report System (ASRS) open database, which is considered as a good validation tool for framework effectiveness out of the training data distribution.

4.1. Knowledge Graph Construction and Learning

To transform unstructured narratives into a computable format, this study uses DeepSeek-R1-671B to generate structured JSON data and organize events into a three-layer hierarchy: L1 Physical (objective states), L2 Information (cockpit displays/warnings), and L3 Cognitive (human decision-making). As illustrated in Figure 2 using NTSB report "20170902X54656", this hierarchy reveals accident logic from environmental states to subjective responses. Processing 22,800 NTSB reports produced 306,020 nodes and 312,165 edges, which were merged into a global aviation accident knowledge graph following the method in Section 3.2. This unified graph supports the discovery of cross-incident risk patterns hidden in isolated narratives.

Table 1. Statistical properties and topological metrics of the generated Knowledge Graph.

Metric description	Value
Scale and Efficiency	
Total Canonical Nodes ($\|V\|$)	162,991
Total Merged Edges ($\|E\|$)	293,453
Compression Ratio	1.84×
Topological Features	
Average Degree ($\langle k \rangle$)	3.6008
Network Density (D)	1.105×10^{-5}
Weakly Connected Components (WCC)	2,934
Average Path Length (APL)	6.52
Layer-wise Node Distribution	
L3: Cognitive Nodes	54,516
L2: Information Nodes	33,526
L1: Physical Nodes	74,949

Table 1 summarizes the Knowledge Graph's statistical and topological properties. In terms of Scale and Efficiency, the KG integrates $|V| = 162,991$ canonical nodes and $|E| = 293,453$ edges, achieving a 1.84× compression ratio. The Topological Features reveal a sparse network (density: 1.105×10^{-5}) with an average degree of 3.6008 and 2,934 weakly connected components. An Average Path Length of 6.52 reflects its underlying connectivity. The Layer-wise Distribution underscores a multi-tiered architecture: the Physical Layer (L1) is the most populous (74,949 nodes),

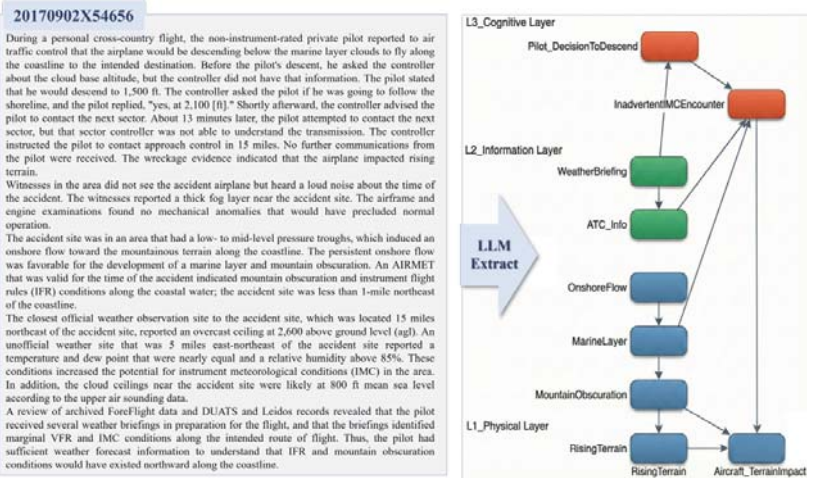

Fig. 2. Extract entities and relationships from aviation accident reports.

followed by the Cognitive (L3: 54,516) and Information (L2: 33,526) layers. This distribution confirms the graph's capacity to model complex interactions between physical environments, informational flows, and cognitive processes.

This study constructs a manually-labeled benchmark for uncertainty-aware aviation risk analysis, where six categories are considered, including six epistemic or aleatory uncertainty labels. The dataset is split into training and validation sets (9:1), and accuracy is computed as the node-level prediction accuracy of these labels.

As shown in the learning curves (Fig. 3), the impact of structured data is decisive: while the raw text baseline (DNN-Text) plateaued at 73–74% accuracy, graph-based models (DNN-Graph and GNN-Graph) consistently achieved 75–77%. This ≈2.5% performance gap validates that LLM-based extraction effectively filters background noise to preserve the semantic core. Furthermore, while the GNN-Graph demonstrates accuracy comparable to the high-quality feature-driven DNN-Graph, its critical advantage lies in incorporating graph topology. This enables the model to retain interpretability via attention mechanisms without sacrificing the predictive power lost in the "black-box" feature aggregation of traditional DNNs.

4.2. Illustrative Example: In-flight Engine Shutdown and Fuel Leak During Cruise

The case study focuses on an ASRS case in Figure 4, where a cruise-phase emergency involving a No. 1 engine fuel leak, characterized by rapid instrument alerts followed by an uncommanded engine shutdown within 60 seconds. Although the flight crew successfully managed an emergency diversion and achieved a safe landing at ZZZ airport, the standard incident report primarily documents explicit mechanical failures and procedural compliance.

In this case study, an LLM was first utilized to extract ten pivotal event nodes from the raw accident report, forming the initial core graph. To uncover latent risks, a matching mechanism retrieved 20 candidate paths from the expert knowledge base, which were subsequently prioritized by the GNN model based on the entire path structure. Due to the inherent sparsity of the operational data, the model yielded conservative confidence scores; therefore, a threshold of 0.2 was established for path filtering. This process successfully

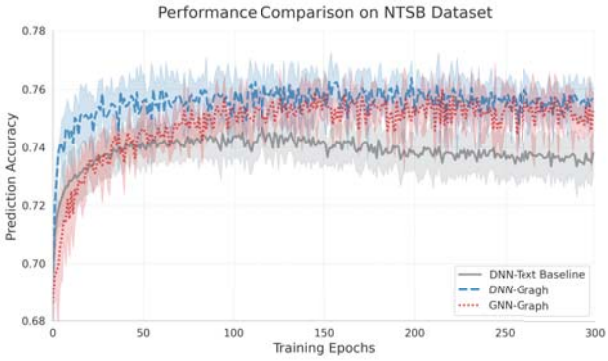

Fig. 3. Prediction Accuracy during model training.

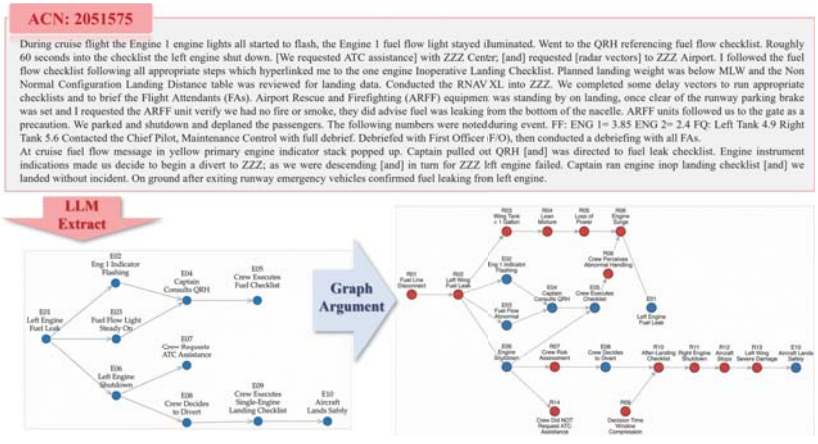

Fig. 4. Graph augment on out-distribution aviation accident report.

expanded the incident graph to 22 nodes, with newly identified entities highlighted in red in Fig. 4.

Table 2 provides a granular breakdown of the top eight augmented paths, including their respective confidence scores, logic types, and systemic contributions. These increments range from single-node root cause identification (e.g., specific hardware failure modes) to multi-node procedural restorations. By strategically classifying these paths into dimensions such as cognitive logic, physical state transitions, and latent hardware interactions, the framework collectively enhances the diagnostic resolution of the incident and provides a traceable evidence chain that exceeds the scope of the original unstructured narrative.

The qualitative value of this augmentation lies in uncovering latent factors omitted from standard reports but critical for high-fidelity assessment. For example, Path 5 introduces "Decision Time Window Compression," providing a human-factors explanation for operational urgency often obscured in procedural logs. Path 4 identifies "Fuel Boost Pump High Setting," suggesting a systemic conflict between standard operations and the high-pressure leak. Together with tactile feedback (Path 6) and the specific failure mode (Path 8: Fuel Line Disconnect), the framework extends the analysis from surface descriptions to unobserved systemic risks. These insights

Table 2. Knowledge Graph Augmentation: Incremental Nodes, Logic Types, and Supplemental Explanations.

ID	Key Incremental Nodes	Rank	Logic Type	Supplemental Explanation
Path 1	Crew Risk Assessment	0.71	Causal backtracking	Cognitive logic supplement: Fills the gap of mental assessment between fault detection and decision-making.
Path 2	Engine Thrust Reduction	0.62	Causal backtracking	Physical state supplement: Smooths the thrust degradation process between engine operation and shutdown.
Path 3	After-Landing Checklist, Right Engine Shutdown, Aircraft Stops, Left Wing Severe Damage	0.44	Gap filling	Procedure and risk supplement: Fully restores the safety sequence post-landing and predicts potential structural damage risks.
Path 4	Fuel Boost Pump High Setting, Fuel Flow Abnormal	0.36	Causal backtracking	Latent component supplement: Identifies hardware control nodes highly related to fuel leaks not mentioned in the original text.
Path 5	Decision Time Window Compression	0.33	Causal backtracking	Human factors supplement: Introduces psychological stress factors that lead to crew operational urgency.
Path 6	Crew Perceives Abnormal Handling, Crew Executes Checklist	0.32	Causal backtracking	Feedback loop supplement: Enriches information acquisition dimensions (from visual instruments to tactile handling feel).
Path 7	Engine Surge	0.30	Causal backtracking	Failure symptom supplement: Refines physical signs before flameout, providing evidence for human factor performance analysis.
Path 8	Fuel Line Disconnect	0.21	Causal backtracking	Root cause supplement: Specifies the general "leak" as a concrete physical failure mode of the fuel line.

reveal misalignments between the crew's mental model and physical reality, supporting the development of more resilient safety protocols.

5. Conclusion

This paper presents a novel data-driven framework for systemic risk identification in aviation, bridging the gap between unstructured operational data and structural safety analysis. By synergizing LLM-based semantic extraction with GNN-based topological reasoning, the study achieves a multi-dimensional deconstruction of complex failure evolution, providing a scalable and objective supplement to traditional expert-led FHA/PSSA processes.

Case validation demonstrates that the transition from unstructured logs to a structured knowledge graph effectively uncovers emergent risks and latent causal links—such as cognitive pressure and implicit operational conflicts—that are frequently omitted in standard reports. This approach enables a paradigm shift from focusing on isolated component failures to identifying complex,

interaction-based risks through automated path mining and uncertainty estimation.

Future research will focus on integrating physical priors to improve model generalization in low-data regimes, developing LLM-based reasoning for safety constraint, and incorporating insights into MBSE workflows for closed-loop safety governance. Ultimately, this framework provides a transferable template for other safety-critical industries to leverage large-scale operational data for more effective, intelligent safety solutions.

Acknowledgement

This work has been supported by the National Natural Science Foundation of China (Grant No. U2541220).

References

Bertero, C., M. Roy, C. Sauve, and D. Trudel (2017). Experience report: Log mining using natural language processing and application to integral technology incidents. Software Reliability Engineering (ISSRE), 2017 IEEE 28th International Symposium on, 351–360.

Chen, Z., D. Chen, X. Zhang, Z. Yuan, and X. Cheng (2021). Learning graph structures with transformer for multivariate time-series anomaly detection in iot. IEEE Internet of Things Journal 9(12), 9179–9189.

Ericson, C. A. (2015). Hazard Analysis Techniques for System Safety (2nd ed.). John Wiley & Sons.

Guo, H., J. Li, H. Shen, and J. Zhu (2021). Logbert: Log anomaly detection via bert. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 2825–2832.

Johnson, E. B., A. N. Kopeikin, N. G. Leveson, and A. W. Drysdale (2021). Hazard analysis for human supervisory control of multiple unmanned aircraft systems. In 56th International Symposium on Aviation Psychology.

Kipf, T. N. and M. Welling (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Leveson, N. G. (2011). Engineering a Safer World: Systems Thinking Applied to Safety. Cambridge, MA: MIT Press.

Liu, A., B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, et al. (2024). Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437.

Mohaghegh, Z. (2012). Development of an aviation safety causal model using socio-technical risk analysis (soteria).

Qi, J., S. Huang, Z. Luan, S. Yang, C. Fung, H. Yang, D. Qian, J. Shang, Z. Xiao, and Z. Wu (2023). Loggpt: Exploring chatgpt for log-based anomaly detection. In 2023 IEEE international conference on HPCC/DSS/SmartCity/DependSys, pp. 273–280. IEEE.

Scott, B. B. I. (2019). Aviation law and drones: Unmanned aircraft and the future of aviation. Zeitschrift fuer Luft- und Weltraumrecht(ZLW) 68(2), 342–344.

Stein Dani, V., M. Dees, H. Leopold, K. Busch, I. Beerepoot, J. M. E. van der Werf, and H. A. Reijers (2024). Event log extraction for process mining using large language models. In International Conference on Cooperative Information Systems, pp. 56–72. Springer.

Wang, L., A. Adiga, J. Chen, A. Sadilek, S. Venkatramanan, and M. Marathe (2022). Causalgnn: Causal-based graph neural networks for spatio-temporal epidemic forecasting. In Proceedings of the AAAI conference on artificial intelligence, Volume 36, pp. 12191–12199.

Xue, H. and F. D. Salim (2023). Promptcast: A new prompt-based learning paradigm for time series forecasting. IEEE Transactions on Knowledge and Data Engineering 36(11), 6851–6864.

Zhao, W. X., K. Zhou, J. Li, et al. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223. Highly cited survey on LLMs.