Causal Graph Reasoning for Explainable ADREP Classification

Whoopie Wanjiru
MSEAI27
wwanjiru@andew.cmu.edu

Theophilus Owiti
MSEAI27
towiti@andrew.cmu.edu

Ronnie Delyon
MSEAI26
rdelyon@andrew.cmu.edu

1 Problem Statement

Civil aviation authorities worldwide face challenges in manually processing and classifying safety reports according to the International Civil Aviation Organization (ICAO) Accident/Incident Reporting (ADREP) taxonomy, resulting in inconsistent labeling, processing delays, and missed opportunities for trend detection. The CMU-Africa Aviation Safety Data Processing and Classification System (SD-CPS) project[2] demonstrated effective automation using a hybrid transformer-multi-LLM pipeline with two BERT models and three LLMs with varied temperatures feeding a 7-rule consensus engine achieving 92.96% Longformer accuracy, 89.81% LLM agreement, and 2.15-s processing on 3,600 reports.

However, 26.94% of reports were conservatively classified as **"OTHER"**, limiting downstream safety analysis and trend detection. These cases typically correspond to narratives with implicit, multi-factor, or weakly expressed causal structure, where flat text classification lacks sufficient relational grounding.

Supervised NLP approaches like those in New and Wallace[5] achieve high precision on labeled data but struggle with rare, multi-factor incidents lacking explicit relational modeling. Similarly, Wang et al.'s[7] self-consistency enhances LLM reliability on flat text but overlooks chronological causal chains critical for aviation analysis.

Rather than redesigning the full system or reprocessing all reports, we propose a targeted refinement step applied only to reports labeled OTHER. This step introduces lightweight causal event extraction and graph-based reasoning to recover latent structure and attempt secondary classification into specific ADREP categories. This approach preserves the existing validated pipeline while focusing computational and modeling complexity precisely where ambiguity remains highest.

2 Data

The project utilizes public aviation safety datasets totaling 68,000–76,000 reports from sources like the Aviation Ssafety Network[1] (3,068 silver-labeled for supervised training via GPT-4o CoT prompting to ICAO ADREP's 11 classes, e.g., CFIT, LOC-I), NASA ASRS[4] (24,000 voluntary informal narratives), American National Transportation Safety Board[6] (44,000 formal high-severity cases) and American Federal Aviation Administration[3] subsets stored in a local database after custom preprocessing to handle jargon, long sequences (up to 4k tokens), and taxonomy mismatches. These span the safety pyramid from minor incidents to accidents, enabling hybrid model training and 3,600-report consensus evaluation (95.06% auto-accept, mean 2.15s processing), though EDA revealed challenges like class imbalance (OTHER 26.94%, CFIT 20.78%), no gold-truth labels, and the need for silver-labeling to bridge gaps for consistent ADREP mapping and trend detection.

3 Method

Our approach augments the existing SDCPS architecture with a post-classification causal reasoning module applied exclusively to reports initially classified as OTHER. The goal is not to replace or retrain the core classifier, but to re-express ambiguous narratives in structured causal form and reassess their ADREP alignment.

3.1 System Architecture

Raw narratives are first processed by the existing SDCPS pipeline (supervised transformers, zero-shot semantic classification, and 7-rule multi-LLM consensus). Only reports assigned the OTHER category proceed to the causal refinement stage, which consists of the following steps:

- **Layer 1 – Event Extraction**: Aviation-BERT-NER or RoBERTa-CRF extracts key entities from OTHER reports, including ACTOR, SYSTEM, PHASE, TRIGGER, and OUTCOME.
- **Layer 2 – Causal Graph Construction**: Extracted entities are assembled into a heterogeneous Directed Acyclic Graph (DAG) representing hypothesized cause–effect relationships (e.g., Ice → Engine Stall → Diversion).
- **Layer 3 – Graph Reasoning**: A Graph Attention Network (GAT) or Heterogeneous Graph Transformer (HGT) performs message passing over the graph to learn an "incident signature" capturing relational and temporal dependencies.
- **Layer 4 – ADREP Reclassification**: The graph-level embedding is mapped to ICAO ADREP categories. If confidence exceeds a predefined threshold, the report is reassigned from OTHER to a specific category; otherwise, it remains OTHER and is flagged for analyst review.

This causal refinement operates as an eighth decision rule, invoked only when prior consensus fails to reach a specific ADREP classification.

3.2 Key Improvements

This targeted approach:

- Reduces overuse of the OTHER category by exploiting latent causal structure
- Preserves the original SDCPS accuracy, latency, and validation guarantees
- Provides analysts with interpretable causal graphs for ambiguous cases
- Avoids the cost and risk of reprocessing the full corpus or retraining core models

4 Evaluation

Evaluation focuses specifically on reports initially classified as **OTHER**. Success is measured by:

1. OTHER reduction rate (percentage of OTHER reports reassigned to specific ADREP categories)
2. Classification confidence stability (agreement between graph-based predictions and LLM consensus where available)
3. Structural validity checks (rejection of physically implausible reclassifications)
4. Macro-F1 on silver-labeled subsets of reclassified reports
5. Operational impact metrics, including latency overhead (<0.5 s per OTHER report) and unchanged auto-accept rates for non-OTHER cases.

With this evaluation strategy, we align with the prior aviation NLP work emphasizing robustness and consistency over absolute accuracy when expert gold labels are unavailable.

References

[1] Aviation Safety Network. Aviation safety network database. `https://aviation-safety.net/database/`, 2026. Accessed: 2026-02-03.

[2] Ronnie Delyon, Baraka Manyara, Nasiru Iliya, and Thuo Gachomba. Intelligent aviation safety data analysis using transformer and multi-LLM consensus. In *Proc. CMU-Africa Capstone*, Kigali, Rwanda, 2026.

[3] FAA. Federal aviation administration. `https://www.asias.faa.gov/apex/f?p=100:189:::NO`, 2026. Accessed: 2026-02-03.

[4] NASA ASRS. Nasa asrs database online. `https://asrs.arc.nasa.gov/search/database.html`, 2026. Accessed: 2026-02-03.

[5] M. D. New and R. J. Wallace. Classifying aviation safety reports using supervised natural language processing. *Safety*, 11(1):7, March 2025.

[6] NTSB. National transportation safety board. `https://data.ntsb.gov/avdata`, 2026. Accessed: 2026-02-03.

[7] Xuezhi Wang, Jason Wei, Denny Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proc. ICLR*, Kigali, Rwanda, 2023.