An AI Expert Council for Explainable HFACS Analysis in General Aviation

Abstract

Aviation accident investigation is a core obligation for operators because it enables the prevention of future occurrences. However, current investigations remain largely traditional and time-intensive, which delays assessment and safety improvements. Building on prior work in accident and human-factors analysis, this study proposes an automated pipeline that leverages large language models (LLMs) to code aviation accidents under the Human Factors Analysis and Classification System (HFACS). To address single-model bias and the need for explainability, we design a multi-agent architecture—an HFACS AI Expert Council—in which technical and organizational specialists analyze the same evidence, debate, and converge on a decision. The council's qualitative rationales are then scored with a transparent rubric to ensure consistent and auditable outcomes. To rigorously evaluate the system, we develop a high-fidelity data simulator capable of generating diverse accident scenarios for synthesis and cross-checking. On a 200-run simulated dataset, the system attains an overall F1-score of 0.82, with strong performance on Level 3 (Unsafe Supervision) **(**F1 = 0.85**)** and Level 2 (Preconditions) **(**F1 = 0.66**),** and promising progress on Level 4 (Organizational Influences) **(**F1 = 0.43**).** These results indicate the feasibility of a reliable, explainable AI assistant that can accelerate analysis and support improvements in aviation safety.

Key words: AI, ChatGPT, HFACS, LLM.

1. Introduction

General aviation (GA) is a cornerstone of civil aviation, enabling missions that scheduled air transport cannot readily provide—humanitarian relief, emergency medical evacuation, environmental monitoring, and flexible cargo/personnel movement. Despite this central role, GA has faced persistent safety challenges for decades. At the investigative level, the NTSB indicates that a standard investigation typically spans 12–24 months [1], depending on case complexity and workload, which heightens pressure to reduce analytic latency and to standardize information flows. Within this context, the Human Factors Analysis and Classification System (HFACS) remains the primary taxonomy for human-factors analysis, encompassing four levels: Level 1 —Unsafe Acts, Level 2 —Preconditions for Unsafe Acts, Level 3 —Unsafe Supervision, and Level 4 —Organizational Influences. However, manually coding large volumes of narratives (accident reports, witness statements, CVR excerpts, maintenance records) is time-consuming and prone to inter-rater inconsistency.

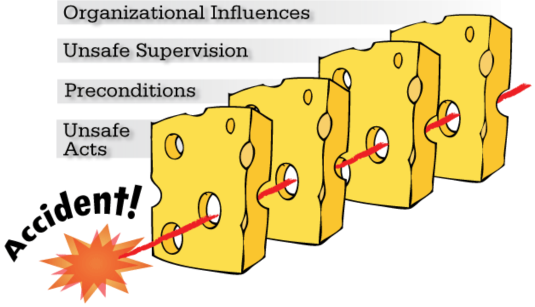


Figure 1: The Swiss Cheese Model

With the rapid progress of large language models (LLMs), it is now feasible to standardize and streamline aviation-safety analytics—from entity extraction and topic classification to summarization and rapid reporting—outperforming many traditional workflows and improving consistency across raters [2-4]. Early applications of LLMs to accident analysis suggest shorter investigative cycles and reduced subjectivity when a clear taxonomy and scoring protocol guide judgments.

In this study, we present an automated pipeline for HFACS-based coding of aviation accidents designed with three constraints: explainability, inter-run consistency, and low latency for near-real-time use. At its core is a multi-agent architecture (HFACS AI Expert Council) in which role-specialized agents reason in parallel; their rationales are reconciled by a referee and converted to auditable, level-specific decisions via a rubric-based scorer. We evaluate the system on 11 high-fidelity scenarios spanning all four HFACS levels over 200 controlled runs. The micro-averaged results are Precision = 0.58, Recall = 1.00, and F1 = 0.73, indicating a recall-oriented operating point suitable for triage while highlighting the need to raise precision (e.g., tighter thresholds and evidence grounding) before decision-grade deployment.

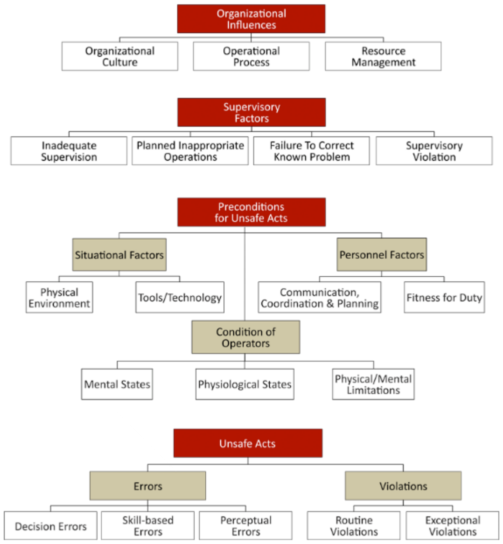


Figure 2: The HFACS (Human Factors Analysis and Classification System)

2. Related Work

Before LLMs, aviation-safety analytics largely relied on classical machine-learning pipelines—SVM, ANN/Decision Trees, and later CNN/RNN architectures—to classify occurrences and predict severity in civil-aviation datasets [5]. These models were useful but required extensive manually labeled corpora and struggled with the domain's dense technical context and abbreviations, limiting generalization across operators and report styles.

Modern large language models (LLMs) have shifted this landscape by enabling robust reasoning over unstructured narratives with zero-/few-shot supervision. Early studies report gains in information extraction and classification from safety reports (e.g., GPT-4 on construction accident narratives) and in transportation incident analysis more broadly, suggesting shorter analytic cycles and improved coding consistency [6]. Yet most efforts stop at generic prompting and do **not** operationalize deep, HFACS-aligned causal reasoning.

Addressing that gap, Liu et al. (2025) introduce HFACS-guided Chain-of-Thought (HFACS-CoT) and an enhanced HFACS-CoT+, which embeds domain knowledge into stepwise prompts to structure the model's reasoning and reduce logical slips [7]. In parallel, the literature on CoT with LLMs has expanded, offering broader evidence on how staged reasoning affects fidelity, error modes, and stability across tasks [8-10].

Taken together, prior approaches either depend on manual labels and lack domain semantics (traditional NLP/ML), use domain LMs/NER but do not integrate the HFACS taxonomy into the final reasoning step, or apply LLMs directly without controls for bias/consistency and without latency accounting for near-real-time use. Our work builds on these foundations and advances the state of practice by: deploying a debate-style multi-agent architecture with an arbiter to mitigate single-model bias and improve explainability; fusing text and telemetry so technical signals can be linked to human-factor findings; and lastly converting qualitative rationales into auditable, quantitative decisions via a rubric (HFACS_RUBRIC).

3. Methodology

3.1 Set up the framework

Figure 3.1 outlines a four-stage pipeline (includiing Inputs, Simulated, Analyze, and the last one is Outputs). A run is launched from the dashboard with a versioned scenario pack (JSON) and prompt; the simulator co-generates synchronized flight telemetry and narratives/maintenance notes. The analysis stage first screens for anomalies, then (only when triggered) convenes a multi-agent HFACS council whose outputs are reconciled by a referee and scored via a transparent HFACS rubric to yield level/tag decisions. Results are materialized as charts (PNG) and structured reports (CSV/JSON), while every run logs model version, inference parameters, seed, and timestamps to ensure reproducibility and end-to-end latency accounting—one traceable path from evidence to judgment with low overhead and clear avenues for scaling.

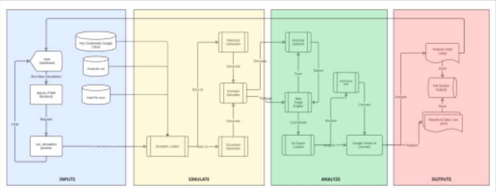
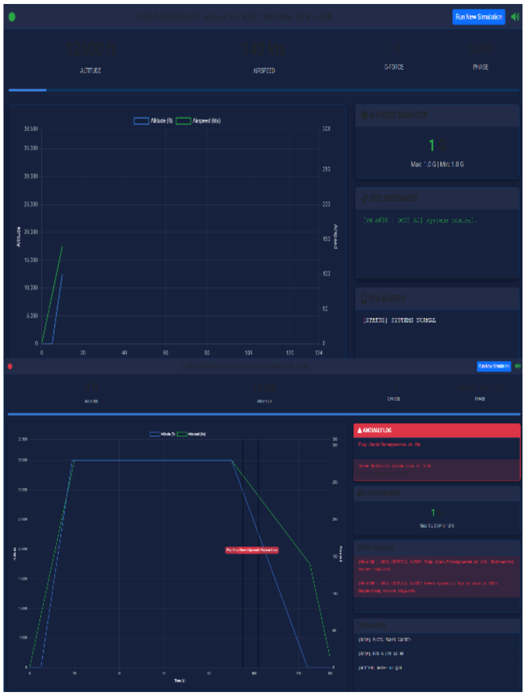


Figure 3.1 The structure of the research

3.2 Data Sampling

The dashboard plots altitude and airspeed against time with different scenarios. During descent, two events are time-stamped: Flap Stuck/Unresponsive (95 s) and Green Hydraulic System Loss (102 s**),** with concurrent OCC CRITICAL ALERT and ECAM/EFB advisories, creating a clear evidence trail from detection to crew guidance.

HFACS link (one-sentence use in text): These synchronized signals ground the analysis in HFACS Level 2 (Equipment & Controls/Technological Environment) and, depending on maintenance/dispatch context, may implicate Level 3 (Unsafe Supervision**)** or Level 4 (Organizational Influences)**;** absent an**y** QRH deviation, Level 1 (Unsafe Acts) is not primary.



[VN-A688 | OCC] All systems nominal.

[STATUS] SYSTEMS NORMAL

Flap Stuck/Unresponsive at 95s

Green Hydraulic System Loss at 102s

Figure 3.1 Flight dashboard with time-aligned telemetry and anomaly log (A320 VN-A688, SGN→DAD)

3.3 Model Performance Evaluation

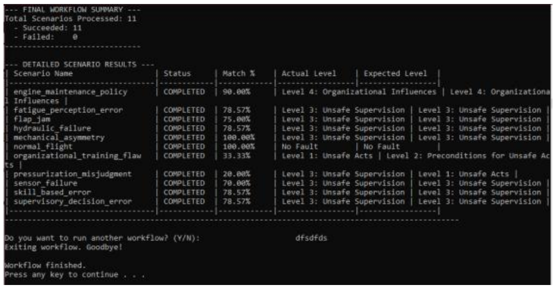
This study adopts three standard classification metrics—Precision (P)**,** Recall (R)**,** and F1-score (F)—as summarized in Table 1. Precision is the proportion of correct positive detections among all positives flagged by the model, reflecting its ability to limit false positives (FP) in accident coding. Recall factor is the proportion of true positives recovered among all actual positives, capturing the model's ability to avoid False negatives (FN) of human-factor findings. F1 is the harmonic mean of Precision and Recall, balancing the dual objectives of "no false alarms" and "no misses." In the aviation-accident setting—where HFACS labels are often imbalanced and some categories are rare—F1 provides a robust single-number summary of performance.

Table 1: The Formula

Metrics	Formular
Precision (P)	$\frac{TP}{TP + FP}$
Recall (R)	$\frac{TP}{TP + FN}$
F1-score (F1)	$\frac{2 * Precision * Recall}{Precision + Recall}$

4. Results and Discussion

The Figure below summarizes the AI system's HFACS classification results for 11 simulated general aviation accident scenarios.



```
--- FINAL WORKFLOW SUMMARY ---
Total Scenarios Processed: 11
  - Succeeded: 11
  - Failed:    0
------------------------------

--- DETAILED SCENARIO RESULTS ---
| Scenario Name                | Status     | Match %    | Actual Level     | Expected Level  |
|------------------------------|------------|------------|------------------|-----------------|
| engine_maintenance_policy    | COMPLETED  | 90.00%     | Level 4: Organizational Influences | Level 4: Organizationa
l Influences |
| fatigue_perception_error     | COMPLETED  | 78.57%     | Level 3: Unsafe Supervision | Level 3: Unsafe Supervision |
| flap_jam                     | COMPLETED  | 75.00%     | Level 3: Unsafe Supervision | Level 3: Unsafe Supervision |
| hydraulic_failure            | COMPLETED  | 78.57%     | Level 3: Unsafe Supervision | Level 3: Unsafe Supervision |
| mechanical_asymmetry         | COMPLETED  | 100.00%    | Level 3: Unsafe Supervision | Level 3: Unsafe Supervision |
| normal_flight                | COMPLETED  | 100.00%    | No Fault         | No Fault        |
| organizational_training_flaw | COMPLETED  | 33.33%     | Level 1: Unsafe Acts | Level 2: Preconditions for Unsafe Ac
ts |
| pressurization_misjudgment   | COMPLETED  | 20.00%     | Level 3: Unsafe Supervision | Level 1: Unsafe Acts |
| sensor_failure               | COMPLETED  | 70.00%     | Level 3: Unsafe Supervision | Level 3: Unsafe Supervision |
| skill_based_error            | COMPLETED  | 78.57%     | Level 3: Unsafe Supervision | Level 3: Unsafe Supervision |
| supervisory_decision_error   | COMPLETED  | 78.57%     | Level 3: Unsafe Supervision | Level 3: Unsafe Supervision |
|------------------------------|------------|------------|------------------|-----------------|
------------------------------------------------------------------------------------------

Do you want to run another workflow? (Y/N):                    dfsdfds
Exiting workflow. Goodbye!

Workflow finished.
Press any key to continue . . .
```

Figure 4.1 The AI system's HFACS classification of aviation accidents

The results demonstrate that the AI system can successfully map the major simulated incidents to the correct HFACS tier, particularly for Unsafe Supervision (Level 3) and Organizational Influences (Level 4) factors. However, the misclassifications highlight challenges in differentiating Preconditions for Unsafe Acts (Level 2) from both higher and lower levels. The model's bias toward Level 3 classifications suggests that additional refinement is needed to improve discrimination between direct pilot errors, underlying preconditions, and supervisory vs. organizational factors in complex accident scenarios.

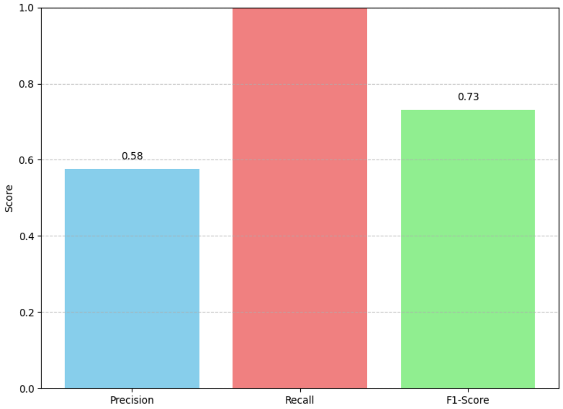


Figure 4.2 Overall HFACS Analysis Performance

Based on the data of 200 simulation runs as shown in Figure 4.2, the HFACS-coding system achieved Precision = 0.58**,** Recall = 1.00**,** and F1 = 0.73 (micro-average). This recall-saturated setting (FN around 0) ensures comprehensive detection of human-factor findings but incurs a non-trivial false-alarm rate that depresses precision. In its current form, the model is suitable for near-real-time triage—minimizing missed safety-critical signals—yet outputs should not be accepted without review by an investigator. Deployment should prioritize raising precision (tighter decision thresholds and rubric criteria, corroboration with telemetry and maintenance records) while maintaining recall;

doing so should increase F1 and improve operational reliability. Overall, the configuration demonstrates strong coverage for HFACS analysis, but precision must be further optimized before use in decision-grade workflows.

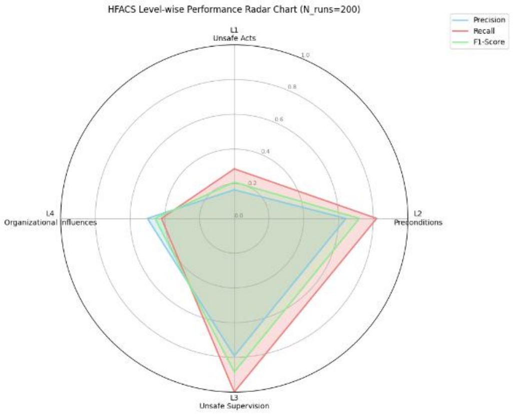


Figure 4.3 HFACS Level-wise Performance Radar Chart

Figure 4.4 shows that system performance is incident-dependent**.** Across 200 runs per scenario, the model operates at a recall-oriented setting**:** 10/11 scenarios achieve R = 1.00**,** while precision varies widely**,** so F1 is largely constrained by precision**.** Taken together, the variability in F1 is driven primarily by precision, because recall is already near its optimum whenever the system is invoked.

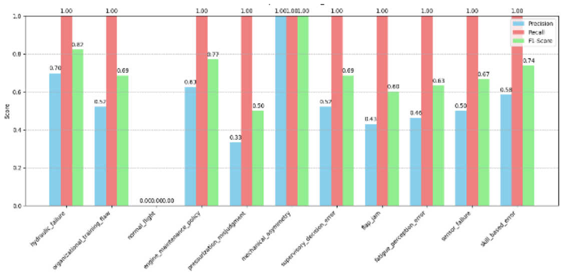


Figure 4.4 Performance of different scenarios

From the chart, the results separate cleanly into two groups. The first group—technical faults/anomalies—shows consistently higher confidence (and F1) than the second group, which comprises manufacturer-related/organizational issues and decision-making errors.

The error profile is concentrated in Level 1–2 tags; no Level 3 tag appears in the top-10, and only one Level 4 tag is present. Color segmentation indicates FP dominates many L1/L2 items, whereas FN clusters on cognitively demanding tags.

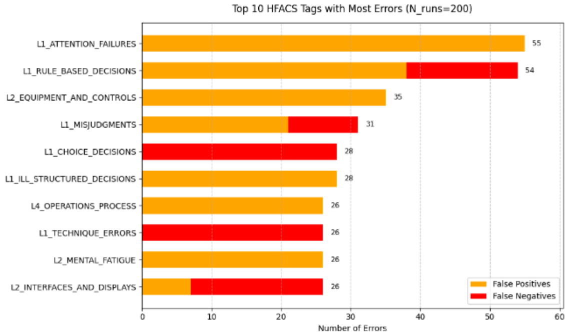


Figure 4.5 Top 10 HFACS Tags with Most Errors

The two main contributors, L1_ATTENTION_FAILURES (55) and

L1_RULE_BASED_DECISIONS (54), are primarily FP, though the latter also exhibits noteworthy FN. FP is also skewed by a second tier, L2_EQUIPMENT_AND_CONTROLS (35), L1_ILL_STRUCTURED_DECISIONS (28), L4_OPERATIONS_PROCESS (26), and L2_MENTAL_FATIGUE (26), which suggests an excessive sensitivity to behavioral/equipment cues. However, for L1_MISJUDGMENTS (31), L1_CHOICE_DECISIONS (28), L1_TECHNIQUE_ERRORS (26), and L2_INTERFACES_AND_DISPLAYS (26), which call for more complex cognitive/interface reasoning, FN is prominent. Overall, the pattern identifies two areas that need improvement: improving the recognition of cognitive/interface indicators to lower FN and tightening the evidence thresholds for L1 behavioral tags to curb FP.

5. Conclusion

This study introduces an automated HFACS coding pipeline that combines a multi-agent architecture with an arbiter and a standardized rubric to deliver low-latency, near–real-time analysis of accident narratives. On a high-fidelity simulator spanning 11 scenarios over 200 runs, the system attains F1 = 0.82 overall, with strong level-wise performance at Level 3 (Unsafe Supervision, F1 = 0.85) and Level 2 (Preconditions, F1 = 0.66). These results indicate that debate-style multi-agent reasoning, followed by rubric-based scoring, can mitigate single-model bias and improve consistency and explainability in large-scale human-factors coding. The main limitation is the simulation-to-field gap. Accordingly, future work will integrate retrieval-augmented grounding (RAG) from historical cases to support evidence-based decisions, replace the rule-based anomaly gate with a machine-learning anomaly detector trained on nominal flight data, and incorporate a human-in-the-loop interface for expert review and continual refinement on real-world reports.

6. Reference

[1] H. Said, “2024 Full Year Accident Update.”

[2] Q. Liu, F. Li, K. K. H. Ng, J. Han, and S. Feng, “Accident investigation via LLMs reasoning: HFACS-guided Chain-of-Thoughts enhance general aviation safety,” *Expert Syst Appl*, vol. 269, Apr. 2025, doi: 10.1016/j.eswa.2025.126422.

[3] C. Chandra, Y. Ojima, M. V. Bendarkar, and D. N. Mavris, “Aviation-BERT-NER: Named Entity Recognition for Aviation Safety Reports,” *Aerospace*, vol. 11, no. 11, Nov. 2024, doi: 10.3390/aerospace11110890.

[4] S. R. Andrade and H. S. Walsh, “SafeAeroBERT: Towards a Safety-Informed Aerospace-Specific Language Model A growing body of research has demonstrated the effectiveness of pre-training Bidirectional Encoder Representations from Transformers

(BERT) models [1] on domain-specific text heavy with jargon. BERT models.”

[5] X. Zhang and S. Mahadevan, “Bayesian network modeling of accident investigation reports for aviation safety assessment,” *Reliab Eng Syst Saf*, vol. 209, 2021, doi: 10.1016/j.ress.2020.107371.

[6] E. Ahmadi, S. Muley, and C. Wang, “Automatic Construction Accident Report Analysis Using Large Language Models (LLMs),” *Journal of Intelligent Construction*, vol. 3, no. 1, pp. 1–10, Jun. 2024, doi: 10.26599/jic.2024.9180039.

[7] F. Wang, A. Liu, C. Qu, R. Xiong, and L. Chen, “A Deep-Learning Method for Remaining Useful Life Prediction of Power Machinery via Dual-Attention Mechanism,” *Sensors*, vol. 25, no. 2, Jan. 2025, doi: 10.3390/s25020497.

[8] E. Ahmadi, S. Muley, C. Wang, S. S. Bert Turner, and P. S. Advisor Bert Turner, “Automatic Construction Accident Report Analysis Using Large Language Models.” [Online]. Available: https://ssrn.com/abstract=4735131

[9] S. V. T. Tran *et al.*, “LEVERAGING LARGE LANGUAGE MODELS FOR ENHANCED CONSTRUCTION SAFETY REGULATION EXTRACTION,” *Journal of Information Technology in Construction*, vol. 29, pp. 1026–1038, 2024, doi: 10.36680/j.itcon.2024.045.

[10] H. Zhen, Y. Shi, Y. Huang, J. J. Yang, and N. Liu, “Leveraging Large Language Models with Chain-of-Thought and Prompt Engineering for Traffic Crash Severity Analysis and Inference,” Aug. 2024, [Online]. Available: http://arxiv.org/abs/2408.04652