Article

Not peer-reviewed version

LLM-Assisted Incident Coding for UAS Safety: Reliability-Aware Human-Factor Extraction and Operational Risk Analytics

Youla Yang *

Posted Date: 4 February 2026

doi: 10.20944/preprints202602.0324.v1

Keywords: UAS safety; drone incidents; human factors; operational risk; large language models; incident narratives

Article

LLM-Assisted Incident Coding for UAS Safety: Reliability-Aware Human-Factor Extraction and Operational Risk Analytics

Youla Yang

Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, IN, USA; yangyoul@iu.edu

Simple Summary

Unmanned aircraft systems generate large volumes of safety incident reports written as free-text narratives, which makes systematic safety analysis time-consuming and subjective. This study presents a reliability-aware framework that uses large language models to automatically extract key human, technical, and environmental safety factors from UAS incident reports. By integrating structured extraction, explicit evidence tracing, and reliability auditing, the proposed approach supports efficient and consistent identification of dominant operational risks, enabling more proactive UAS safety management.

Abstract

Safety analysis for Unmanned Aircraft Systems (UAS) relies heavily on incident and occurrence reports that document operational anomalies, environmental conditions, and human-factor contributors in free-text narrative form. While these narratives contain rich safety-relevant information, transforming them into structured and analyzable knowledge remains labor-intensive, inconsistent, and difficult to scale. This paper proposes a reliability-aware framework for large language model (LLM)-assisted incident coding tailored to UAS safety analysis. A UAS-specific safety factor taxonomy encompassing human, system, environmental, and organizational contributors is first developed. Using constrained prompting, LLMs are guided to extract structured safety factors from incident narratives together with explicit supporting evidence spans. To address trust and robustness concerns in safety-critical applications, a multi-level reliability audit is introduced, integrating self-consistency analysis, evidence stability assessment, and agreement with expert annotations. Finally, the extracted safety factors are incorporated into a risk-weighted operational analytics pipeline to identify dominant contributors and emerging safety patterns across different mission contexts. The proposed approach substantially reduces manual coding effort while maintaining strong alignment with expert judgment, demonstrating the potential of reliability-aware LLM analytics to support scalable and proactive UAS safety management.

Keywords: UAS safety; drone incidents; human factors; operational risk; large language models; incident narratives

 doi:10.20944/preprints202602.0324.v1

1. Introduction

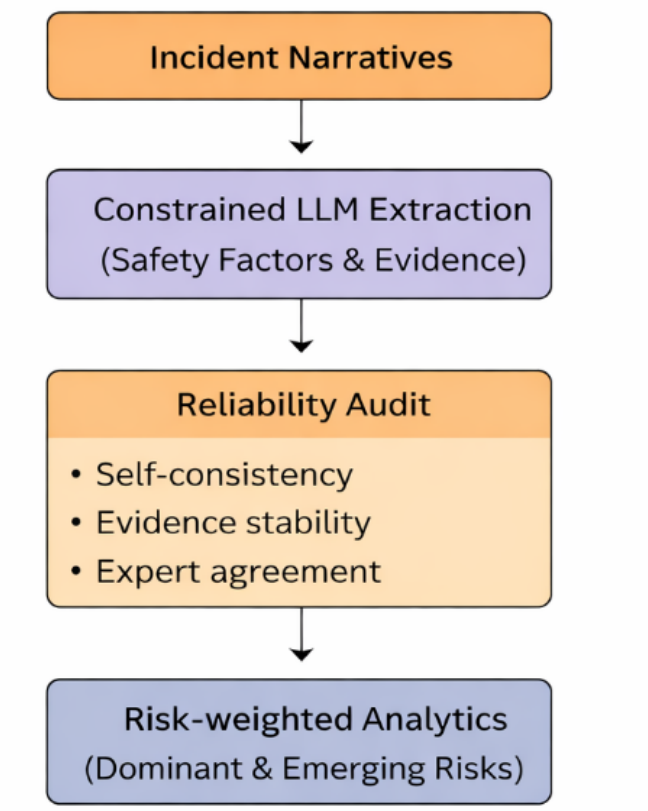

Figure 1. Overview of the proposed reliability-aware LLM-assisted incident coding framework for UAS safety analytics. Incident narratives are processed using constrained extraction to identify safety factors with supporting evidence spans. A multi-level reliability audit evaluates self-consistency, evidence stability, and agreement with expert annotations. Extracted factors are subsequently integrated into a risk-weighted operational analytics pipeline to identify dominant and emerging safety patterns.

Unmanned Aircraft Systems (UAS) are increasingly deployed across a wide range of civilian and commercial applications, including aerial inspection, environmental monitoring, logistics, emergency response, and infrastructure surveillance. As UAS operations continue to scale in complexity and frequency, ensuring operational safety has become a critical concern for regulators, operators, and system designers. Safety management for UAS relies heavily on the systematic analysis of incident and occurrence reports, which document operational anomalies, environmental conditions, system malfunctions, and human-factor contributors observed during flight operations.

In practice, a substantial portion of UAS safety knowledge is embedded in free-text incident narratives. These narratives provide rich contextual detail describing event sequences, contributing factors, and operational circumstances that are often not captured by structured metadata alone. However, the unstructured nature of narrative reports poses significant challenges for large-scale safety analysis. Manual incident coding by domain experts is time-consuming, subjective, and difficult to scale, while inconsistencies in terminology and reporting style further complicate systematic analysis across datasets and operational contexts.

Prior research in aviation and UAS safety has explored a range of natural language processing (NLP) and machine learning techniques to extract recurring themes and contributing factors from safety narratives. Classical approaches, including rule-based systems, statistical text classification, topic modeling, and network-based analytics, have demonstrated potential for identifying dominant patterns in historical data. More recent deep learning methods have improved performance on complex classification tasks but often operate as black boxes, producing factor labels without explicit evidence

 doi:10.20944/preprints202602.0324.v1

grounding or transparency. These limitations restrict their suitability for safety-critical applications, where interpretability, traceability, and accountability are essential.

Recent advances in Large Language Models (LLMs) have renewed interest in leveraging generative AI for safety analysis tasks, including narrative summarization, incident reasoning, and human-factor interpretation. When combined with structured safety frameworks or domain-specific taxonomies, LLMs offer the potential to streamline incident coding and reduce expert workload. However, LLM outputs are inherently stochastic and can be sensitive to prompt formulation and decoding strategy. Without appropriate constraints and validation mechanisms, LLM-based systems may produce unsupported or inconsistent factor assignments, raising concerns about reliability and trustworthiness in high-stakes safety settings.

Addressing these challenges requires not only improved extraction capability but also explicit mechanisms for assessing the reliability of model outputs. In safety-critical domains such as UAS operations, automated analytics should support, rather than replace, expert judgment. Reliability-aware design principles—such as consistency analysis, evidence grounding, and agreement with expert annotations—are therefore crucial for responsible deployment of LLM-assisted safety analytics.

In this work, we propose a reliability-aware framework for LLM-assisted incident coding tailored specifically to UAS safety analysis. The framework integrates a domain-specific safety factor taxonomy, constrained prompting for structured factor extraction, and a multi-level reliability audit that evaluates self-consistency, evidence stability, and alignment with expert judgment. By explicitly coupling extraction with reliability assessment, the proposed approach enables transparent inspection of model outputs and systematic identification of uncertain or ambiguous cases.

Beyond incident-level extraction, the framework incorporates a risk-weighted operational analytics pipeline that aggregates extracted safety factors across mission contexts and operational conditions. This aggregation supports identification of dominant contributors, recurring interaction patterns, and emerging safety risks that may not be apparent from isolated incident reports. Rather than focusing solely on predictive accuracy, the proposed framework emphasizes interpretability, robustness, and actionable insight for proactive UAS safety management.

The main contributions of this paper are threefold. First, we introduce a UAS-specific safety factor taxonomy designed to support consistent and interpretable incident coding from free-text narratives. Second, we develop a reliability-aware LLM-assisted extraction framework that integrates structured prompting with multi-level reliability auditing. Third, we demonstrate how extracted and audited safety factors can be incorporated into a risk-weighted analytics pipeline to support data-driven identification of dominant and emerging UAS safety risks. Together, these contributions highlight the potential of reliability-aware language model analytics to enhance the scalability, transparency, and effectiveness of UAS safety analysis.

2. Related Work

Related research relevant to this study can be broadly grouped into three areas: (i) NLP and machine learning for aviation and UAS safety narratives, (ii) automated incident coding and causal factor identification, and (iii) reliability assessment for LLM-assisted analytics in safety-critical domains.

2.1. NLP for Aviation and UAS Safety Narratives

Aviation safety analysis has long relied on narrative incident reports collected in voluntary reporting systems such as ASRS. Prior studies have applied classical NLP techniques, including text classification, causal factor detection, topic modeling, and network-based analytics, to extract recurring themes and contributing factors from aviation safety narratives [1–3].

Survey studies further highlight persistent challenges in this domain, including noisy text, severe class imbalance, and domain-specific terminology [4]. Similar issues have been observed in UAS safety records, where heterogeneous report formats and incomplete metadata complicate large-scale analysis [5,6]. These findings motivate robust methods capable of extracting comparable safety signals across diverse operational contexts.

 doi:10.20944/preprints202602.0324.v1

2.2. Automated Incident Coding and Causal Factor Identification

Automating the identification of causal or contributing factors from safety reports has been a long-standing research focus. Early work on ASRS-style narratives explored supervised and active learning approaches to reduce expert labeling effort [1]. Subsequent studies formulated incident coding as a multilabel classification problem and demonstrated improved decision support for aviation safety management [2,7].

More recent approaches incorporate deep learning models to address complex label dependencies and long-tail distributions [8,9]. However, most supervised methods output factor labels without explicit evidence grounding, limiting interpretability and systematic auditing—key requirements for safety-critical deployment [4]. Our work addresses this gap by prioritizing evidence-based extraction and reliability auditing.

2.3. LLM-Based Extraction and Aviation Safety Analytics

Recent advances in Large Language Models have prompted growing interest in their application to aviation and UAS safety analysis, including narrative summarization, incident reasoning, and human-factor interpretation [6,10]. When combined with structured frameworks such as HFACS or domain-specific taxonomies, LLMs show potential for streamlining investigation workflows.

Nevertheless, LLM outputs can be sensitive to stochastic decoding and may produce unsupported factor assignments if unconstrained [11]. Prompt engineering and constrained generation have therefore been widely studied to steer model outputs toward valid formats and domain constraints [12,13]. Such constraints are particularly important in safety analytics, where interpretability and accountability are essential.

2.4. Reliability Audits: Self-Consistency and Evidence Grounding

Reliability and trustworthiness remain central challenges for deploying LLMs in safety-critical pipelines. Self-consistency approaches aggregate multiple generations to reduce variance and improve robustness [14,15], while reliability-aware extensions adapt generation strategies based on agreement or confidence signals [16].

Complementary to consistency, evidence grounding—requiring stable textual support for model outputs—facilitates human verification and error detection [17,18]. These ideas align with broader work on interpretable and accountable AI for high-stakes domains [19]. Building on this literature, the present study integrates self-consistency, evidence stability, and expert agreement into a unified reliability audit for LLM-assisted incident coding.

3. Materials and Methods

3.1. Incident Narrative Dataset

A corpus of Unmanned Aircraft Systems (UAS) incident narratives was compiled from publicly available safety reports and operational occurrence records released by regulatory agencies and safety reporting systems. Each incident record consists of a free-text narrative describing the event sequence, contributing circumstances, operational context, and observed outcomes. When available, structured metadata fields were also collected, including mission type, operational environment (e.g., urban, rural, or mixed), platform characteristics, and reported outcome severity.

To ensure data privacy and ethical use, all narratives were anonymized prior to analysis by removing or masking identifying information related to operators, organizations, locations, or specific platforms. Incident reports with empty, extremely short, or incomplete narrative descriptions were excluded, as such records do not provide sufficient contextual information for reliable factor extraction. After filtering, the resulting dataset spans a diverse range of operational contexts, mission profiles, and incident types, providing a representative basis for evaluating automated incident coding methods in UAS safety analysis.

 doi:10.20944/preprints202602.0324.v1

3.2. UAS Safety Factor Taxonomy

To enable consistent, interpretable, and domain-relevant incident coding, a structured UAS safety factor taxonomy was defined. The taxonomy comprises four primary categories: (i) *human factors*, such as decision-making, situational awareness, training adequacy, workload, and human–machine interaction; (ii) *system and avionics factors*, including navigation, communication, sensing, control, and propulsion subsystems; (iii) *environmental conditions*, such as weather, visibility, terrain, and electromagnetic interference; and (iv) *procedural and organizational factors*, including standard operating procedures, supervision, coordination, and organizational practices.

The taxonomy was informed by established aviation safety and human-factors frameworks and refined to reflect UAS-specific operational characteristics and reporting practices. Definitions and category boundaries were iteratively refined through pilot annotation and error analysis to reduce overlap between factors and to improve inter-category separability. This taxonomy serves as both a constrained label space for automated extraction and a common semantic basis for downstream risk aggregation and operational analysis.

3.3. LLM-Assisted Safety Factor Extraction

Incident narratives were processed using a constrained prompting strategy that guides a Large Language Model (LLM) to extract safety factors strictly within the predefined taxonomy. Prompts explicitly specified the allowed factor categories and required the model to return structured outputs consisting of (i) identified safety factors and (ii) corresponding evidence spans quoted verbatim from the original narrative text.

For each incident narrative, multiple independent decoding runs were performed using identical prompts but different random seeds. This design captures the inherent stochasticity of LLM generation and enables analysis of output variability. Rather than relying on a single model response, this multi-run strategy provides a richer basis for assessing extraction robustness and identifying ambiguous or unstable factor assignments.

3.4. Reliability and Consistency Audit

Given the safety-critical nature of UAS incident analysis, a multi-level reliability audit was conducted to systematically evaluate the trustworthiness and robustness of extracted safety factors. Rather than treating LLM outputs as deterministic predictions, the proposed audit framework explicitly accounts for variability arising from stochastic decoding, narrative ambiguity, and interpretive uncertainty. Three complementary criteria were considered to assess extraction reliability from different perspectives.

First, *self-consistency* was evaluated by examining the frequency with which the same safety factor was extracted across multiple independent LLM runs for a given incident narrative. Multiple decoding passes were performed using identical prompts but different random seeds. Safety factors that appeared consistently across runs were regarded as more reliable, while factors with sporadic occurrence were flagged as potentially unstable. This criterion captures the robustness of factor identification under stochastic generation and helps distinguish dominant contributors from marginal or uncertain assignments.

Second, *evidence stability* was assessed by analyzing the consistency of textual grounding across runs. For each extracted safety factor, the corresponding evidence spans quoted from the narrative were compared using token-level overlap metrics. High overlap indicates that the model relies on stable and explicit textual cues within the incident description, whereas low overlap suggests sensitivity to phrasing variations, narrative structure, or contextual ambiguity. Evidence stability thus complements self-consistency by evaluating not only whether a factor is repeatedly extracted, but also whether it is supported by consistent portions of the source text.

Third, *agreement with expert annotations* was evaluated on a manually labeled subset of incident narratives. Expert annotations were produced following the same UAS safety factor taxonomy and

 doi:10.20944/preprints202602.0324.v1

annotation guidelines used to constrain the LLM outputs. Agreement between automated extraction and expert labels was quantified using standard inter-rater agreement metrics. This comparison provides an external reference point for assessing the practical validity of extracted factors and highlights categories where automated extraction aligns well with expert judgment versus those requiring greater interpretive caution.

Taken together, these three criteria form a holistic reliability audit that addresses both internal model robustness and external validity. By integrating self-consistency, evidence stability, and expert agreement, the audit framework enables differentiation between robust factor assignments suitable for downstream risk analysis and uncertain cases that may warrant expert review. This reliability-aware design supports transparent, accountable use of LLM-assisted analytics in UAS safety management rather than fully autonomous decision-making.

As illustrated in Figure 2, the proposed reliability audit evaluates extraction robustness from complementary perspectives of consistency, evidence grounding, and expert agreement.

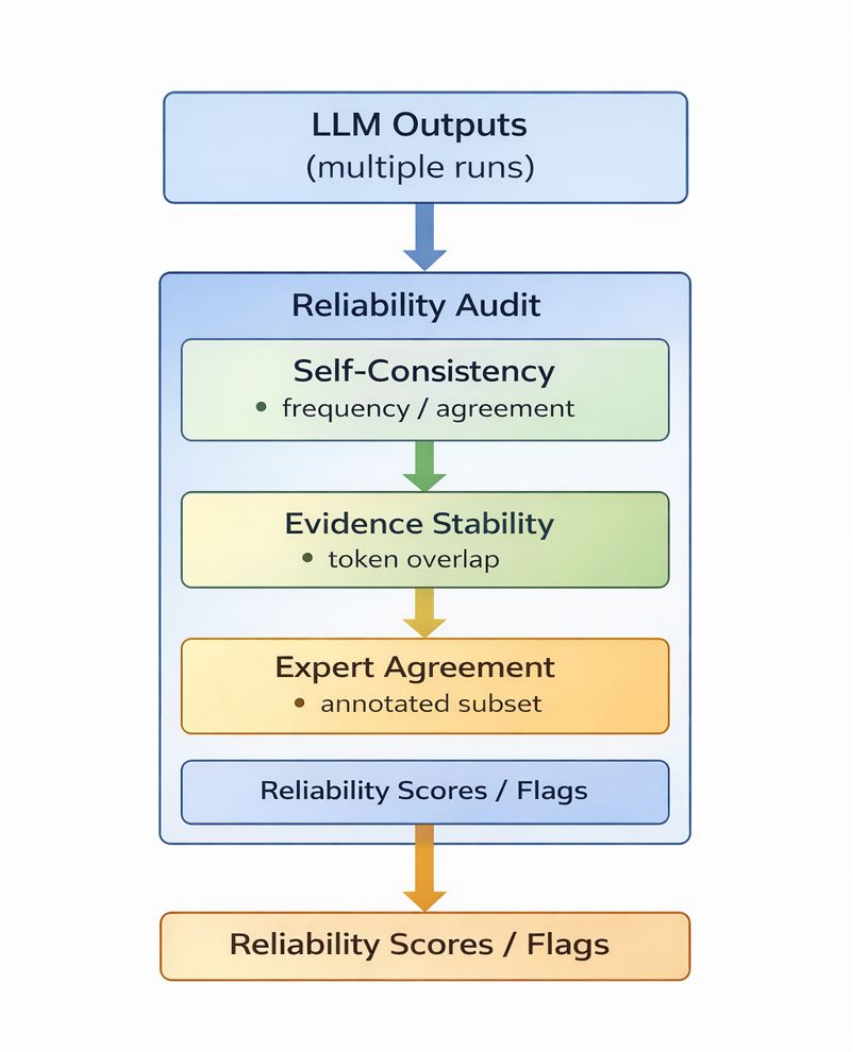

Figure 2. Reliability audit pipeline for LLM-assisted incident coding. Multiple independent LLM generations are evaluated through complementary criteria, including self-consistency across decoding runs, evidence stability based on token-level overlap of extracted spans, and agreement with expert annotations. The audit produces reliability scores or flags that support transparent and accountable downstream safety analysis.

3.5. Risk-Weighted Operational Analytics

Extracted safety factors were integrated into a risk-weighted operational analytics pipeline to support higher-level safety insights. Severity and likelihood proxies were derived from available

 doi:10.20944/preprints202602.0324.v1

incident metadata, including reported outcomes and contextual indicators, and used to compute factor-level risk scores.

Risk scores were aggregated across incidents to identify dominant contributors within specific mission types and operational contexts. Temporal and contextual aggregation further enabled identification of emerging safety patterns and interaction effects between human, system, and environmental factors. This risk-weighted analysis supports proactive UAS safety management by highlighting recurring contributors and prioritizing areas for intervention that may not be evident from individual incident reports alone.

4. Results

4.1. Incident Coding Performance

The proposed framework was evaluated on a held-out test split of the UAS incident narrative dataset under a multi-label classification setting. Two baseline methods were considered for comparison: (i) a lightweight rule-based approach using direct substring matching against frequent safety factor labels, and (ii) a statistical text classification baseline based on TF–IDF representations with a One-vs-Rest (OvR) linear classifier.

The rule-based baseline achieved negligible performance across all evaluation metrics, indicating that direct string matching is ineffective for extracting structured safety factors from unstructured narrative text. This result highlights the inherent limitations of heuristic approaches in complex safety reporting scenarios, where contributing factors are often described implicitly, distributed across multiple sentences, or expressed using domain-specific terminology.

In contrast, the TF–IDF baseline demonstrated substantially stronger performance. As summarized in Table 1, the TF–IDF + OvR model achieved a micro-F_1 score of approximately 0.57, with a micro-precision of 0.65 and a micro-recall of 0.51. These results indicate that classical text representations are capable of capturing meaningful correlations between narrative content and safety factor labels, despite the presence of severe label imbalance and heterogeneous reporting styles.

Notably, the gap between micro- and macro-averaged metrics reflects the long-tail distribution of safety factors. A small number of dominant contributors account for a large proportion of labeled instances, while many factors occur infrequently. As a result, frequent safety factors are detected with reasonable reliability, whereas rare or highly context-specific contributors remain challenging for purely statistical models. This observation motivates the need for structured extraction and reliability-aware analysis beyond conventional classification performance.

Table 1. Multi-label incident coding performance on the UAS narrative dataset.

Method	Micro-F_1	Macro-F_1	Micro-Precision	Micro-Recall
Rule-based baseline (substring)	0.00	0.00	0.00	0.00
TF–IDF + OvR (SGD)	0.57	0.08	0.65	0.51

4.2. Reliability and Consistency Analysis

Beyond aggregate predictive performance, the reliability of extracted safety factors is critical for safety-critical applications. Analysis across multiple decoding runs reveals systematic differences in extraction stability across factor types. High-support safety factors, particularly those related to human decision-making and system malfunctions, exhibit strong self-consistency across runs. In contrast, lower-frequency or context-dependent factors show greater variability.

Self-consistency analysis indicates that factors consistently extracted across multiple runs are typically associated with explicit narrative cues, such as direct statements of operator error, equipment malfunction, or adverse environmental conditions. Conversely, factors with low consistency often arise in narratives describing complex event chains or indirect contributing circumstances, where causal attribution is inherently ambiguous.

 doi:10.20944/preprints202602.0324.v1

Evidence stability analysis further shows that, when the same safety factor is extracted across runs, the associated evidence spans exhibit substantial token-level overlap. This finding suggests that the model relies on stable textual cues within the narrative rather than generating unsupported or spurious explanations. Reduced evidence overlap is primarily observed in narratives containing multiple plausible interpretations or overlapping factor categories.

Agreement with expert annotations is strongest for dominant human and system factors, where narrative descriptions are relatively explicit. Discrepancies are more common for higher-level organizational or procedural contributors, which are often implicit and require interpretive judgment. These results demonstrate the value of reliability auditing for identifying uncertain cases and prioritizing expert review.

4.3. Operational Risk Analytics

Integrating extracted safety factors with severity and likelihood proxies enables risk-weighted operational analytics across mission contexts. Aggregated results show that human-factor contributors dominate across a wide range of operational scenarios, consistent with prior findings in aviation and UAS safety research. System and environmental factors frequently act as amplifiers, particularly under adverse operational or environmental conditions.

Risk-weighted aggregation reveals distinct patterns across mission types. Routine operations tend to exhibit lower overall risk scores with isolated contributing factors, whereas complex or non-standard missions show higher concentrations of interacting human, system, and environmental contributors. Temporal aggregation further indicates that certain combinations of factors recur across incidents, suggesting structural vulnerabilities rather than isolated failures.

Overall, these results demonstrate that structured safety factor extraction, combined with reliability-aware analysis, supports data-driven identification of dominant and emerging UAS safety risks. By enabling systematic comparison across operational contexts, the proposed framework provides actionable insights for proactive safety management.

5. Discussion

The results underscore the importance of reliability-aware design when applying large language models to safety-critical analytics. While conventional statistical baselines provide useful reference points, they primarily optimize predictive performance and offer limited transparency. In contrast, the proposed framework emphasizes evidence grounding and consistency analysis, enabling human experts to inspect, validate, and contextualize model outputs.

The observed discrepancies between automated extraction and expert annotations highlight the continued necessity of human oversight. Rather than replacing expert judgment, the framework functions as a decision-support tool that reduces manual coding effort while preserving accountability. Reliability signals such as self-consistency and evidence stability provide practical mechanisms for identifying cases that warrant closer expert review.

Several limitations should be acknowledged. First, risk weighting relies on proxy indicators derived from available metadata, which may not fully capture operational severity or likelihood. Second, the safety factor taxonomy is fixed and may not capture emerging or context-specific contributors. Finally, while reliability auditing improves robustness, it does not eliminate ambiguity inherent in narrative-based safety reporting.

Future work will explore adaptive taxonomy refinement, integration of additional operational signals, and longitudinal analysis of safety trends. Extending the framework to support real-time monitoring and cross-domain generalization also represents a promising direction.

 doi:10.20944/preprints202602.0324.v1

6. Future Research Directions

While the proposed framework demonstrates the feasibility and utility of reliability-aware LLM-assisted incident coding for UAS safety analysis, several promising directions remain for future research.

6.1. Adaptive Taxonomy Refinement

The current framework relies on a fixed UAS safety factor taxonomy designed to ensure interpretability and consistency. However, operational environments, platforms, and mission profiles continue to evolve. Future work may explore adaptive taxonomy refinement mechanisms that allow new safety factors or subcategories to emerge over time. Such extensions could leverage clustering, concept discovery, or expert-in-the-loop feedback to incrementally expand the taxonomy while preserving semantic coherence and auditability.

6.2. Integration of Multimodal and Contextual Signals

Incident narratives represent only one source of safety-relevant information. Future research may integrate additional operational signals, including flight telemetry, weather data, airspace constraints, and sensor logs, to provide richer contextual grounding for safety factor extraction. Multimodal integration has the potential to reduce ambiguity in narrative interpretation and improve both extraction reliability and downstream risk assessment.

6.3. Real-Time and Near-Real-Time Safety Monitoring

The present study focuses on retrospective analysis of historical incident reports. An important future direction is the extension of the framework to real-time or near-real-time safety monitoring. By coupling streaming incident reports or operational alerts with lightweight LLM inference and reliability auditing, the framework could support early warning systems that flag emerging risks before they escalate into serious incidents.

6.4. Human-in-the-Loop Reliability Calibration

Although reliability auditing improves robustness, expert judgment remains essential in safety-critical contexts. Future work may investigate tighter human-in-the-loop integration, where expert feedback dynamically calibrates reliability thresholds, adjusts prompting strategies, or informs selective re-analysis of uncertain cases. Such interaction would further align automated analytics with operational decision-making practices.

6.5. Cross-Domain Generalization

Finally, while this work focuses on UAS safety, the underlying principles of constrained extraction, evidence grounding, and reliability auditing are broadly applicable. Future research may examine cross-domain generalization to other safety-critical domains, such as manned aviation, maritime operations, autonomous driving, and industrial process safety, thereby evaluating the transferability and scalability of the proposed framework.

7. Potential Applications

The proposed reliability-aware LLM-assisted incident coding framework enables a range of practical applications across UAS safety management, regulation, and system design.

An overview of representative deployment and application scenarios is summarized in Figure 3.

 doi:10.20944/preprints202602.0324.v1

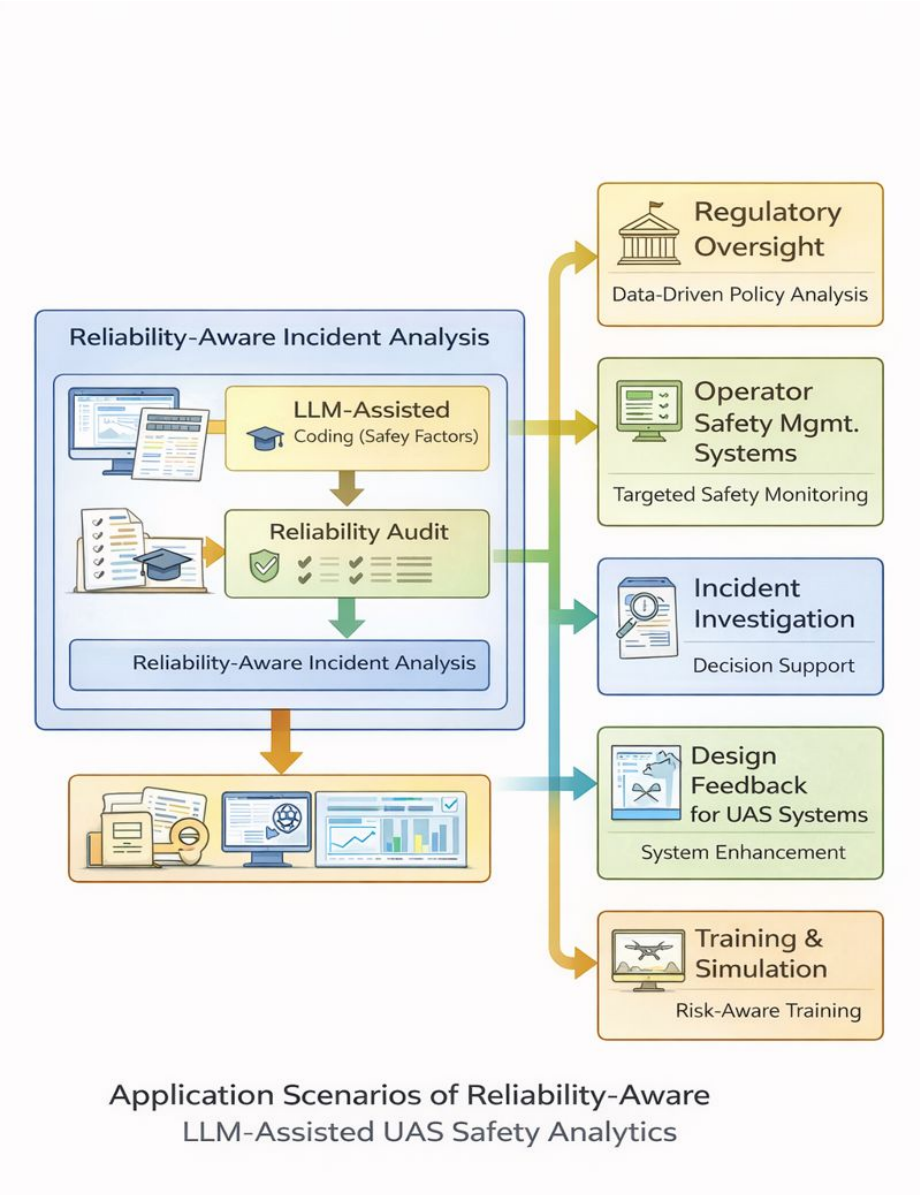

Figure 3. Application scenarios of reliability-aware LLM-assisted UAS safety analytics. The proposed framework performs LLM-assisted incident coding with explicit evidence support, followed by reliability auditing to provide trustworthy safety-factor extraction. The resulting reliability-aware analytics can be integrated into multiple downstream use cases, including regulatory oversight, operator safety management systems, incident investigation support, design feedback for UAS systems, and risk-aware training and simulation.

7.1. Regulatory Safety Analysis and Oversight

Regulatory agencies and aviation authorities routinely analyze large volumes of incident and occurrence reports to identify systemic risks. The proposed framework can support scalable, transparent preprocessing of narrative reports by automatically extracting safety factors with explicit evidence and reliability indicators. This capability facilitates prioritization of high-risk issues, supports data-driven policy development, and enhances traceability in regulatory decision-making.

7.2. Operator Safety Management Systems

UAS operators can integrate the framework into Safety Management Systems (SMS) to support routine safety monitoring and internal audits. By aggregating reliability-audited safety factors across missions, operators can identify recurring operational weaknesses, training deficiencies, or procedural gaps. The evidence-grounded outputs further enable efficient review by safety officers and support targeted corrective actions.

 doi:10.20944/preprints202602.0324.v1

7.3. Incident Investigation and Expert Support

During post-incident investigation, analysts often face time pressure and large volumes of unstructured narrative data. The proposed framework can serve as an intelligent decision support tool by highlighting candidate contributing factors, providing supporting evidence spans, and flagging uncertain attributions. This functionality accelerates investigation workflows while preserving expert authority over final determinations.

7.4. Design Feedback for UAS Systems

Aggregated safety factor analytics can provide valuable feedback to UAS designers and manufacturers. Recurring system-related contributors identified across incidents may inform design improvements, redundancy strategies, or human–machine interface refinements. Reliability-aware extraction ensures that such insights are grounded in consistent narrative evidence rather than isolated or spurious reports.

7.5. Training and Risk-Aware Simulation

Finally, extracted and categorized safety factors can be used to support risk-aware training and simulation environments. Scenario generation informed by real-world incident patterns enables more realistic training for operators and safety personnel. By emphasizing dominant and emerging risks, the framework supports proactive skill development and safety culture reinforcement.

8. Conclusions

This paper presents a reliability-aware framework for LLM-assisted incident coding and operational risk analytics in UAS safety. By combining domain-specific taxonomy design, constrained extraction with explicit evidence grounding, and multi-level reliability auditing, the proposed approach enables scalable and interpretable analysis of incident narratives.

Experimental results demonstrate that the framework substantially reduces manual coding effort while maintaining strong alignment with expert judgment. Reliability-aware analytics further support the identification of dominant and emerging safety risks across operational contexts. These findings highlight the potential of structured and accountable language model analytics to support next-generation UAS safety management systems.

Future research will investigate tighter integration with operational workflows, adaptive learning mechanisms, and deployment in real-time or near-real-time safety monitoring environments.

References

1. Persing, I.; Ng, V. Improving Cause Detection Systems with Active Learning. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2009. Applied to ASRS-style aviation safety reports.
2. Dong, T.; et al. Identifying Incident Causal Factors to Improve Aviation Safety. *Complexity* **2021**, *2021*, 1–12. https://doi.org/10.1155/2021/5540046.
3. Xing, Y.; et al. Discovering Latent Themes in Aviation Safety Reports Using Text Mining and Network Analytics. *Results in Engineering* **2024**, *21*, 100742.
4. Nanyonga, A.; Joiner, K.; Turhan, U.; Wild, G. Applications of Natural Language Processing in Aviation Safety: A Review and Qualitative Analysis. *Aerospace* **2025**, *12*, 34.
5. Cunningham, K.; et al. Exploratory Analysis of Unmanned Aircraft Sightings Using Text Mining. *Transportation Research Record* **2021**, *2675*, 121–132. https://doi.org/10.1177/0361198120987230.
6. Yan, Y.; et al. UAV Accident Forensics via HFACS-Based Large Language Model Reasoning. *Drones* **2025**, *9*, 704. https://doi.org/10.3390/drones9100704.
7. Dong, T.; et al. Multilabel Classification for Aviation Incident Causal Factor Identification. *Safety Science* **2021**, *139*, 105260.
8. Xu, Z.; et al. Deep Learning for Aviation Safety Incident Analysis. *IEEE Access* **2020**, *8*, 215265–215276.
9. Wang, H.; et al. A Deep Learning Approach for Aviation Incident Report Classification. *Safety Science* **2022**, *149*, 105654.

 doi:10.20944/preprints202602.0324.v1

10. Puranik, A.; et al. Large Language Models for Aviation Safety: Enhancing Incident Analysis through ASRS Report Summarization. *Aerospace* **2025**, *12*, 89.
11. Wei, J.; et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* **2022**, *35*, 24824–24837.
12. Chen, B.; et al. Unleashing the Potential of Prompt Engineering for Large Language Models: A Comprehensive Review. *Software Impacts* **2025**, *23*, 100603.
13. Liu, P.; et al. A Survey of Prompting Methods for Large Language Models. *ACM Computing Surveys* **2023**, *55*, 1–35.
14. Wang, X.; et al. Self-Consistency Improves Chain-of-Thought Reasoning in Language Models. *International Conference on Learning Representations* **2023**.
15. Xie, Y.; et al. Self-Consistency for Robust Large Language Model Reasoning. *arXiv preprint* **2023**, [2302.11427].
16. Zhang, L.; et al. Reliability-Aware Adaptive Self-Consistency for Large Language Models. *arXiv preprint* **2024**, [2403.01829].
17. Jacovi, A.; Goldberg, Y. Towards Faithfully Interpretable NLP Systems. *Proceedings of the ACL* **2020**, pp. 419–433.
18. Wiegreffe, S.; et al. Measuring Faithfulness in Abstractive Summarization. *Proceedings of NAACL* **2021**, pp. 1908–1920.
19. Doshi-Velez, F.; Kim, B. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint* **2017**, [1702.08608].