# LLM-DRIVEN LIFE-CYCLE ACCIDENT INVESTIGATION IN GENERAL AVIATION: A MULTI-STAGE FRAMEWORK FOR CAUSATION GRAPH CONSTRUCTION

**LIU QINGLI**

**PhD**

**The Hong Kong Polytechnic University**

**2025**

The Hong Kong Polytechnic University

Department of Aeronautical and Aviation Engineering

# LLM-Driven Life-Cycle Accident Investigation in General Aviation: A Multi-Stage Framework for Causation Graph Construction

LIU Qingli

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Aug 2025

## CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

LIU Qingli (Name of student)

# Abstract

General aviation (GA) plays a vital role in the low-altitude economy but continues to suffer from disproportionately high accident rates. Accident investigation, the foundation of aviation safety, remains expert-dependent and slow, with its core challenge lying in reconstructing the causal sequence of accidents rather than identifying a single root cause. Recent advances in large language models (LLMs) offer new opportunities to automate this process through natural language understanding and multi-step reasoning, yet their application is limited by hallucination, unstable reasoning, and poor adaptability across investigation stages.

This thesis proposes an LLM-driven life-cycle framework for GA accident investigation, aligned with the preliminary, detailed, and post-investigation stages of real-world practice. To reduce hallucinations and enhance reasoning performance, the framework integrates three complementary methods: (1) domain-informed Chain-of-Thought (HFACS-CoT) prompting for preliminary investigations, guiding LLMs to reliably identify and classify human errors from pilot narratives; (2) a Dynamic Aviation Risk Field–guided workflow (DARF-ACR) for detailed investigations, enabling structured reasoning to reconstruct multi-event causal chains; and (3) random-parameter bivariate probit modeling with heterogeneity in means, coupled with a structured knowledge base and LLM semantic retrieval, for post-

investigation analyses, augmenting causal graphs with statistically validated exogenous risk factors.

The framework is validated on a self-constructed dataset derived from the U.S. National Transportation Safety Board (NTSB). Results demonstrate that: (1) HFACS-CoT prompting significantly improves the accuracy of human error inference, achieving higher F1 scores than baseline prompts; (2) the DARF-ACR workflow produces more complete and temporally consistent causal graphs, outperforming existing methods in both node identification and edge reconstruction; and (3) the integration of statistically validated risk factors through econometric modeling and LLM retrieval, further verified through a case study, enriches causal graphs with contextually grounded explanations, extending the analysis from how accidents occurred to why they were more likely under specific conditions.

Overall, this thesis advances the theory and practice of GA accident investigation by presenting an innovative LLM-based framework for life-cycle causation analysis. It shifts the focus from static root-cause identification to dynamic causal graph construction, offers a paradigm that can accelerate investigations and improve decision support, and demonstrates a transferable methodology for applying trustworthy and explainable AI in other safety-critical domains.

# Publications arising from the thesis

*Journal articles (in reverse chronological order)*


[1] Liu, Q., Song, P., Li, F. "Exploring the dynamic determinants of general aviation accidents across flight phases and time: A random parameter bivariate probit approach with heterogeneity in means". In: *Analytic Methods in Accident Research* 47 (2025), p. 100386. ISSN: 2213-6657. DOI: https://doi.org/10.1016/j.amar.2025.100386. URL: https://www.sciencedirect.com/science/article/pii/S221366572500017X.

[2] Liu, Q., Li, F., Ng, K. K., Han, J., Feng, S. "Accident investigation via LLMs reasoning: HFACS-guided Chain-of-Thoughts enhance general aviation safety". In: *Expert Systems with Applications* 269 (2025), p. 126422. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2025.126422. URL: https://www.sciencedirect.com/science/article/pii/S0957417425000442.

[3] Liu, Q., Li, F., Ng, K. K. "Unveiling the determinants of injury severities across age groups and time: A deep dive into the unobserved heterogeneity among pedestrian crashes". In: *Analytic Methods in Accident Research* 43

(2024), p. 100336. ISSN: 2213-6657. DOI: https://doi.org/10.1016/j.amar.2024.100336. URL: https://www.sciencedirect.com/science/article/pii/S2213665724000204.

*Manuscript under review*

[1] Liu, Q., Li, F. "From Accident Evidence to Causal Graphs: Field-Theory–Guided General Aviation Accident Investigation with LLMs". In: *Reliability Engineering and System Safety (under review)* (2025).

*Conference paper (in reverse chronological order)*

[1] Liu, Q., Yan, Y., Li, F., Feng, S. "Show the Way: Accelerating General Aviation Accident Investigations through LLMs and HFACS". In: *Advances in Human Factors of Transportation* 148.148 (2024).

[2] Liu, Q., Han, S., Li, F., Lee, C.-H. "Group Dynamics and Air Traffic Controllers' Safety Behaviors: Self-Efficacy as a Mediator". In: *Leveraging Transdisciplinary Engineering in a Changing and Connected World*. Advances in Transdisciplinary Engineering. 2023. ISBN: 9781643684406. DOI: 10.3233/atde230613.

# Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Fan Li. Her patience, wisdom, and keen academic insight have guided me through every stage of this research. She always identified the core of my confusion with ease and offered insightful suggestions that greatly inspired my thinking. Beyond her academic guidance, her approachable and calm personality made my doctoral journey both rewarding and enjoyable.

I am also sincerely grateful to my co-supervisor, Professor Gangyan Xu, for his continuous encouragement and thoughtful advice, which helped me regain direction whenever I felt uncertain. My heartfelt thanks also go to Professor Kam K.H. Ng, whose generous support in funding, manuscript revision, and resource sharing made this work possible. I would further like to thank the kind strangers who helped me while I was learning the random parameter logit model. Though we have never met, their shared references and patient explanations were invaluable to me. I am equally indebted to Professor Fred Mannering, Advisory Editor of *Analytic Methods in Accident Research*. Despite the time difference between us, he consistently responded to my emails with remarkable promptness and detail. His encouragement gave me great confidence to persist in my research journey.

I am also grateful to all the members of the SHINE Group. Whether during

hiking trips, badminton games, daily meals, or research discussions, their companionship, support, and generosity made my PhD life both meaningful and joyful. With their presence, my three years of doctoral study were filled with warmth and laughter.

Finally, I wish to thank my family—my parents, my sister, and my grandmother. They have been both my soft spot and my armor. Their unconditional love has been my responsibility as well as my motivation, giving me the strength to face every challenge. Last but not least, I am grateful to all the people who have shown kindness to me along the way. This thesis is dedicated to them.

# Contents

# List of Figures

# List of Tables

# Nomenclature

| | |
|---|---|
| GA | General Aviation |
| NTSB | National Transportation Safety Board |
| LLM | Large Language Model |
| RAG | Retrieval-Augmented Generation |
| CoT | Chain of Thought |
| HFACS | Human Factors Analysis and Classification System |
| GAHFACS | General Aviation Human Factors and Accident Causes |
| DARF | Dynamic Aviation Risk Field |
| ACR | Accident Causation Reconstruction |
| DARF-ACR | Dynamic Aviation Risk Field–guided Accident Causation Reconstruction |
| CER | Causal edge recall |
| NLP | Natural Language Processing |
| HCI | Human–Computer Interaction |
| DAG | Directed Acyclic Graph |
| KG | Knowledge Graph |
| JSON | JavaScript Object Notation |
| CNN | Convolutional Neural Network |
| SVM | Support Vector Machine |

| | |
|---|---|
| PoT | Program of Thought |
| ToT | Tree of Thought |
| GoT | Graph of Thought |

# Chapter 1

# Introduction

This introductory chapter comprises five sections. Section 1.1 provides the research background on general aviation (GA) accident investigation, highlighting the challenges of timely and accurate causal analysis and introducing the potential of large language models (LLMs). Section 1.2 defines the overall objective of the thesis and specifies three sub-objectives that align with the stages of real-world investigations. Section 1.3 outlines the scope of the research, clarifying its focus on LLM-driven causal graph reconstruction and enhancement while excluding data collection and organizational-level analyses. Section 1.4 discusses the academic, practical, and methodological significance of the work. Finally, Section 1.5 presents the organization of the thesis.

## 1.1 Background

General aviation (GA, Part 91) plays a critical role in disaster relief, medical evacuation, environmental monitoring, and cargo transportation, especially in the cur-

Figure 1.1: GA accident investigation process (adapted from NTSB guidelines)

rent era of the low-altitude economy. Despite its extensive benefits, GA presents considerable safety risks, with a higher accident rate compared to commercial aviation. National Transportation Safety Board (NTSB) shows that hundreds of fatalities and thousands of injuries occur each year in GA accidents in the United States, accounting for the overwhelming majority of civil aviation accidents [1–3].

Accident investigation is the foundation for identifying vulnerabilities and preventing recurrence [4]. In GA, however, timely and accurate analysis remains highly challenging. According to the National Transportation Safety Board (NTSB), determining the probable cause of a GA accident typically requires 12 to 24 months. As shown in Fig. 1.1, this process involves on-site examination, evidence collection, causal reconstruction, and multi-stage report validation, which can be broadly categorized into initial, detailed, and post-investigation phases [5]. In practice, GA accidents are rarely the result of a single factor; rather, they emerge from the interaction and escalation of multiple contributors across different stages of flight. Thus, the core task of GA accident investigation is not simply to identify a discrete "root cause," but to reconstruct the causal process that explains *how* the accident unfolded. Achieving this requires synthesizing diverse forms of evidence—such as flight logs, maintenance records, and witness statements—to identify causal nodes and their interrelationships, a task that is both knowledge-intensive and reasoning-intensive, and consequently expert-dependent and time-consuming.

Recent advances in large language models (LLMs) offer new opportunities to automate accident causal graph reconstruction through their powerful natural lan-

guage understanding and multi-step reasoning [6], potentially accelerating the entire GA accident investigation process. Yet a key challenge remains: hallucination. Hallucination refers to the generation of fluent but factually incorrect or logically inconsistent content, as LLMs are trained to predict the most probable next token rather than to guarantee factual accuracy or causal validity [7–9]. This problem is especially pronounced in specialized domains where data are limited and reasoning tasks are complex [10]. GA accident investigation exemplifies these difficulties: accident evidence is highly unstructured and heterogeneous—narratives often mix factual descriptions with subjective impressions and emotional language, while also containing domain-specific terminology and abbreviations uncommon in general corpora. Moreover, GA accidents typically involve intertwined human, mechanical, and environmental factors that evolve into multi-event causal chains. These characteristics increase the likelihood that LLMs fabricate evidence, misrepresent factor importance, or generate spurious causal links, undermining their reliability in constructing causal graphs.

Currently, three major strategies are widely adopted to mitigate hallucination in LLMs: prompt engineering, workflow design, and retrieval-augmented generation (RAG). However, these methods are mostly validated on general-purpose tasks and lack adaptation to the context of GA accident investigation. Specifically, applying these strategies to GA accident investigation still faces the following unique challenges:

- **Stage-specific investigation process.** GA accident investigation proceeds through multiple stages, from initial fact-finding to detailed causal reconstruction and post-accident risk assessment. Each stage is characterized by

distinct evidence sources, reasoning requirements, and reporting goals. To apply LLMs effectively across the entire investigation pipeline, hallucination-mitigation strategies must be adapted to these stage-specific demands. A uniform approach risks failing to capture the unique needs of each stage, reducing the reliability of investigation outcomes.

- **Explicit and exogenous factors.** GA accidents arise not only from explicit causal factors but also from exogenous risk factors such as pilot demographics, aircraft type, and operational mode. Although these factors do not directly trigger accidents, statistical evidence shows they significantly increase accident likelihood. LLMs, however, lack the statistical reasoning required to capture such exogenous risks, and simple retrieval-augmented generation is inadequate for this purpose.

- **Dynamic causal chains.** GA accidents are not single-point or single-cause events but evolve through dynamic, multi-event causal chains. LLMs are prone to hallucinations or omissions in long-chain reasoning, making it difficult to capture temporal progressions and cross-event dependencies. Static root-cause models and existing hallucination-mitigation strategies cannot address such dynamics, revealing their inadequacy in handling the complexity of GA accident causation.

- **Reliability and explainability.** GA accident investigation is a high-stakes domain where reliability and transparency are essential. Basic hallucination-mitigation strategies often result in black-box reasoning without domain-grounded interpretability, which undermines trust in LLM-generated causal graphs by investigators and policymakers.

To address these challenges, this thesis proposes an LLM-driven life-cycle framework for GA accident causal reconstruction. The framework aligns LLM reasoning with the three stages of real-world investigations and integrates accident investigation models with prompt engineering, workflow design, econometric modeling, and RAG. Following a progressive pipeline, it advances from causal node identification to causal graph reconstruction and finally to graph augmentation, thereby providing stronger decision support for GA accident investigators.

## 1.2 Objectives

The primary objective of this thesis is to design and validate an LLM-based life-cycle framework for GA accident investigation. The framework follows the initial, detailed, and post-investigation stages of real-world practice and focuses on building and improving causal graphs. The aim is to increase the accuracy, clarity, and completeness of causal reconstruction, and to provide investigators and policy-makers with clear and reliable decision support. In line with this overall objective, the thesis pursues the following three specific research objectives:

**Objective 1: Enhance identification of human errors in initial investigation.** Pilot narratives are often unstructured, emotionally influenced, and filled with technical terms, which makes it difficult to extract reliable causal information. This objective is to guide LLMs to more consistently identify and classify human errors from such early-stage evidence by using structured domain knowledge. Achieving this will allow more accurate extraction of initial causal nodes, speed up the initial investigation, and provide a strong basis for later causal recon-

struction.

**Objective 2: Ensure accurate and complete causal graph reconstruction in detailed investigation.** GA accidents usually develop through several interacting events rather than a single failure. Static models cannot capture these dynamic and cross-event relations, and LLMs are often prone to mistakes or omissions in long-chain reasoning. This objective is to build a structured reasoning process that helps LLMs reconstruct clear and time-consistent causal chains across events. The expected result is the creation of complete causal graphs that accurately reflect how accidents progress, thereby improving the reliability of detailed investigations.

**Objective 3: Achieve augmented causal graph in post investigation.** GA accident analysis must explain not only how an accident happened but also why it was more likely under certain conditions. Exogenous risk factors such as pilot background, aircraft type, and mode of operation strongly influence accident likelihood but are often overlooked in causal reconstruction because LLMs lack statistical reasoning ability. The goal of this objective is to achieve augmented causal graphs in post-investigation analysis by incorporating statistically validated risk factors, thereby improving interpretability and providing a more comprehensive basis for safety management and policy decisions.

## 1.3 Scope

This research is confined to GA accident investigation, covering the initial, detailed, and post-investigation stages. The focus is on reconstructing the accident process from already available evidence, such as pilot narratives and investigation records, by employing LLMs for causal graph construction and enhancement. The

scope emphasizes text-based reasoning with LLMs, while excluding activities related to data collection, such as witness interviews, on-site inspections, or crash simulations, as well as broader organizational-level analyses. To ensure alignment with the research objectives, the scope is specified as follows:

**Scope of Objective 1: Domain-informed prompting for human error identification**

At the initial stage of GA accident investigation, the scope is centered on pilot narratives and early witness statements as the primary evidence. The work is confined to extracting and classifying human errors from these unstructured texts by leveraging structured domain knowledge to guide LLMs. The scope does not extend to quantitative data such as sensor readings or flight recorder outputs.

**Scope of Objective 2: Dynamic reasoning workflow for causal chain reconstruction**

For the detailed investigation stage, the scope lies in reconstructing the causal evolution of accidents. The emphasis is on enabling LLMs to capture multi-event chains and organize human, mechanical, and environmental factors into coherent and time-consistent causal graphs. The scope does not include higher-level organizational or systemic modeling beyond the accident sequence.

**Scope of Objective 3: Risk factor integration for post-investigation augmentation**

In the post-investigation stage, the scope expands to enriching causal graphs with contextual risk conditions. This involves incorporating statistically supported external factors—such as pilot characteristics, aircraft type, and operational context—to explain why accidents are more likely under specific circumstances. Broader considerations such as economic evaluation or direct policy implementation re-

main outside the scope.

## 1.4 Significance

By achieving these objectives, this research makes contributions on three levels:

From an academic perspective, it broadens the application of LLMs to GA accident investigation. Previous studies have largely concentrated on extracting information from finalized reports or analyzing isolated factors. In contrast, this work places causal graph construction at the center of inquiry and covers the initial, detailed, and post-investigation stages. By moving beyond static root-cause lists to dynamic, multi-event causal reconstruction, it offers a new perspective for understanding accident evolution and directly addresses the challenges of hallucination and unstable reasoning in specialized, high-stakes domains.

In practice, the thesis introduces a novel paradigm for accident investigation that leverages LLMs to support life-cycle causation reconstruction. Instead of relying solely on lengthy, expert-driven procedures, the framework builds on existing evidence to identify human errors in the initial stage, reconstruct multi-event causal chains during detailed investigation, and integrate contextual risk factors in post-investigation analysis. This approach has the potential to shorten investigation timelines, enhance the consistency of findings, and provide investigators and policymakers with timely and reliable decision support in aviation safety management.

On a methodological level, the research shows how embedding domain knowledge into LLM reasoning processes can improve reliability and interpretability in specialized contexts. The combination of structured knowledge, constrained rea-

soning strategies, and statistical validation establishes a transferable paradigm that extends beyond aviation. It facilitates the shift from static, factor-based accident analysis toward dynamic, graph-based causal narratives, and offers a model for applying trustworthy and explainable AI to other safety-critical fields such as railways, maritime transport, and healthcare.

## 1.5 Structure of the thesis

This thesis consists of six chapters, with the overall organization illustrated in Figure 1.2.

Chapter 1 introduces the background and motivation of this study, defines the research objectives, scope, and significance, and outlines the overall structure of the thesis.

Chapter 2 reviews the literature on general aviation accident analysis, including risk factors, accident causation frameworks, and methods for identifying determinants. It further examines the emerging applications of large language models (LLMs) in accident analysis and investigation, and summarizes mainstream strategies for mitigating hallucination, thereby identifying the methodological gaps that motivate this research.

Chapter 3 focuses on the initial investigation stage. It proposes a domain-knowledge–guided prompting approach that enables LLMs to identify and classify human errors from pilot narratives and other early-stage evidence. Comparative experiments are conducted to evaluate its performance and to demonstrate the value of structured knowledge in enhancing LLM reasoning for accident investigation.

Chapter 4 addresses the detailed investigation stage. It develops a dynamic workflow to reconstruct multi-event causal chains, guiding LLMs to identify event nodes, merge evidence across events, and generate temporally consistent causal graphs. A curated dataset and new evaluation metrics are introduced to validate the effectiveness of the proposed approach in modeling complex accident scenarios.

Chapter 5 investigates the post-investigation stage. It integrates statistically validated exogenous risk factors into LLM-generated causal graphs, thereby extending the analysis from explaining how an accident occurred to why it was more likely under certain conditions. By constructing a structured risk knowledge base and linking it with unstructured narratives, this chapter enhances the interpretability and policy relevance of causal reconstructions.

Chapter 6 concludes the thesis by summarizing the key findings and contributions, highlighting the limitations of the current work, and outlining directions for future research in aviation safety and beyond.

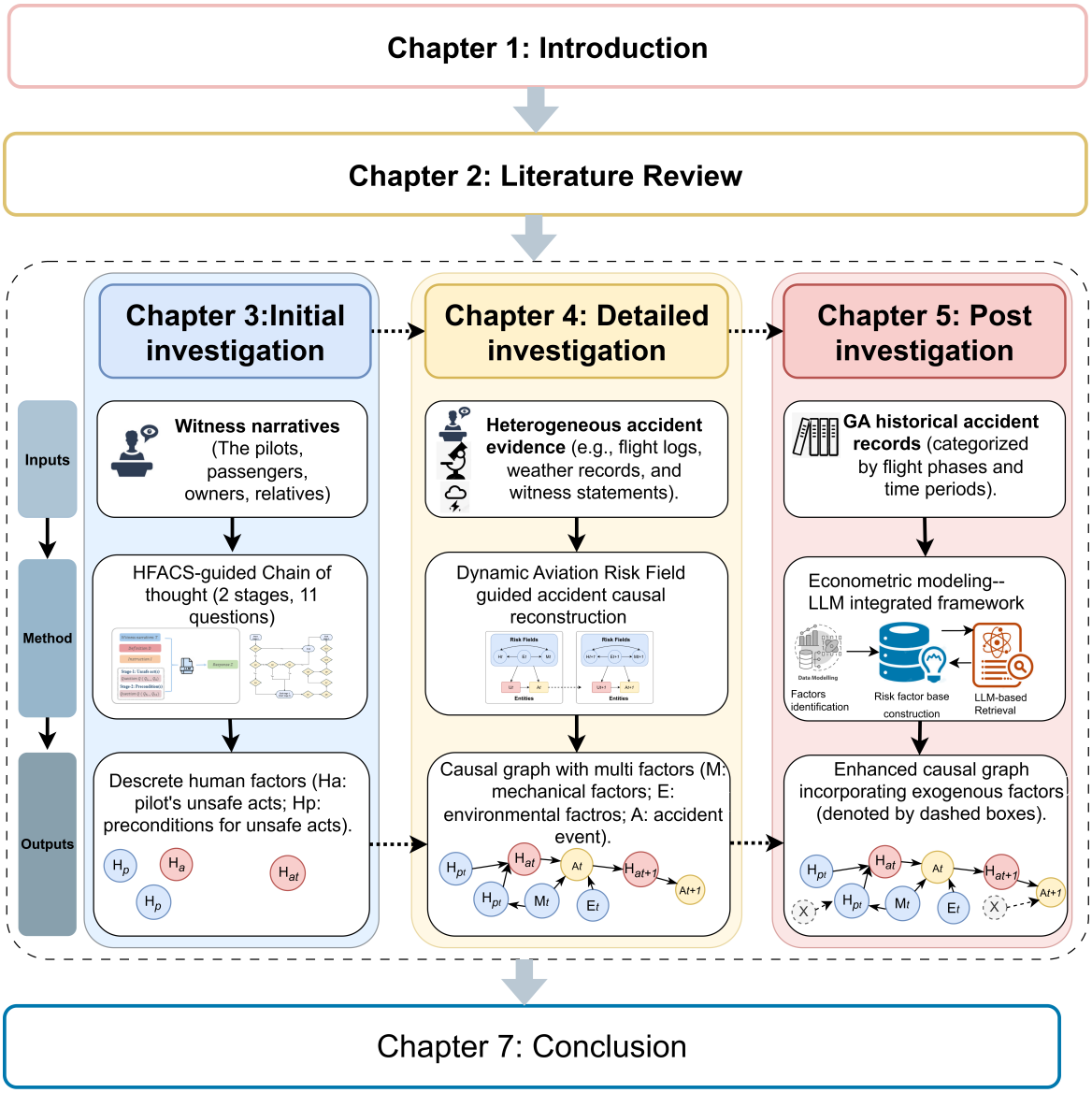



Figure 1.2: Organization of the thesis

# Chapter 2

# Literature Review

This chapter reviews the relevant literature on GA accident risk analysis and the use of LLMs in accident investigation, with the goal of identifying methodological gaps and motivating the framework proposed in this thesis. Section 2.1 examines GA risk factors and accident causation frameworks, highlighting both the determinants of accidents and the theoretical models used to explain causal chains. Next, Section 2.2 discusses the methodological approaches for identifying GA risk factors, ranging from descriptive statistics to econometric regression models, and emphasizes their strengths and limitations. Section 2.3 then introduces the applications of LLMs in accident analysis and investigation, focusing on their roles in information extraction and causal inference. Finally, Section 2.4 outlines the major strategies for hallucination mitigation in LLM reasoning, including prompt engineering, workflow design, and retrieval augmentation, which are essential for ensuring reliable deployment in accident investigation.

## 2.1 Risk factors and causation frameworks in GA

### 2.1.1 Risk factors

Risk factors are elements that significantly influence both the frequency and severity of accidents. Research has progressively highlighted human, mechanical, environmental, and operational determinants. Among these, human factors are consistently regarded as the most influential [11, 12]. Pilot demographics such as age, gender, certification, and experience have been widely examined, but their association with accident severity remains inconclusive. Some studies report that elderly pilots are disproportionately represented in fatal accidents [13–15], whereas others find no significant correlation with age [16]. Similarly, while certain analyses suggest more experienced pilots face higher fatal accident risk [17], other evidence contradicts this claim [16]. Pilot's explicit errors, such as alcohol consumption [18, 19], medication use [20], and IFR certification [21], along with implicit cognitive failures, such as low prospective memory and situational awareness, increase accident likelihood [22] (see Table 2.1).

Aircraft- and environment-related factors further complicate the risk landscape in GA. Unlike commercial aviation, GA involves a large proportion of homebuilt aircraft, and studies consistently show that experimental and amateur-built airplanes are associated with elevated accident probabilities [23]. Aircraft characteristics such as engine number [24] and type [25, 26] have also been identified as significant predictors of fatal GA accidents. Environmental conditions, including lighting [27] and wind velocity [14, 28], have also been found to shape GA accidents.

These studies have advanced the understanding of GA accident determinants;

however, most analyses concentrate on pilot injury severity while neglecting aircraft damage and its correlation with injury outcomes, thereby limiting the comprehensiveness of preventive strategies. Moreover, few investigations explicitly address the phase-specific transferability or temporal stability of risk factors, which restricts dynamic assessment and reduces the precision of targeted safety interventions.

### 2.1.2 Accident causation frameworks

While identifying individual risk factors is fundamental, increasing attention has been paid to understanding the process of how and why a safe flight deteriorates into an accident, which is a critical aspect for effectively guiding accident investigations. To address this, various causation frameworks have been proposed to reveal the causal chains and systemic interactions among risk factors. These frameworks have evolved from linear to more system-oriented approaches. Early models, such as Heinrich's Domino Theory, conceptualized accidents as sequential events [29], while Reason's Swiss Cheese Model highlighted latent failures across multiple defense layers and the interplay of organizational and human factors [30]. Building on this foundation, Shappell and Wiegmann introduced the Human Factors Analysis and Classification System (HFACS), which operationalized the Swiss Cheese Model by explicitly defining its "holes" as unsafe acts, preconditions, supervisory factors, and organizational influences [31]. More advanced systemic models include Rasmussen's AcciMap, emphasizing multi-level system interactions [32], and Leveson's STAMP, which applies control theory to interpret accidents as failures of system control [33]. Collectively, these frameworks

provide essential theoretical tools for understanding the complexity of aviation accidents.

In the specific domain of GA accident analysis, several frameworks have been adapted or developed to better reflect the characteristics of GA operations. For instance, Xue. et al. [34] proposed the Swiss Cheese Model for General Aviation (SCM-GA), which integrates classic accident causation theories with GA-specific laws, regulations, and safety management practices. Chi et al. [35] introduced the human–machine–environment–procedure (HMEP) classification scheme for coding the root causes of helicopter accidents. Among the existing frameworks, HFACS remains the most widely used in GA safety research due to its clarity and structured taxonomy, and it has been extensively applied to GA accident reports [1, 2, 36].

These studies have advanced the understanding of GA accident causes and their interrelationships, providing a solid foundation for investigation. However, unlike commercial operations under Part 135 or Part 121, GA rarely involves complex organizational layers, raising concerns about the applicability of frameworks that emphasize organizational factors [24]. Moreover, traditional frameworks remain primarily static and retrospective, making it difficult to capture the dynamic evolution of accident chains. Their inherent linearity often leads to rigid structures that fail to reflect the complexity and non-linearity of real-world processes. Applied to GA—where small samples, high heterogeneity, and multi-causal interactions across flight phases are common—these frameworks show limited adaptability. Such shortcomings underscore the need for more flexible, dynamic, and data-driven approaches to support GA accident investigation.

## 2.2 Risk factor identification methods in GA

Identifying risk factors seeks to reveal common contributors to accidents through extensive historical data analysis. Identifying GA risk factors is crucial for accident investigation and prevention. The methods used to identify GA risk factors can be primarily divided into two categories: descriptive statistical analysis and regression analysis. Descriptive statistical analysis, such as frequency and proportion analysis, is widely used in studying GA accidents. For instance, Fultz and Ashley (2016) [37]analyzed the proportions of spatiotemporal risk factors in weather-related fatal GA accidents to identify key influences, finding that most occurred between October and April, on weekends, and during early mornings or evenings. Similarly, de Voogt and Louteiro (2024) [38] analyzed 134 NTSB-reported nose-up and nose-down accidents, finding 35% caused by ground loss of control and 58% involving tail-wheel aircraft.

In addition, econometric statistical analysis, especially logistic and Poisson regression, is prevalent in GA determinant studies. Researchers have used these approaches across various datasets and accident categories to explore the relationship between contributors and accident severity. For example, Bazargan and Guzhva (2007)[14] employed logistic regression on U.S. GA accidents (1983–2002) and found that older pilots, darker environments, and retractable landing gear were significantly associated with higher fatality risk. More recent work applied logistic and Poisson regression to turbine-powered aircraft accidents, showing that inadequate preflight and weather planning, checklist deviations, and violations of regulations were strong predictors of fatal or serious injuries [24].

Table 2.1 provides further details. It is important to note that descriptive sta-

tistical analysis can only identify correlations, not infer or explain causal relationships between accident outcomes and risk factors. Econometric methods can overcome this limitation, but the econometric techniques commonly used in GA safety studies, such as logistic and Poisson regression, often fail to account for unobserved heterogeneity, assuming the impact of determinants remains constant across observations. Considering this, advanced methods that account for unobserved heterogeneity are needed to enhance accuracy in analyzing accident determinants.

Table 2.1: Summary of GA risk factors

| Variables | Findings | Data Source | Methodology | Authors |
|---|---|---|---|---|
| Pilot attributes | [Fatal accident] Older pilots↑ | All accidents, US, 1983-2002 | Logistic regression | [14] |
| | [Fatal accident] Pilot age >or = 50 yr↑ | All accidents, US, 1985-1994 | Logistic regression | [13] |
| | [Fatal accident] Advanced age↑ | Single, reciprocating engine accidents, US, 2002-2012 | Logistic regression | [15] |
| | [Accident] Instrument flight rules certification↑ | Single, reciprocating engine accidents, US, 2002-2012 | Logistic regression | [15] |
| | [Fatal accident] Pilot's medical certificate↑ | All accidents, US, 1983-2002 | Logistic regression | [14] |
| | [Fatal accident] Female- | All accidents, US, 1983-2002 | Logistic regression | [14] |
| | [Fatal accident] Age and female- | All accidents, US, 1983–2002 | Logistic regression, Chi-square tests | [16] |
| | [Fatal accident] Airmen with more experience↓ | All accidents, US, 1983–2002 | Logistic regression, Chi-square tests | [16] |
| | [Fatal accident] Airmen with more experience↑ | All accidents, US, 1983–1996 | Multivariate logistic regression | [17] |
| | [Fatal injury] Failed to use lap belt and shoulder harness↑ | Crash landings, US, 1983-1992 | Logistic regression | [39] |
| | [Fatal or serious injury] Checklist/flight manual not followed↑ | Turbine-powered aircraft accidents, US, 1989-2013 | Logistic regression, Poisson regression | [24] |
| | [Fatal or serious injury] Inadequate pre-flight planning↑ | Turbine-powered aircraft accidents, US, 1989-2013 | Logistic regression, Poisson regression | [24] |
| | [Fatal or serious injury] Inadequate weather planning↑ | Turbine-powered aircraft accidents, US, 1989-2013 | Logistic regression, Poisson regression | [24] |

| Variables | Findings | Data Source | Methodology | Authors |
|---|---|---|---|---|
| | [Fatal or serious injury] Violation FARs/AIM deviation↑ | Turbine-powered aircraft accidents, US, 1989-2013 | Logistic regression, Poisson regression | [24] |
| Aircraft attributes | [Fatal accident] Retractable landing gear↑ | All accidents, US, 1983-2002 | Logistic regression | [14] |
| | [Fatal accident] Two-engine aircraft↑ | Go-around accidents, US,2000-2004 and 2013-2017 | Poisson regression | [27] |
| | [Fatal accident] Twin-engine aircraft↑ | Takeoff accidents, US, 2013-2018 | Logistic regression | [40] |
| | [Fatal accident] Multi-engine aircraft- | All accidents, US, 1983-2002 | Logistic regression | [14] |
| | [Accident] Multi-engine aircraft↑ | Turbine-powered aircraft accidents, US, 1989-2013 | Logistic regression, Poisson regression | [24] |
| | [Severe accident] Hot-air balloons↑ | All accidents, German, 1993-2007 | Descriptive statistical analysis | [26] |
| | [Fatal accident] Helicopter↓ | Abrupt maneuver accidents, US, 2008-2020 | Logistic regression | [25] |
| | [Fatal accident] Aircraft of 2–5.7 tons↑ | All accidents, German, 1982-2007 | Descriptive statistical analysis | [26] |
| | [Accident] Homebuilt aircraft ↑ | Sports aviation accidents, US, 1993-2007 | Descriptive statistical analysis | [23] |
| Flight attributes | [Fatal accident] Approach, landing, climb and descent↑ | All accidents, US, 1983-2002 | Logistic regression | [14] |
| | [Fatal accident] Instructional flight ↓ | Go-around accidents, US, 2000-2004 and 2013-2017 | Logistic regression | [27] |
| | [Fatal accident] Longer flight distance↑ | Business accidents, US, 1996-2015 | Poisson regression | [41] |
| | [Fatal accident] Cross-country flights↑ | Cross-country crashes, New Zealand, 1988-2000 | Chi-square tests | [42] |
| | [Fatal accident] The presence of a student pilot ↓ | Abrupt maneuver accidents, US, 2008-2020 | Logistic regression | [25] |
| | [Fatal accident] Off-airport landing↑ | Twin-piston engine airplane accidents, US, 2002-2012 | Logistic regression | [43] |
| | [Fatal accident] Off-airport location↑ | All accidents, US, 1985-1994 | Logistic regression | [13] |
| | [Fatal accident] Distance over 10km from the runway↑ | Fixed-wing aircraft accidents, German, 2000-2019 | Logistic regression | [44] |
| | [Fatal accident] Fire or explosion on the ground↑ | Crash landings, US 1983-1992 | Logistic regression | [39] |

| Variables | Findings | Data Source | Methodology | Authors |
|---|---|---|---|---|
| | [Fatal accident] Post-impact fire↑ | Twin-piston engine airplane accidents, US, 2002-2012 | Logistic regression | [43] |
| | [Fatal accident] Post-impact fire↑ | All accidents, US, 1985-1994 | Logistic regression | [13] |
| | [Accident] Flight into terrain↑ | Mountainous terrain accidents, US, 2001–2014 | Poisson regression | [28] |
| | [Fatal accident] Spring and summer↑ | Takeoff accident, US, 2013-2018 | Logistic regression | [40] |
| | [Fatal accident] October and April↑ | Weather-related accidents, US, 1982-2013 | Descriptive statistical analysis | [37] |
| | [Fatal accident] Early morning and evening periods↑ | Weather-related accidents, US,1982-2013 | Descriptive statistical analysis | [37] |
| Environment attributes | [Fatal accident] Instrument meteorological condition↑ | All accidents, US, 1983-2002 | Logistic regression | [14] |
| | [Fatal accident]Instrument meteorological condition↑ | Go-around accidents, US, 2000-2004 and 2013-2017 | Logistic regression | [27] |
| | [Fatal accident]Instrument meteorological condition↑ | Takeoff accident, US, 2013-2018 | Logistic regression | [40] |
| | [Fatal accident]Instrument meteorological condition↑ | Weather-related accidents, US, 1982-2013 | Descriptive statistical analysis | [37] |
| | [Fatal accident] Night↑ | Go-around accidents, US, 2000-2004 and 2013-2017 | Logistic regression | [27] |
| | [Fatal accident] Nighttime↑ | All accidents, US, 1985-1994 | Multivariate logistic regression | [13] |
| | [Fatal accident] Non-daylight hours↑ | Business accidents, US, 1996-2015 | Poisson regression | [41] |
| | [Fatal accident] Darker environment↑ | All accidents, US, 1983-2002 | Logistic regression | [14] |
| | [Fatal accident] Operations at night↑ | Twin-piston engine airplanes, US, 2002-2012 | Poisson regression | [43] |
| | [Fatal accident] Degraded visibility↑ | Business accidents, US, 1996-2015 | Poisson regression | [41] |
| | [Fatal accident] Higher wind velocity↓ | All accidents, US, 1983-2002 | Logistic regression | [14] |
| | [Fatal accident] Wind gusts/shear↑ | Mountainous terrain accidents, US, 2001–2014 | Poisson regression | [28] |

Note: The upward arrow (↑) represents a positive correlation with accident severity or occurrence, as indicated in the square brackets, the downward arrow (↓) indicates a negative correlation, and a dash (-) denotes no correlation.

## 2.3 LLMs in accident analysis and investigation

### 2.3.1 Information extraction

Historically, safety regulations have been shaped by analyzing risk factors derived from extensive accident data [45]. The first step in this process is extracting crucial information from accident investigation reports. To automate this process and free experts from complex information extraction tasks, previous research has utilized traditional NLP techniques such as Support Vector Machines (SVM), word embedding, and Convolutional Neural Networks (CNNs) for extracting accident information and predicting outcomes [46–48]. However, the domain-specific terminology, abbreviations, colloquial expressions, and extraneous details common in accident investigation reports pose significant challenges to conventional NLP techniques.

Conversely, LLMs excel in tasks such as named entity recognition, text classification, and summarization, and demonstrate remarkable reasoning capabilities with minimal pretraining [49–52]. As such, they hold great promise for improving information extraction from accident reports. Ahmadi et al. [49] utilized GPT-3.5, GPT-4, and Gemini Pro to extract key details from construction accident reports, finding GPT-4 to be the most accurate across most attributes. Studies [53–56] explored the performance of LLMs in extracting details from traffic accident narratives and answering accident-related queries. LLMs can also process accident information based on specific rules. For instance, specialized models like SafeAeroBERT [57] are explored to categorize civil aviation accident reports into four causal categories, while ChatGPT [58] can generate event summaries from accident narratives. Such applications of LLMs in accident analysis are post hoc:

they process finalized investigation reports. While effective for retrospective review, they bypass the more demanding scenario of ongoing investigations, when no final report is available.

### 2.3.2 Causal inference

The ability of LLMs to process multimodal information and perform causal reasoning has been validated across various domains, indicating strong potential for accident investigation. Beyond text, LLMs can integrate images, video, and audio, thereby enhancing both the depth and automation of investigations. For example, Zhang et al. [59] proposed *SeeUnsafe*, a framework that employs multimodal LLM agents to analyze traffic accident videos, transforming the conventional "extract-then-explain" workflow into a more interactive and conversational process. Similarly, Chowdhury et al. [60] developed a multimodal LLM pipeline to automatically scrape, classify, and structure accident data from online news, demonstrating the feasibility of automating complex data workflows through combined visual and language processing. Moreover, different versions of *AccidentGPT* have been proposed. One version leverages collaborative sensing from multiple vehicles and roadside devices to perform intelligent real-time safety analysis involving pedestrians, vehicles, road conditions, and the surrounding environment, ultimately generating detailed reports on accident causes and liabilities [61]. Another version incorporates multimodal inputs to automatically reconstruct accident scenes and generate comprehensive multi-task outputs, including visualizations, summaries, and cause analysis [62].

Regarding causal reasoning, counterfactual inference, and root cause analysis,

LLMs have demonstrated the ability not only to identify and interpret causal relationships in natural language but also to perform more complex tasks such as intervention-based reasoning and counterfactual analysis [63, 64]. These reasoning capabilities have been successfully applied in various real-world contexts such as news event analysis, scientific impact tracking, and social behavior modeling [65, 66]. Given that accident investigations heavily rely on reconstructing causal graphs and identifying root causes, LLMs show promising potential in this domain as well. For example, prior research has explored the use of LLMs for predicting causal factors from highway crash records [67], constructing causal knowledge graphs to support complex event explanation and source attribution [68], and generating causal graphs as domain experts [69].

Existing studies have made preliminary attempts to apply LLMs to accident investigation and construct the causal graph. However, the related research has only used basic, generalized prompt strategies, failing to fully leverage the reasoning capabilities of LLMs in this field. Additionally, while LLMs perform well in isolated cases, they lack statistical modeling strength and struggle to extract insights from large-scale, domain-specific historical data. Combining LLMs with statistical or machine learning models may help bridge this gap, enabling more accurate and scalable accident analysis.

## 2.4 Hallucination mitigation in LLM reasoning

### 2.4.1 Prompt engineering

LLMs hold great potential in accident investigation and analysis, while the hallucination issue in complex logical reasoning presents a significant obstacle [70–72]. This challenge arises from LLMs' bounded internal knowledge since the domain-specific corpus is insufficient during training processes [73]. To address this limitation, recent studies highlight three main strategies for mitigating hallucinations in LLMs: (1) prompt engineering to guide reasoning more effectively, (2) workflow or multi-agent designs that distribute tasks across specialized components, and (3) retrieval augmentation generation that grounds model outputs in external domain knowledge to enhance factual accuracy.

One-shot and few-shot strategies, by offering in-context input-output examples, have been shown to improve LLMs' reasoning [74]. ReAct interleaves reasoning with actions, enabling LLMs to interact with external environments and thereby improve factuality and interpretability [75]. Additionally, the Chain-of-Thought (CoT) prompt, which guides LLMs to decompose complex tasks into simpler steps, enhances their reasoning even further, surpassing few-shot approaches [76]. Strategies such as Tree of Thought (ToT) [77], Graph of Thought (GoT) [78], and Program of Thought (PoT) [79] have been developed to further enhance the global reasoning capabilities of LLMs, while their essence remains aligned with CoT. Wei et al [76] showcased that manually providing a sequence of CoT examples (manual-CoT) can significantly boost LLMs' performance, and even merely adding "think step by step" in the prompt (auto-CoT) markedly enhanced the LLM's reasoning abilities [80]. To address the missing-step errors in auto-

CoT, another enhanced version of auto-CoT, the Plan-and-Solve prompt has been proposed. This prompt guides LLMs to break down the entire task into smaller subtasks (plan) and then solve them according to a plan (solve), achieving performance comparable to few-shot CoT in some dataset tests [81]. Overall, manual-CoT outperforms auto-CoT in practice [76, 80], while this superiority largely depends on the quality of the designed demonstrations. This is particularly true for specific tasks, which require carefully crafted, domain-specific CoT examples because reasoning tasks and complexities vary across domains [82].

Incorporating domain knowledge has been shown to effectively enhance the performance of domain-specific NLP tasks, such as aspect-based sentiment analysis and fault diagnosis [51, 83]. Similarly, some researchers advocate for integrating domain knowledge into Chain-of-Thought (CoT) designs to fully leverage the potential of CoT strategies in guiding LLMs through specialized reasoning tasks. For instance, Chen et al. (2024) [84] integrated API knowledge into the Chain-of-Thought (CoT) process to guide code generation, significantly improving domain-specific code generation performance. Jiang et al. (2024) [9] employed the Bloom Cognitive Model (BCM) to develop a Pedagogical Chain-of-Thought (PedCoT), which aids LLMs in self-correction for mathematical reasoning tasks.

### 2.4.2 Workflow design or multi-agent systems

Beyond prompt optimization, researchers are increasingly adopting structured approaches such as workflow-based design and multi-agent frameworks to mitigate hallucinations in LLMs. These methods decompose complex tasks into modular roles or stages, reflecting standard operating procedures to enhance transparency,

traceability, and reliability. For instance, MetaGPT [85] encodes SOPs into agent workflows by assigning roles such as planner, reviewer, and executor, effectively reducing cascading hallucinations in complex reasoning. Similarly, workflow architectures improve multi-step reasoning by combining ensemble outputs, self-reflection, and deliberation among agents, particularly in robotic planning domains [86]. Multi-agent strategies further enhance robustness by distributing reasoning across interacting agents. Yu et al. [87] introduced a system integrating self-correction, external feedback, and agent debates to improve factual accuracy in vision-language tasks. Duan and Wang [88] demonstrated that involving third-party LLMs in agent consensus increases reliability under uncertainty, while Yang and Ma [89] proposed an adversarial multi-agent framework with voting mechanisms to improve decision consistency in high-stakes domains such as healthcare and law.

Together, these findings underscore the growing consensus that embedding human-like workflows and collaborative roles into LLM execution pipelines is an effective and scalable strategy for mitigating hallucination—especially in high-stakes, structured, or multi-modal domains.

### 2.4.3 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) has recently emerged as a key strategy for mitigating hallucination in large language models (LLMs). By incorporating external knowledge sources, RAG decomposes generation into two stages—retrieval and synthesis—which significantly reduces the likelihood of fabricated outputs, especially in domain-specific or low-resource settings. Despite its ad-

vantages, hallucinations in RAG often arise from two sources: inaccuracies in retrieved documents and semantic drift during synthesis. To address these challenges, researchers have explored a variety of solutions. One line of work focuses on refining the retrieval process, such as the Decomposed Retrieval of Constraint (DRoC) framework [73], which decomposes task constraints and leverages documentation and example codes to better support complex optimization problems. Other studies emphasize detection and evaluation, including the Context Sensitivity Ratio (CSR), a metric designed to identify hallucinated spans that are disconnected from retrieved content [90], and the RAGTruth benchmark, a large-scale human-annotated dataset for training and assessing hallucination-resistant RAG models [91].

In practice, hallucination mitigation in LLMs often combines multiple strategies, as their integration provides greater flexibility and effectiveness. For instance, multi-agent frameworks can be augmented with prompt-based approaches such as the ReAct agent, which employs reasoning–action loops to improve interpretability and faithfulness in complex reasoning tasks, thereby reducing hallucinations [92]. Similarly, retrieval-augmented generation (RAG) can be embedded into multi-agent systems to ground collaborative reasoning in external knowledge. Building on these advances, this study incorporates domain-specific knowledge from accident investigation into the workflow design of LLMs, while leveraging RAG to retrieve external resources such as the GA risk factor knowledge base, with the goal of enhancing their capacity for accident analysis and causal graph generation.

# Chapter 3

# Study 1: HFACS-guided CoT for initial investigation

In this chapter, the classical Human Factors Analysis and Classification System (HFACS) framework is incorporated into the design of Chain-of-Thought (CoT) prompts to enhance the performance of LLMs during preliminary investigations. Based on HFACS, we propose a novel prompt strategy, HFACS-CoT, along with its variant HFACS-CoT+, which aim to guide structured reasoning and reduce hallucinations. Furthermore, a new dataset, GAHFACS, is constructed to support model evaluation and serve as a reusable resource for future research. The structure of this chapter is as follows: Section 3.1 introduces the background and related work. Section 3.2 presents the proposed HFACS-CoT and HFACS-CoT+ prompts. Section 3.3 describes the development of the GAHFACS dataset. Section 3.4 reports a case study with experiments and results. Finally, Section 3.5 summarizes the findings and discusses the limitations of this study.

## 3.1 Background

As discussed in Chapter 1, GA accident investigation is a lengthy and expert-dependent process. Among the three stages of investigation, the preliminary stage plays a critical role. In this stage, deriving preliminary accident causes can offer investigation experts valuable insights, helping to narrow the investigation scope and reduce the time required. Inferring human errors from witness statements gathered during the initial investigation is a practical approach. As human errors are the leading cause of aviation accidents [11, 93], they serve as essential initial leads. Witness narratives, primarily derived from pilots' self-reports or interviews, can yield critical insights into events surrounding accidents. Furthermore, these narratives are valuable for identifying errors related to emotional states or decision-making, which are challenging for onsite investigators to detect.

LLMs have demonstrated impressive capabilities in various natural language processing (NLP) tasks. The unstructured text format of witness narratives presents an opportunity for LLMs. However, witness narratives often intertwine factual accounts of the accident with the personal feelings and emotional reactions. Additionally, these narratives are filled with specialized civil aviation terminology. Analyzing these narratives goes beyond mere basic NLP, requires domain-specific reasoning, a challenge for LLMs that usually struggle with complex reasoning tasks [9]. The Chain of Thought (CoT) prompt has been shown to be effective in enhancing LLMs' reasoning capabilities by breaking down complex tasks into multiple steps [76]. However, this advancement is primarily evident in well-designed steps. How to design CoT for GA accident investigation remains an open challenge. Some studies confirmed the benefit of integrating domain-specific

knowledge into CoT to address specialized reasoning problems [9, 84]. Inspired by these developments, our research addresses the challenges of incorporating domain knowledge into CoT design for GA accident investigation. Specifically, the Human Factors Analysis and Classification System (HFACS) has been applied, a structured accident investigation tool that categorizes human errors into four layers: unsafe acts, preconditions for unsafe acts, unsafe supervision, and organizational influence [94, 95] (see Fig. 5.3). Given its hierarchical structure, HFACS is well-suited to guide CoT design. Thus, an HFACS-based CoT (HFACS-CoT) prompt has been proposed in this research.

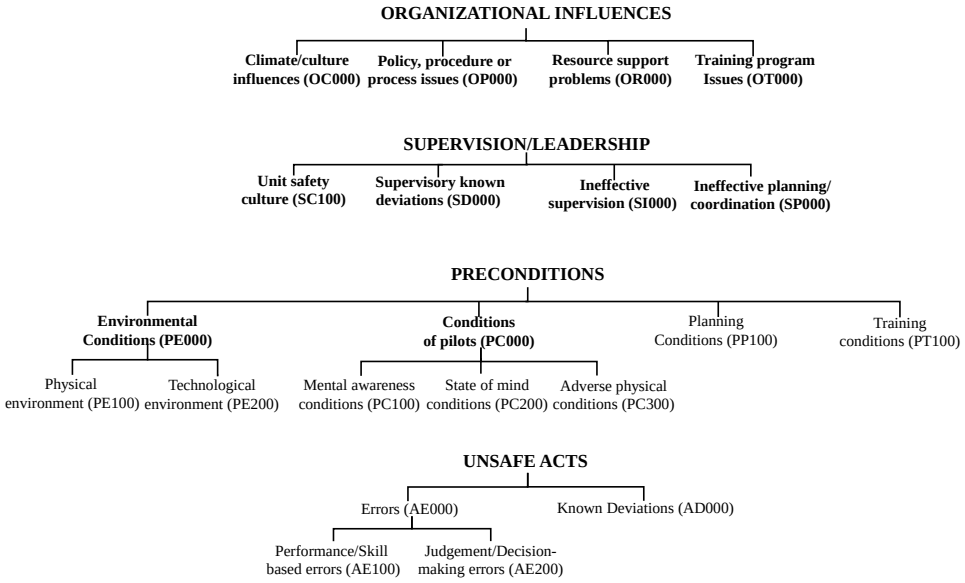



Figure 3.1: Main Structure of HFACS 8.0

Notably, the complete HFACS is not fully applicable to GA accidents, which often involve private flights and rarely include organizational factors [11, 24]. Therefore, the HFACS-CoT focuses on the first two layers: the pilot's unsafe acts

and their preconditions. Accordingly, the HFACS-CoT comprises two stages and multiple steps: the first stage guides LLMs to sequentially determine if the GA accident was due to the pilot's performance/skill-based errors, judgment/decision-making errors, or known deviations. The second stage then identifies the specific preconditions that led to the pilot's unsafe acts. To further enhance LLMs' reasoning capabilities, we developed HFACS-CoT+ by enabling LLMs to process each step individually and integrating programmatic logic statements. Experiments on a self-constructed dataset with GPT-4o demonstrated that our proposed prompts significantly outperform the four baseline prompts. Notably, under the HFACS-CoT+, the f1 scores across all preconditions show a significant improvement compared to HFACS-CoT.

## 3.2 HFACS-CoT and HFACS-CoT+

### 3.2.1 HFACS framework

Our proposed HFACS-CoT and HFACS-CoT+ are based on the HFACS framework. It is a classical tool designed to identify significant human factor trends obscured by the original data structure [96]. The original HFACS framework has undergone several major revisions. The most recent version, HFACS 8.0 proposed by the US Department of Defense [95], is the one adopted in this study. As shown in Fig. 5.3, HFACS 8.0 classifies human errors into four major layers in a structured manner, with each level dependent on the previous one. The layers directly related to accidents begin with the unsafe acts of the pilot. Next are the preconditions leading to these unsafe acts. The following two layers pertain to or-

ganizational factors: supervision leadership from smaller units and organizational influences from larger entities. The prompts proposed in this study focus on the first two layers: unsafe acts and preconditions. The unsafe acts involve errors and known deviations that directly lead to accidents, such as performance/skill-based errors, judgment, and decision-making errors. Preconditions include environmental conditions, conditions of pilots, and planning conditions[1], and training conditions. Within these layers, the environmental conditions and conditions of pilots are further subdivided into smaller categories. For instance, environmental conditions can be classified into the physical environment and the technological environment. Each of these sub-categories presented in Fig. 5.3 also contains smaller subcategories that are not displayed due to space constraints. For more details, refer to [95].

### 3.2.2 HFACS-CoT

Referencing the HFACS 8.0 guidelines, this study proposes HFACS-CoT, an innovative prompt designed to systematically analyze unsafe acts and their preconditions based on witness narratives. As illustrated in Fig 3.2, the two-stage HFACS-CoT prompt generates a series of questions that guide LLMs to sequentially infer the relevant information.

[1]This category was slightly adjusted from 'Team coordination/communication conditions', considering that GA rarely involves team coordination.

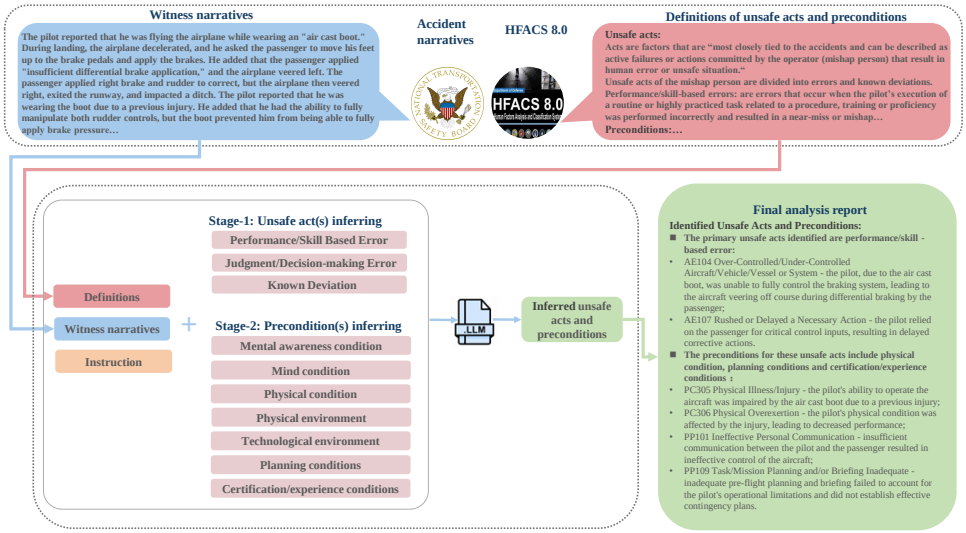



Figure 3.2: The two stages of HFACS-CoT

**Stage 1: Inferring Unsafe Acts**

In stage 1, HFACS-CoT determines the unsafe acts by answering four preliminary questions specifically formulated to probe the narrative $T$. Questions $Q_2$ and $Q_3$ inquire about the pilot's errors, while $Q_4$ addresses the pilot's known deviation. It should be noted that error and known deviation are considered mutually exclusive in this context.

**Stage 2: Inferring preconditions**

The second stage focuses on identifying the preconditions contributing to the unsafe acts identified in Stage 1. Seven questions $Q_i$ are formulated to explore these preconditions. Questions $Q_5$, $Q_6$, and $Q_7$ address conditions related to the pilot, $Q_8$ and $Q_9$ focus on environmental conditions, $Q_{10}$ examines planning-related factors, and $Q_{11}$ pertains to training-related factors.

All the questions $Q$ ($Q = \sum_i Q_i$), the instructions $I$, the narrative $T$, and the definition of HFACS classification $D$ are simultaneously combined and inputted

into the LLMs to obtain the response $Z$ as follows:

$$Z = \text{LLM}(Q, I, T, D) = \{(A_i, C_i, R_i) \mid i = 1, \ldots, 11\} \tag{3.1}$$

where $A_i$ indicates whether the answer to question $Q_i$ is affirmative ('1' for 'yes') or negative ('0' for 'no'). If 'yes', $C_i$ identifies the most applicable error or precondition code associated with $Q_i$, and $R_i$ explicates the reasoning behind each answer.

After obtaining $Z$, the $Q_i$ where $A_i$ =1, along with their corresponding code $C_i$ and reason $R_i$ were selected and then compiled into the final report. That is:

$$\text{Final report} = \{(Q_i, C_i, R_i) \mid A_i = 1, i = 1, \ldots, 11\} \tag{3.2}$$

Specifically, the question set $Q$ and the instructions $I$ can be seen in Table 3.1.

Table 3.1: The question set and response instructions

| Question $Q_i$ | Content | Main instructions $I$ |
|---|---|---|
| UNSAFE ACTS | | 1)If the answer to $Q_1$ is no, then |
| $Q_1$ | Did the mishap individual do or fail to do something that allowed the near-miss or mishap to occur? | do not answer the following questions. |
| | | 2) Only if the answer to both |
| $Q_2(AE100)$ | Was the error a Performance/Skill Based Error? | $Q_2$ and $Q_3$ is no, do you need to answer $Q_4$. |
| $Q_3(AE200)$ | Was the error a Judgment and Decision-making Error? | 3)If the answer to any of $Q_2$, $Q_3$, or $Q_4$ is yes, then start |
| $Q_4(AD000)$ | Was the act a Known Deviation? | analyzing the preconditions of |
| PRECONDITIONS | | the identified unsafe acts. |
| $Q_5(PC100)$ | Did a mental awareness condition influence the unsafe act? | 4) After answering each question, the most applicable |
| $Q_6(PC200)$ | Did a state of mind condition influence the unsafe act? | error code and a brief justification must also be |
| $Q_7(PC300)$ | Did a physical condition influence the unsafe act? | provided. |
| | | 5)Answer in sequence. |
| $Q_8(PE100)$ | Did the physical environment negatively affect performance? | |
| $Q_9(PE200)$ | Did the technological environment negatively affect performance? | |
| $Q_{10}(PP100)$ | Did coordination/communication conditions contribute to the unsafe act? | |
| $Q_{11}(PT100)$ | Did certification/experience conditions contribute to the unsafe act? | |

The instructions $I$ outlined the logical sequence for answering the 11 questions, as clearly illustrated in Fig. 3.3.

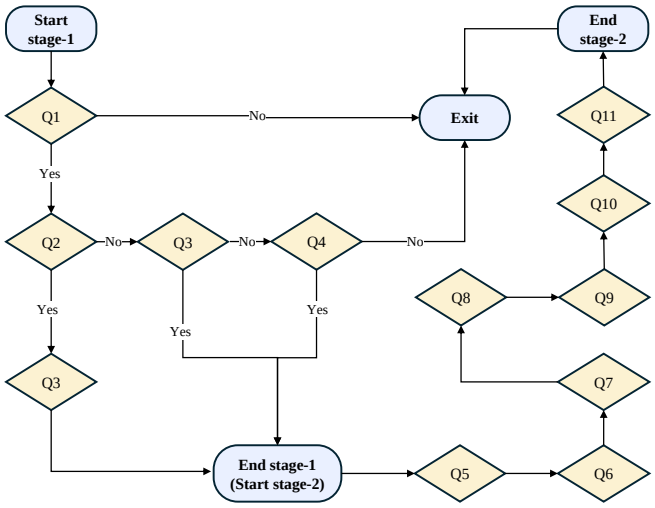



Figure 3.3: Answering logic in two-stage HFACS-CoT

### 3.2.3 HFACS-CoT+

As described in Section 3.2, HFACS-CoT guides LLMs through a systematic chain involving 11 steps (questions) across two stages, and the set of inputs—including all questions $Q$ ($Q = \sum_i Q_i$), the instructions $I$, the narrative $T$, and the definition of HFACS classification information $D$ is fed into the LLM simultaneously. However, responses generated under HFACS-CoT have exhibited instances of hallucinations or logical inconsistencies, such as the concurrent inference of 'error' and 'known deviation' despite instructions indicating their mutual exclusivity.

To address these issues, an enhanced version has been put forward named HFACS-CoT+. The main difference between the two lies in how each step is fed to the LLMs (see Fig. 3.4). In HFACS-CoT+, steps (questions) are fed into the

LLM individually and sequentially, ensuring that the LLMs are not overwhelmed by excessive information at once but instead focus more on processing individual steps. Additionally, HFACS-CoT+ replaces textual instructions $I$ with structured programming logic statements to enhance the logical processing as represented in Fig. 3.2.

The specifics of the two-stage HFACS-CoT+ are as follows:

**Stage 1: Identifying Unsafe Acts**

In the first stage of HFACS-CoT+, different from HFACS-CoT, each question $Q_i$ and its associated HFACS classification information $D_i$ are paired with the constant witness narrative $T$ and fed sequentially into the LLMs, generating responses $Z_i$ as follows:

$$Z_i = \text{LLM}_1(Q_i, T, D_i) = \{(A_i, C_i, R_i) \mid i = 1, 2, 3, 4\} \tag{3.3}$$

Based on the answers $A_i$ and the extracted error codes $C_i$, the unsafe act $E$ is then determined as follows:

$$E = \begin{cases} C_2 & \text{if } A_2 = 1 \text{ and } A_3 = 0 \\ C_3 & \text{if } A_3 = 1 \text{ and } A_2 = 0 \\ C_2 \text{ and } C_3 & \text{if } A_2 = A_3 = 1 \\ C_4 & \text{if } A_2 = A_3 = 0 \text{ and } A_4 = 1 \\ \text{None} & \text{if } A_1 = 0 \text{ or } A_1 = 1, A_2 = A_3 = A_4 = 0 \end{cases} \tag{3.4}$$

**Stage 2: Identifying Preconditions**

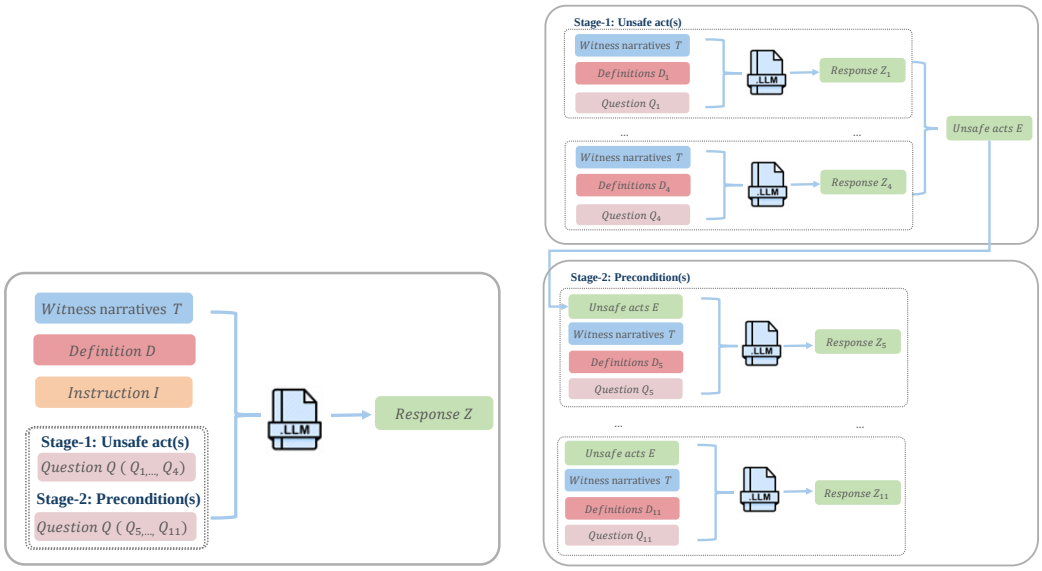
The inferring in Stage 2 is based on the unsafe act(s) $E$ identified in Stage 1.

Seven questions $Q_i$ are formulated to explore the preconditions:

These questions are posed to the narrative $T$, considering the identified unsafe act $E$, and answers $Z_i$ are obtained:

$$Z_i = \text{LLM}_2(E, Q_i, T, D_i) = \{(A_i, C_i, R_i) \mid i = 5, 6, 7, 8, 9, 10, 11\} \tag{3.5}$$

Similar to Eq. (2), by organizing all the $Q_i$ where $A_i$ =1, along with their corresponding code $C_i$ and reason $R_i$, we obtained the final report. The complete HFACS-CoT+ can be found in the appendix.




(a) HFACS-CoT: single-step invocation of LLM

(b) HFACS-CoT+:multi-step invocation of LLM

Figure 3.4: Comparison of question inputs between HFACS-CoT and HFACS-CoT+

## 3.3 GAHFACS dataset construction

Given the current lack of witness narratives and HFACS-coded accident causes, we developed a dataset named GAHFACS to verify the proposed prompts. The GAHFACS dataset was sourced from the National Transportation Safety Board (NTSB) [2], which provides public records of civil aviation accidents that occurred in the United States from 1982 to the present in Microsoft Access (MDB) format [3]. In this study, we focus solely on accidents, excluding the less frequently recorded incidents. Additionally, we compiled records of accidents from January 2008 to April 2024 due to differences in the accident coding system before and after 2008.

The raw dataset includes multiple tables, primarily "accident narratives" and "findings", matched using the "ev-id" column as an identifier. By filtering with the condition "far-part=091", we formed the initial version of General Aviation Human Factors and Accident Causes (GAHFACS) dataset, comprising 18,237 records. We then processed this initial version by extracting witness narratives from the accident narratives and recoding accident findings according to the Human Factors Analysis and Classification System (HFACS).

### 3.3.1 Extraction of witness narratives

Since the NTSB seldom provides raw witness narratives, we employed GPT-3.5-turbo to extract these narratives. Specifically, the NTSB's "narrative" table contains two columns relevant to accident-related descriptions: narr_accp and narr_accf. The narr_accp column includes a comprehensive account of the accident, detailing

[2]The Aviation Safety Reporting System (ASRS) database provides the original witness narratives; however, many accident causes in this database are missing, so we did not choose it in this study.

[3]The complete dataset is available on the NTSB website (https://app .ntsb.gov/avdata/).

the events, locations, aircraft model, sequence of occurrences, witness statements, post-accident examination results, and potential causes. In contrast, narr_accf exclusively stores witness narratives, occasionally supplemented by a brief summary of post-accident findings.

To focus on the witness statements, GPT-3.5-turbo, known for its ability to text process, was used to further process the narr_accf column, keeping only the relevant witness content and removing unrelated information. In cases where only narr_accp was available, GPT-3.5 was used to extract the witness statements directly from narr_accp. The extracted statements were then manually checked to ensure no relevant content was missed, and to remove any irrelevant material, such as statements from FAA investigators or post-accident examination details.

### 3.3.2 HFACS coding of human errors

While the NTSB provides the probable causes of each GA accident, these are often brief and directly state the unsafe acts and outcomes, such as: "The pilot's improper landing flare resulting in a bounced landing and his delay in performing a go-around, which resulted in a collision with trees." Additionally, the "findings" table records various factors related to the causes and outcomes of accidents. Some of these factors can be categorized as unsafe acts and preconditions; however, they are often redundantly coded, and some are not coded using HFACS.

To re-code the GA human errors, two human factors experts were invited. They reviewed and adjusted the previous codes using the HFACS 8.0 framework, referring to the detailed accident narratives to ensure accurate coding. Specifically:

The original accident factors were divided into five major categories: aircraft, personnel issues, environmental issues, organizational issues, and undetermined. Our focus was on the first three categories, especially factors related to the pilot's unsafe acts. For aircraft-related factors, we excluded equipment malfunctions (such as failures, wear, fatigue, or design issues). In terms of environmental issues, we considered only those affecting operations, personnel, response, compensation, and the ability to respond, rather than those affecting equipment, condition awareness, and contributions to the outcome, as the latter do not directly impact pilot performance. Personnel issues were a key focus, encompassing various factors related to pilot behavior, cognition, emotional state, knowledge, and training. Specific findings and their corresponding codes can be found in Table 3.2.

Table 3.2: HFACS re-labeled human errors*

| Category | Subcategory | Subsection | HFACS-Code |
|---|---|---|---|
| Aircraft | Aircraft handling/service | -** | AE100 |
| | Aircraft systems | - | AE100 |
| | Aircraft structures | - | AE100 |
| | Aircraft propeller/rotor | - | AE100 |
| | Aircraft power plant | - | AE100 |
| | Aircraft oper/perf/capability | - | AE100 |
| | Fluids/misc hardware | - | AE100 |
| Personal issue | Physical issue | - | PC300 |
| | Psychological issue | Personality/attitude | PC200 |
| | | Attention/monitoring | PC100 |
| | | Perception/orientation/illusion*** | PC100 |
| | | | PC300 |
| | | | AE200 |
| | | Mental/emotional state | PC200 |
| | | Cognitive overload | PC100 |
| | Experience/knowledge | - | PT100 |
| | Action/decision | Action | AE100 |
| | | Info processing/decision | AE200 |
| | | Miscellaneous | AD000 |
| | | Planning/preparation | PP100 |
| | Task performance | Inspection | AE100 |
| | | Maintenance | AE100 |
| | | Use of equipment/info | AE100 |
| | | Communication (personnel) | PP100 |
| Environmental issues | Physical environment | - | PE100 |
| | Conditions/weather/phenomena | - | PE100 |
| | Task environment | Physical workspace | PE200 |
| | | Pressures/demands | PC100 |

* The meanings of the HFACS-coded accident causes can be found in Fig. 5.3.

** "-" indicates all subsections within the preceding subcategory.

*** In this subsection, geographic disorientation is re-labeled as PC100; visual illusions/disorientations, temporal disorientation, and spatial disorientation as PC300; situational awareness and perception as AE200.

After the initial HFACS coding of the factors, we cross-referenced these re-coded factors with the accident's probable causes. If discrepancies were found, the probable causes were used to adjust the initial coding. Further modifications were made as needed based on the details in the accident narratives.

Ultimately, we removed accidents without any coded causes or witness narra-

tives, resulting in a final GAHFACS dataset of 15,230 records. Each record includes an 'ev-id', 'witness narratives', and 'accident causes coded with HFACS'. The GAHFACS not only provides a reliable dataset for validating our prompt strategies but also serves as a valuable resource for other researchers. The retention of accident IDs allows researchers to extend the dataset by matching additional information to meet diverse research needs, such as pilot details, aircraft information, and environmental conditions.

## 3.4 Case study

### 3.4.1 The selected GAHFACS

The proposed prompts are evaluated using a subset of the GAHFACS dataset. Considering the sample sizes in similar reasoning tasks of LLMs [50, 51, 97], as well as the constraints of time and cost, a total of 835 accident records were used in this research. This dataset consists of 760 randomly selected entries and an additional 75 samples, aimed at ensuring adequate representation of underrepresented sub-tasks (e.g., AD000, PC200, and PE200) by balancing the distribution of positive cases (1s) across all sub-tasks. Although AD000 remains infrequent, its retention facilitates the evaluation of prompts in testing the logical reasoning capabilities of LLMs. As AD000 is mutually exclusive with AE100 and AE200, it serves as a critical test for assessing prompt performance. Fig. 3.5 shows the dataset statistics.

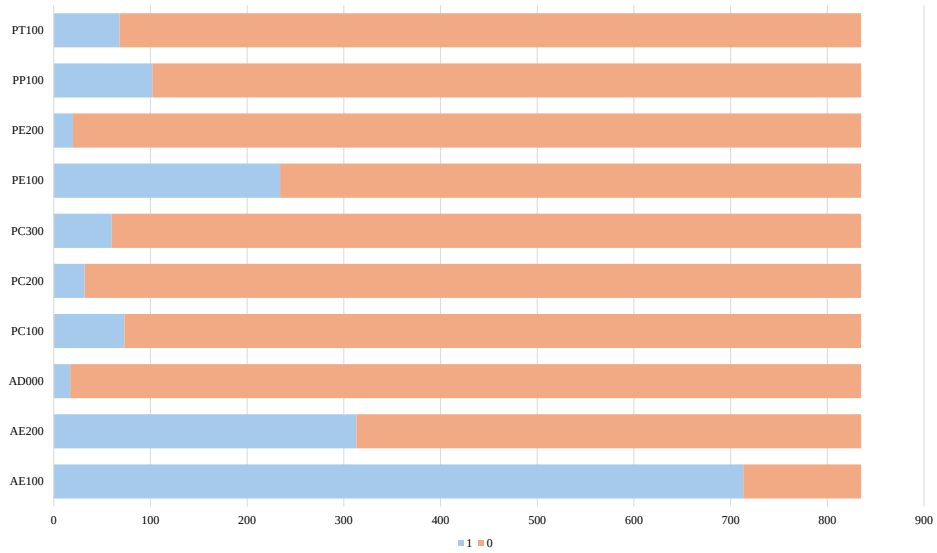



Figure 3.5: Frequency distribution of parameters

### 3.4.2 Baseline prompts

In this study, we compare our proposed HFACS-CoT and HFACS-CoT+ with four baseline prompts: basic zero-shot, basic few-shot, auto-CoT, and plan-and-solve. (1)**Basic zero-shot**: The basic zero-shot prompt only includes the general task requirements, that is, inferring the potential unsafe acts and preconditions that led to the accident based on the witness narrative and referencing the relevant HFACS information. It does not provide additional information, relying solely on the LLM's inherent text inference capabilities. (2)**Basic few-shot**: This prompt extends the basic zero-shot prompt by incorporating three carefully selected examples. These examples consist of three witness narratives along with their corresponding output. Considering the powerful learning capabilities of LLMs, these additional examples are expected to equip the model with some prior experience in GA accident investigation. (3) **Auto-CoT**: This prompt builds upon the basic zero-shot by sim-

ply adding the phrase "let us think step by step." It guides LLMs to decompose the task of accident investigation into subtasks and address each one sequentially until a final conclusion is reached. (4) **Plan-and-solve**: This prompt strategy also extends the basic zero-shot prompt by initially asking the model to create a plan to solve the task and then to execute it step by step. Similarly to HFACS-CoT and HFACS-CoT+, auto-CoT and plan-and-solve are expected to enhance the reasoning ability of LLMs by breaking down complex tasks. The key difference is that the two prompts rely on the model itself to devise the steps, rather than using domain knowledge to guide the step designs.

### 3.4.3 Experimental setting and evaluation

For our experiments, we use GPT-4o accessed through its respective API as the backbone language model, which is one of the state-of-the-art LLMs. The temperature was set to 0.1, and the maximum token limit was configured to 1024 to ensure output consistency. We also evaluated the prompts on other advanced LLMs, including llama-3-8b, GPT-3.5-turbo, and qwen-turbo. However, these models exhibited suboptimal performance on both the HFACS-CoT and HFACS-CoT+ tasks. Specifically, GPT-3.5-turbo consistently answered "yes" to all queries, while qwen-turbo and llama-3-8b frequently generated non-existing or redundant HFACS codes, deviating from the prescribed task requirements[4]. As a result, we focused our evaluation on GPT-4o, which provided more reliable results. The comparison results from the other models are given in Appendix Table6.1 and Table6.2.

The performance of all inferring tasks for the proposed prompts was assessed using common machine learning metrics: f1 score, recall, precision, and accu-

[4]All the sub-tasks requires selecting the most suitable code from a given code set.

racy, with a focus on metrics for the positive class. This focus on the positive class arises from the practical need to accurately identify potential causes during the initial phase of an accident investigation, where the priority is recognizing causes. In this phase, false positive can be ruled out in subsequent stages, whereas failing to identify a true cause (true negative) could result in overlooking critical information. Thus, we also prioritized F1 score and recall to provide comprehensive insights for investigators. Moreover, to evaluate the overall performance of each prompt, we calculated the average metrics across all 10 classification tasks, referred to as average (Ave). To reduce the impact of extreme values from the AD000 inferring task, we calculated the average metrics for the other nine classifications, referred to as average+ (Ave+).

Each sample undergoes 10 binary classification tasks, comprising a total of 11 questions, with $Q_1$ serving as an introductory query that does not contribute to the classification process. For each of the classification tasks, the LLM's response (either "yes" corresponding to 1 or "no" corresponding to 0) is evaluated by comparing it to the ground truth. For example, if an accident is attributed to the pilot's Performance/Skill Based Error (AE100, $Q_2$) and the LLM correctly responds with "yes" to $Q_2$, the response is considered accurate; otherwise, it is deemed incorrect. In addition to answering "yes" or "no," the LLM is required to provide a corresponding error code or precondition code when responding "yes". While these codes are not used in the evaluation due to the lack of ground truth for them[5], they are included to enhance the interpretability of the results and offer additional insight into the model's reasoning process.

[5]The original accident data lacks the necessary detail to support encoding into finer HFACS categories, therefore, it cannot be recoded into more specific error or precondition codes.

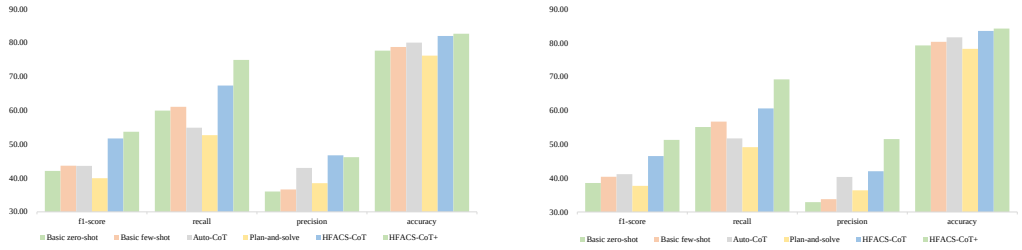
### 3.4.4 Evaluation results

#### 3.4.4.1 Overall evaluation

Table 3.3 presents the performance comparison among six different prompts: basic zero-shot, basic few-shot, auto-CoT, plan-and-solve, HFACS-CoT, and HFACS-CoT+, evaluated on the dataset. As shown in Fig. 3.6a and Fig.3.6b, regarding comprehensive performance (Ave and Ave+), both basic zero-shot and few-shot settings yield moderate results across all metrics, with the latter showing slight improvement due to the additional context provided by the examples. In particular, under the auto-CoT and plan-and-solve, LLM performed even worse than with the basic few-shot prompt in terms of Ave+(F1 score) and Ave+(recall). This suggests that, in complex reasoning tasks, relying on the LLM's inherent ability to generate reasoning steps may lead to inconsistencies or errors in the reasoning process, thereby compromising performance rather than enhancing it.

HFACS-CoT consistently outperforms the four baseline prompts in all average metrics and their adjusted counterparts. HFACS-CoT+ further enhances the reasoning ability of LLMs, reaching the highest performance across all comprehensive metrics. Specifically, HFACS-CoT+ enhances the adjusted average f1 score by 27.51% over the basic zero-shot. Furthermore, in terms of adjusted average recall, HFACS-CoT shows an improvement of 12.34%, while HFACS-CoT+ exhibits a significant increase of 24.98% compared to the basic zero-shot. Our proposed prompt strategies demonstrate the benefits of using domain knowledge to guide CoT design, especially compared to auto-CoT and plan and solve, both rely on automatically generated steps to break down tasks but perform poorly. In addition, the leading performance of HFACS-CoT+ suggests that presenting reasoning

steps one at a time, rather than all at once, may enhance LLMs' reasoning ability by improving focus and reducing information overload. Furthermore, HFACS-CoT+ employs a procedural programming language to represent the answering logic rather than textual instructions, thereby eliminating logical inconsistencies (e.g., outputting both AE100 and AD000 simultaneously) and enhancing the logical inference capabilities of LLMs.




(a) The overall performance on the adjusted average

(b) The overall performance on average

Figure 3.6: Comparison of overall performance: f1 Score, recall, precision, and accuracy

#### 3.4.4.2 Sub-tasks evaluation

The performance of the six prompts across ten sub-tasks was also evaluated (see Table 3.3 and Fig. 3.7a-3.7d). As for the identification to unsafe acts, all prompts performed well in identifying performance/skill-based errors (AE100) and judgment and decision-making errors (AE200), with f1 scores and recall scores exceeding 85 for identifying performance/skill-based errors. It should be noted that all prompts performed poor in inferring known deviations (AD000), with f1 scores below 30. HFACS-CoT even lead to an f1 score, recall, and precision all at 0. These poor results may arise from witnesses' reluctance to report intentional deviations due to potential repercussions. Furthermore, such deviations often involve

complex motivations and psychological states, requiring the model to infer subtle cues from implicit narratives, which remains a challenge for current LLMs.

As for preconditions, HFACS-CoT and HFACS-CoT+ demonstrated significant improvements over the baseline prompts in identifying preconditions, as reflected by higher F1 scores across all precondition identification subtasks. HFACS-CoT+ outperformed HFACS-CoT in six out of seven precondition identification tasks based on F1 scores. Notably, auto-CoT outperforms HFACS-CoT in identifying both physical environment (PE100) and training conditions (PT100), with the highest recall achieved in PT100 identification. Additionally, HFACS-CoT+ surpassed the other four prompts in recall for all subtasks, except for PE100 and PT100.

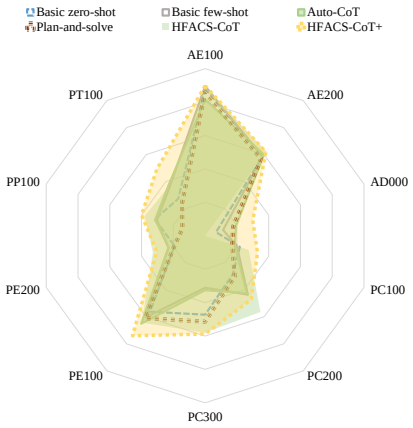
Notably, all prompts underperformed for PC100 in terms of f1 scores, with the best prompt only achieving a 33.02 (see Table 3.3), although recall was relatively high. This suggests that the LLM tends to associate unsafe acts with attention-related preconditions based on the pilots' narratives. Furthermore, for conditions related to pilots' preparedness or communication status, all prompts showed poor F1 scores and recall. This may stem from pilots' reluctance to report their lack of preparation, making it difficult to infer from their narratives that inadequate preparation contributed to unsafe acts. In summary, HFACS-CoT+ achieved a recall exceeding 70 in six out of ten subtasks. Further, as presented by Fig. 3.7a-3.7d, the radar charts demonstrate that HFACS-CoT nearly covers the other four baseline prompts in terms of f1 score, while HFACS-CoT+ almost entirely covers all other prompts across f1 score, recall, precision, and accuracy. This further emphasizes the robustness of our proposed prompts, demonstrating the superiority of well-designed strategies and the potential of LLMs in preliminary GA accident

investigations.

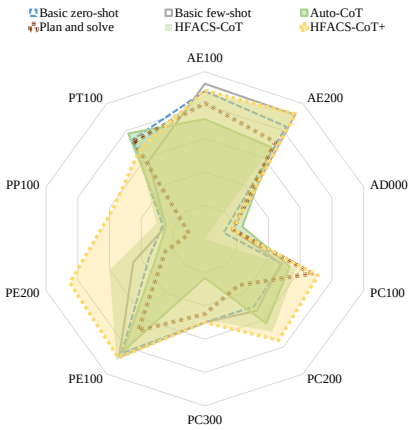
Table 3.3: Performance comparison of four prompts

| Matrix | Prompt | AE100 | AE200 | AD000 | PC100 | PC200 | PC300 | PE100 | PE200 | PP100 | PT100 | Ave | Ave+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| f1-score | Basic zero-shot | 88.42 | 58.57 | 6.90 | 21.88 | 28.32 | 47.24 | 56.17 | 19.18 | 31.76 | 27.32 | 38.58 | 42.10 |
| | Basic few-shot | 89.58 | 58.57 | 11.32 | 20.06 | 43.59 | 31.09 | 60.44 | 23.08 | 30.30 | 36.11 | 40.41 | 43.65 |
| | Auto-CoT | 81.69 | 59.97 | 19.51 | 18.66 | 42.55 | 32.56 | 65.10 | 23.08 | 31.45 | 37.32 | 41.19 | 43.60 |
| | Plan-and-solve | 85.84 | 57.00 | 17.65 | 17.13 | 30.14 | 51.43 | 60.92 | 19.23 | 14.29 | 23.61 | 37.72 | 39.95 |
| | HFACS-CoT | 89.03 | **62.46** | 0.00 | 27.45 | **56.41** | 58.00 | 64.33 | 32.88 | 38.04 | 36.78 | 46.54 | 51.71 |
| | HFACS-CoT+ | **89.80** | 61.95 | **30.00** | **33.02** | 47.06 | **58.82** | **73.97** | **30.36** | **40.00** | **49.18** | **51.31** | **53.68** |
| recall | Basic zero-shot | 88.36 | 82.96 | 11.76 | 48.61 | 50.00 | 50.00 | 84.62 | 35.00 | 26.47 | 73.53 | 55.13 | 59.95 |
| | Basic few-shot | **92.85** | 91.69 | 17.65 | 47.95 | 53.13 | 50.00 | 87.18 | 45.00 | 24.51 | 57.35 | 56.73 | 61.07 |
| | Auto-CoT | 71.67 | 67.73 | **23.53** | 53.42 | 62.50 | 23.33 | 82.91 | 30.00 | 24.51 | **77.94** | 51.75 | 54.89 |
| | Plan-and-solve | 81.21 | 71.57 | 17.65 | 67.12 | 34.38 | 45.00 | 67.95 | 25.00 | 9.80 | 72.06 | 49.17 | 52.68 |
| | HFACS-CoT | 87.10 | **93.29** | 0.00 | 57.53 | 68.75 | 50.00 | **90.17** | 60.00 | 38.04 | 70.59 | 60.62 | 67.35 |
| | HFACS-CoT+ | 88.80 | 92.33 | 17.65 | **71.23** | **75.00** | **75.00** | 88.03 | **85.00** | **57.84** | 66.18 | **69.21** | **74.93** |
| precision | Basic zero-shot | 88.48 | 45.26 | 4.88 | 14.11 | 19.75 | 44.78 | 42.04 | 13.21 | 39.71 | 16.78 | 32.90 | 36.01 |
| | Basic few-shot | 86.54 | 43.03 | 8.33 | 12.68 | 36.96 | 22.56 | 46.26 | 15.52 | 39.68 | 26.35 | 33.79 | 36.62 |
| | Auto-CoT | **94.98** | **53.81** | 16.67 | 11.30 | 32.26 | 53.85 | 55.21 | 18.75 | 43.86 | 24.54 | 40.36 | 42.99 |
| | Plan-and-solve | 91.04 | 47.36 | 17.65 | 9.82 | 26.83 | 60.00 | 55.21 | 15.63 | 26.32 | 14.12 | 36.40 | 38.48 |
| | HFACS-CoT | 91.06 | 46.95 | 0.00 | 18.03 | **47.83** | 72.50 | 50.00 | **22.64** | **50.82** | 20.70 | 42.05 | **46.72** |
| | HFACS-CoT+ | 90.83 | 46.61 | **100.00** | **21.49** | 34.29 | **71.43** | **63.78** | 18.48 | 30.57 | **39.13** | **51.56** | 46.17 |
| accuracy | Basic zero-shot | 80.24 | 56.29 | 93.53 | 70.06 | 90.30 | 91.98 | 62.99 | 92.93 | 86.11 | 68.14 | 79.26 | 77.67 |
| | Basic few-shot | 81.56 | 51.38 | 94.37 | 66.59 | 94.73 | 84.07 | 68.02 | 92.81 | 86.23 | 83.47 | 80.32 | 78.76 |
| | Auto-CoT | 72.57 | **66.11** | 96.05 | 59.28 | 93.53 | 93.05 | 75.09 | **95.21** | 86.95 | 78.68 | 81.65 | 80.05 |
| | Plan-and-solve | 77.13 | 59.52 | 96.65 | 43.23 | 93.89 | 93.89 | 75.57 | 94.97 | 85.63 | 62.04 | 78.25 | 76.21 |
| | HFACS-CoT | 81.68 | 57.96 | 97.25 | 73.41 | **95.93** | **94.97** | 71.98 | 94.13 | **87.90** | 80.24 | 83.54 | 82.02 |
| | HFACS-CoT+ | **82.75** | 57.49 | **98.32** | **74.73** | 93.53 | **94.97** | **82.63** | 90.66 | 78.80 | **88.86** | **84.24** | **82.67** |

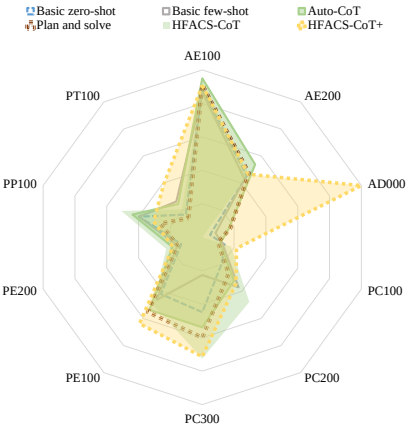
Note:**Bold** values represent the highest in each category, and underlined values represent the second highest. This applies to Table 3.4 as well.



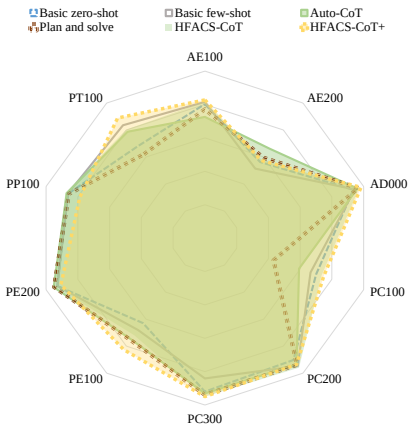

(a) The f1 scores for five prompts across 10 subtasks




(b) The recall for five prompts across 10 subtasks




(c) The precision for five prompts across 10 subtasks




(d) The accuracy for five prompts across 10 subtasks

Figure 3.7: Comparison of 10 sub-task performances for five Prompts

### 3.4.5 Ablation experiment

To further verify the significance of integrating domain knowledge into CoT design and to evaluate the contribution of each stage in HFACS-CoT+, we designed two ablation experiments on half of the experimental dataset. The ablation experiments progressively ablated the two stages of HFACS-style questioning to

observe how these modifications impact performance. During the evaluation of results, we merged the categories of preconditions due to their small sample sizes and assessed the correctness of the model's outputs based on these merged metrics. Specifically, PC100, PC200, and PC300 were combined into PC000 (conditions of pilot); PE100 and PE200 were combined into PE000 (environmental conditions); and PP100 and PT100 were combined into PPT100 (others). The results are shown in Table 3.4.

#### 3.4.5.1 Stage 2 ablation

In the first ablation experiment, we simplified Stage-2 of the HFACS-CoT+ by combining the seven questions related preconditions into a single question:

Single Question for Stage 2: "Considering the witness narratives and the identified unsafe acts, were there any preconditions that led to the unsafe act(s)?"

That is,

$$Q_6, Q_7, Q_8, Q_9, Q_{10}, Q_{11} \rightarrow Q_{\text{preconditions}}$$

This dissolved several steps in Stage 2 that are used to infer preconditions for unsafe acts, while other components remained unchanged. The results in Table 3.4 show that while the f1 score for PPT000 showed a slight increase, the f1 scores for PC000 and PE000 declined significantly, experiencing drops of 17.06 and 9.95, respectively. Both the average F1 score and recall, as well as the adjusted average F1 score and recall, also decreased to some extent. This indicates that step-by-step guidance for LLMs can effectively enhance their reasoning performance.

#### 3.4.5.2 Stage 1 ablation

The second ablation experiment builds on stage 2 ablation by further simplifying stage 1. Here, we merged the four questions in Stage 1 into a single question:

Single Question for Stage 1: "Considering the witness narratives, are there any pilot's unsafe acts that led to this accident?"

$$Q_1, Q_2, Q_3, Q_4 \rightarrow Q_{\text{unsafe acts}}$$

This further simplification aims to assess the feasibility of identifying unsafe acts with a broad, undifferentiated question. The results in Table 3.4 reveal significant drops in the performance metrics for the three unsafe acts and preconditions.

The two ablation experiments demonstrate the importance of each stage and that decomposing complex tasks into multiple steps based on domain knowledge is more effective than asking a single comprehensive question.

Table 3.4: Performance of ablation experiments on HFACS-CoT+

| Matrix | Prompt | AE100 | AE200 | AD000 | PC000 | PE000 | PT100 | Ave | Ave + |
|---|---|---|---|---|---|---|---|---|---|
| F1 score | HFACS-CoT+ | **89.13** | **72.73** | **80.00** | **65.9**5 | **80.00** | 54.21 | **73.67** | **72.40** |
| | Stage 2 ablation | 86.96 | 71.11 | **80.00** | 48.89 | 70.05 | **55.13** | 68.69 | 66.43 |
| | Stage 1 ablation | 76.43 | 68.92 | 0.00 | 46.96 | 67.37 | 50.00 | 51.61 | 61.94 |
| Recall | HFACS-CoT+ | **89.91** | **96.97** | **66.67** | 83.56 | **90.20** | **76.32** | **83.94** | **75.31** |
| | Stage 2 ablation | 87.72 | **96.97** | **66.67** | **91.67** | 67.65 | 56.58 | 77.87 | 69.86 |
| | Stage 1 ablation | 67.54 | 67.83 | 0.00 | 80.56 | 62.75 | 47.37 | 54.34 | 68.39 |

### 3.4.6 LLMs vs. Human expert evaluation

To further evaluate the performance of LLMs in GA accident investigation, we compared the performance of GPT-4o with that of human experts, focusing on identifying human errors from witness narratives. It's important to note that the ground truth in the GAHFACS dataset, used to evaluate the prompt's performance,

is determined by NTSB investigators after considering a wide range of information beyond just witness narratives. This process involves multiple rounds of verification and revision to ensure accuracy. Therefore, to truly assess the potential of LLMs, it is more appropriate to compare their performance with that of human experts under similar conditions—using only witness narratives. For this comparison, a randomly selected sample of 100 accident cases was used [6]. Fig. 3.8 shows the differences in performance between the two, where GPT-4o falls slightly short in assessing pilots' conditions, planning, and training. However, it slightly outperforms human experts in inferring unsafe acts and recognizing environmental conditions. This may be due to GPT-4's ability to systematically analyze and connect different pieces of information within the witness narratives, potentially identifying subtle patterns that human experts might overlook. Overall, GPT-4o demonstrates a marginally higher average F1 score compared to human experts. This finding underscores the significant potential of LLMs in GA accident investigations.

[6] In this sample, no cases were marked with AD000 as 1.

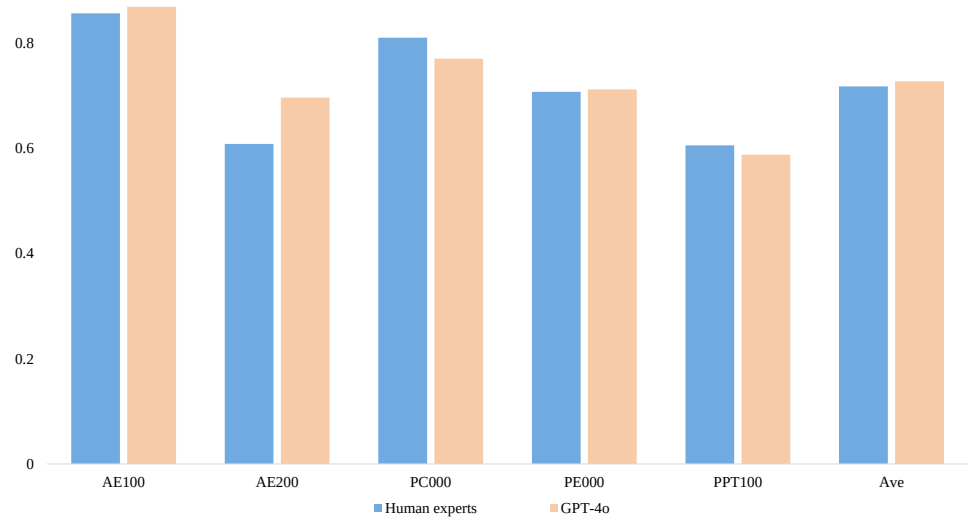



Figure 3.8: Comparison of F1 Score: LLMs vs. Human Experts

## 3.5 Concluding remarks

This research initiates an attempt to apply generative AI within GA accident investigations. Specifically, we utilize LLMs to infer human errors from witness narratives. Guided by the HFACS framework, we integrate this domain knowledge into a CoT design and thus introduce a two-stage HFACS-CoT designed to enhance LLMs' reasoning capabilities. To further refine the performance of LLM, we developed HFACS-CoT+, which delivers each step sequentially and replaces traditional textual instructions with programmatic logic statements. The superior performance of our prompts on a novel GA accident dataset underscores the importance of integrating domain knowledge into prompt design. The two ablation experiments validate the necessity of each stage of the prompts. Additionally, comparisons with human experts further highlight the potential of LLMs in acci-

dent investigations. In summary, the core contributions of this study are:

a) This study explores the potential of LLMs to infer human errors from witness narratives, providing automated insights that accelerate investigations and presenting a new paradigm for accident investigations.

b) We propose the HFACS-CoT and its variant HFACS-CoT+, which integrate domain knowledge into prompt design to enhance the reasoning capabilities of LLMs. The superior performance of these prompts demonstrates how human expertise (e.g., domain-specific knowledge) aids AI in handling specialized tasks, illustrating the potential for future human-AI collaboration in knowledge transfer.

c) This study developed the GAHFACS dataset, which integrates witness narratives from GA accidents with HFACS-coded human errors. Covering records from 2008 to 2024 in the NTSB's GA accident database, this dataset not only validates our proposed prompt performance but also serves as a valuable resource for future research.

d) By comparing the inference results from witness narratives alone, this study confirms that the LLM (GPT-4o) matches or even surpasses the performance of human experts in some human error inferences, revealing the potential of LLMs to assist in accident investigations.

Several limitations require attention: First, since GA accidents with witness narratives usually involve non-fatal accidents, inferring accident causes from these narratives limits our understanding of fatal accident investigations. Second, this study focuses on human errors, neglecting mechanical failures and environmental factors that directly impact aircraft. In future research, we aim to address these limitations by expanding the applications of generative AI in GA accident investigations. For example, by incorporating more comprehensive data such as weather

conditions, pilot logs, and flight plans, our research scope will be expanded to include fatal accidents. Additionally, a multi-agent framework or workflow driven by LLMs, will be developed to investigate different types of accidents and further enhance the performance of AI agents in accident investigations.

# Chapter 4

# Study 2: Risk field for detailed investigation

This chapter introduces a dynamic framework for reconstructing multi-event causal chains in GA accident investigation. The framework is inspired by field theory and models interactions between human, mechanical, and environmental conditions together with unsafe acts and accident events. Guided by this structure, a DARF-based workflow is developed to reduce hallucinations in long-chain reasoning and to generate coherent, temporally consistent causal graphs. In addition, a dedicated dataset and evaluation metric are constructed to support systematic validation. This chapter is organized as follows: Section 4.1 provides the background and motivation of this work by outlining the challenges of dynamic causal reconstruction and the limitations of existing accident investigation frameworks. Section 4.3 presents the proposed DARF-ACR framework in detail. Section 4.2 describes the construction of the AeroCausal dataset for evaluation. Section 4.4 reports a case study with experiments and results. Finally, Section 4.5 summarizes

the findings, discusses the limitations of this study, and highlights directions for future research.

## 4.1 Background

As outlined in Chapter 1, the detailed investigation stage is central to GA accident analysis, where investigators reconstruct how an accident unfolded by integrating more heterogeneous evidence than in the initial investigation stage into causal graphs. Recent advances in LLMs offer new opportunities to automate this process [6], thereby accelerating the whole GA accident investigations. Current applications of LLMs in accident analysis, however, are largely post hoc: they process finalized investigation reports to extract needed information such as structured facts, incident synopses, or causal entities [56, 98, 99]. While effective for retrospective review, they bypass the more demanding scenario of ongoing investigations, when no final report is available. Recent work has begun applying LLMs to accident investigation, for example by predicting causal factors from highway crash records [67] or analyzing witness statements to identify human factors in GA accidents [100]. These studies demonstrate LLMs' potential for early-stage accident investigation but produce isolated factors rather than integrated, multi-event causation chains. In GA, accidents often arise from the connection of multiple factors, unlike commercial aviation. Less stringent training, simpler systems, relaxed procedures, and lighter oversight increase GA's susceptibility to mechanical, environmental, and human risks. Moreover, these accidents result from multi-event causal chains. Causal reconstruction requires reasoning across events while capturing temporal and causal dependencies, a challenge for LLMs that struggle with domain-specific

long-chain reasoning [101].

Injecting domain knowledge has been shown to enhance LLM reasoning in specialized tasks[73, 100, 102]. Building on this insight, we aim to incorporate a dedicated accident investigation framework to guide LLMs in accurately reconstructing GA accident causation graph. However, existing accident investigation frameworks, such as HFACS, SHEL, AcciMap, and STAMP, face notable limitations: (1) incomplete factor coverage, often overemphasizing human factors and treating mechanical and environmental states as background (2) overly linear and rigid structures that fail to capture within-level coupling, cross-level feedback, and concurrent causation [103]; (3) static analyses focusing on single event rather than the dynamic evolution of multiple events; and (4) excessive complexity and control-oriented design in system-theoretic models like STAMP, which prioritize hierarchical control structures over explicit causal chains. Therefore, a concise framework with comprehensive factor coverage, flexible structure, and dynamic feedback is needed for effective GA accident causation reconstruction.

To address this gap, we propose the **Dynamic Aviation Risk Field** (DARF) framework to guide LLM-based causation reconstruction in GA accidents. Inspired by Lewin's field theory, DARF models each accident through two interconnected layers: a static within-event structure where three risk fields (unsafe human, mechanical, and environmental conditions) interact with two entities (unsafe pilot acts and accident events), and a dynamic cross-event structure representing causal links from one event to the next. Guided by this structure, we introduce **DARF-guided Accident Causation Reconstruction workflow** (DARF-ACR) , a three-stage LLM workflow that: **1)** identifies all event nodes in the accident and their associated field and entity nodes; **2)** merge equivalent nodes using narrative

and temporal cues while preserving distinctions for dynamic progression; **3)** link causes within and across events following DARF's dynamic connections to form coherent multi-event causal graph.

To evaluate the effectiveness of DARF-ACR, we construct **AeroCausal**, a GA accident causation dataset of multi-event causal graphs from NTSB narratives. We further propose a new bridge-tolerant evaluation metric that rewards legitimate intermediate expansions while penalizing causal errors, thereby aligning assessment with the generative characteristics of LLMs. Experiments on AeroCausal with multiple LLMs demonstrate that DARF-ACR delivers consistent and significant improvements in GA accident causation reconstruction, with notable gains in both node identification and edge detection.

## 4.2 Dynamic aviation risk field guided Accident Causal Reconstruction (DARF-ACR)

Our approach aims to improve LLM-based multi-event causation reconstruction for GA accidents by embedding domain knowledge into the workflow design. To this end, we propose the **Dynamic Aviation Risk Field** (DARF) framework, which provides broader factor coverage, greater structural flexibility, and streamlined causal logic for dynamic analysis compared to existing investigation frameworks. Building on DARF, we develop the a three-stage process that segments accidents into multiple events and operates within and across events to generate and validate multi-event causal graphs from diverse accident evidences.

### 4.2.1 DARF framework

The **Dynamic Aviation Risk Field** (DARF) framework is inspired by the physical concept of a "field" and Lewin's field theory [104], which states that individual behavior $B$ is determined jointly by the person's internal state $P$ and the surrounding environment $E$ as $B = f(P, E)$. This principle concisely explains how interactions between an individual and their environment shape behavior. Similarly, in GA accident context , unsafe human, mechanical, and environmental conditions can be regarded as pervasive "background fields" that define the accident context and provide latent triggers. The causal entities, that is, unsafe pilot acts and unsafe flight states (accident events) are the observable outcomes of the interacting fields. For example, icing weather, pilot fatigue, and an instrument water-ingress failure may jointly lead to an erroneous control input, which in turn triggers a loss-of-control accident event.

In additon, GA accidents usually evolve through a sequence of events. Each event can influence the risk fields and entities of subsequent events, potentially amplifying hazards, reducing consequences, or even interrupting the accident chain. Therefore, effective accident causal reconstruction must capture both within-event causal mechanisms and influences across events, rather than treating the event sequence as a passive linear chain or focusing on isolated events.

To formalize these ideas, the DARF framework defines a *three-field–two-entity* accident investigation model:

- **Fields**: unsafe human condition $H$ (e.g., fatigue, distraction), unsafe mechanical condition $M$ (e.g., engine failure, instrument malfunction), and unsafe environmental condition $E$ (e.g., low visibility, turbulence).

- **Entities**: unsafe pilot acts $U$ (e.g., erroneous control input) and accident events $A$ (e.g., stall, collision).

Based on these definitions, DARF characterizes accident causation through two complementary structures: a static within-event structure and a dynamic cross-event structure, see Figure 4.1.

**Static within-event structure** Within a single event at time $t$, the three risk fields independently or jointly influence the entities, while also interacting with each other. This can be expressed as:

$$U_t = f_U(H_t, M_t, E_t), \tag{4.1}$$

$$A_t = f_A(H_t, M_t, E_t, U_t), \tag{4.2}$$

where $f_U(\cdot)$ and $f_A(\cdot)$ represent causal functions mapping field states to unsafe acts and accident events, respectively. Additionally, while the mechanical ($M_t$) may influence human ($H_t$) conditions, the environmental condition ($E_t$) acts as an external driver that affects $H_t$ and $M_t$ but is not altered by them. This static structure captures the causal relationships specific to a given event.

**Dynamic cross-event structure** An accident typically unfolds as an ordered sequence of $T$ events $\{A^{(1)}, A^{(2)}, \ldots, A^{(T)}\}$. DARF constrains cross-event links such that the preceding event node $A_t$ may influence the fields and entities of the subsequent event:

$$X_{t+1} = g_X(A_t), \quad X \in \{H, M, E, U, A\}, \quad t = 1, \ldots, T-1, \tag{4.3}$$

where $g_X(\cdot)$ denotes a causal update function that propagates the impact of the prior event into the next event's conditions and outcomes. This design ensures temporal directionality and prevents spurious cross-event links.

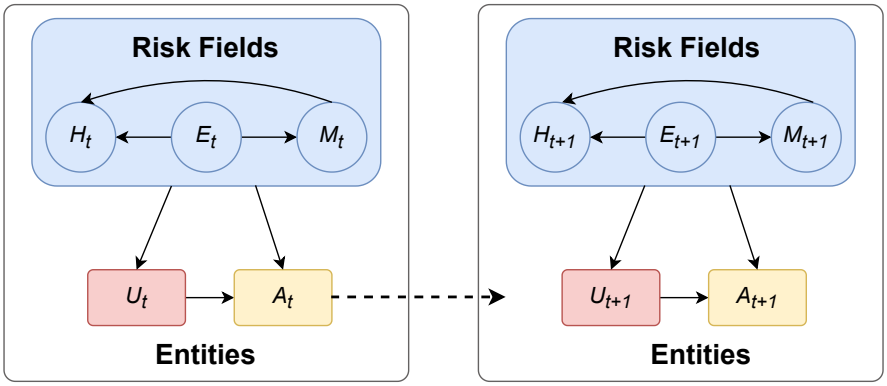



Figure 4.1: Illustration of the DARF framework

### 4.2.2 DARF-ACR workflow

We operationalize the DARF ontology into a three-stage workflow, fully executed by LLMs. Each stage is formulated as an LLM-driven operator, transforming accident evidence into a temporally consistent and causally interpretable accident causation graph, see Fig.4.2.

#### 4.2.2.1 Stage 1: node identification and validation

This stage builds on the DARF "field–entity" ontology to extract a structured node set from the accident evidence collection. Let Evidence $=$ {*narrative*, *findings*, *event_sequence*}, where each element provides complementary information about the accident. The system first identifies all accident event nodes $A$, and then expands to their associated causal factors $\{H, M, E, U\}$, forming the initial node set. Each node is represented inline as a quadruple $n = (C, c, d, p)$,

where $C \in \{A, U, H, M, E\}$ is the high-level category, $c$ is a subclass label (e.g., UC101 physiological state, UA104 operational error, engine failure), $d$ is a natural-language description (e.g., "pilot spatial disorientation," "engine fire"), and $p$ is the corresponding flight phase (e.g., *Enroute*, *Landing*).

**Step 1: accident event extraction.** The LLM extracts accident events from the evidence set:

$$A \;=\; \text{LLM}_{\text{event}}(\text{Evidence}) \tag{4.4}$$

**Step 2: causal factor extraction.** For each event $a \in A$, the LLM identifies associated causal factors such as risk fields or unsafe acts:

$$H,\; M,\; E,\; U \;=\; \text{LLM}_{\text{cause}}(a, \text{Evidence}), \quad a \in A \tag{4.5}$$

**Step 3: unified validation.** The LLM performs global validation to ensure integrity, consistency, and coverage of all nodes:

$$N' = \text{LLM}_{\text{validate}}(N), \quad N = A \cup H \cup M \cup E \cup U \tag{4.6}$$

such that every node is a well-formed quadruple $(C, c, d, p)$ satisfying three constraints: field integrity (all four elements must be present), category–description consistency (e.g., "engine failure" must belong to class $M$), and phase validity with coverage (every node must be assigned a valid flight phase, and each accident event must be linked to at least one upstream causal factor).

#### 4.2.2.2 Stage 2: node deduplication

The LLM removes redundant nodes by semantic similarity checking:

$$N'' = \text{LLM}_{\text{dedup}}(N') \tag{4.7}$$

Two nodes are merged if they share the same subcategory, their descriptions are judged semantically similar by the LLM, and they belong to the same flight phase. In such cases, the LLM either fuses their descriptions or retains the most informative one. This design avoids erroneous deletion: since the same subcategory may appear in different phases or contexts, requiring alignment in both semantics and phase ensures that only genuine duplicates are removed, while distinct but related factors are preserved.

#### 4.2.2.3 Stage 3: edge identification and validation

**Step 1: Intra-event Edges.** LLM generates candidate intra-event edges following DARF static rules:

$$E_{\text{intra}} = \text{LLM}_{\text{edge-intra}}(N'') \tag{4.8}$$

Among this: risk fields $(H, M, E)$ point to unsafe acts or accident events $(U, A)$; each accident event $A_t$ may lead to an unsafe act $U_t$; human and mechanical conditions $(H_t, M_t)$ can be bidirectionally linked; and environmental conditions $(E_t)$ influence both $H_t$ and $M_t$.

**Step 2: Inter-event Propagation.** LLM generates temporal edges across

events:

$$E_{\text{inter}} \;=\; \text{LLM}_{\text{edge-inter}}(N''), \quad X_t \to X_{t+1},\; X \in \{A, U, H, M, E\} \tag{4.9}$$

This restricts edges to forward links between nodes of the same type across successive steps, capturing causal progression while preventing backward or cyclic dependencies.

**Step 3: Edge Validation.** LLM verifies global consistency:

$$E' \;=\; \text{LLM}_{\text{edge-validate}}\big(E_{\text{intra}} \cup E_{\text{inter}}\big) \tag{4.10}$$

removing invalid links, correcting directions, and pruning cycles to ensure a directed acyclic graph.

The final output is:

$$G \;=\; (N'', E') \tag{4.11}$$

where $N''$ is the deduplicated node set and $E'$ is the validated edge set; the resulting DAG preserves domain fidelity, temporal consistency, and causal interpretability.

### 4.2.3 Causal edge recall (CER) metric

In safety-critical domains such as aviation accident analysis, evaluating causal graph predictions requires metrics that balance precision with robustness to incomplete annotations. Existing measures are inadequate: *edge recall* strictly rewards exact matches, ignoring semantically valid indirect links, while *path recall* is overly permissive, treating long chains (e.g., $A \to C \to \cdots \to B$) as equivalent to direct edges ($A \to B$). These limitations are particularly problematic when

ground-truth annotations simplify causal chains by omitting latent mediators.

We introduce **causal edge recall (CER)**, a metric that explicitly accounts for indirect causal connections while penalizing their length. For each ground-truth edge $(u, v) \in E_{\text{gt}}$, CER checks whether a path exists in the predicted graph $G_{\text{pred}}$ and assigns partial credit inversely proportional to the squared path length:

$$\text{CER} = \frac{1}{|E_{\text{gt}}|} \sum_{(u,v)\in E_{\text{gt}}} \mathbb{1}(\text{path}(u \to v) \in G_{\text{pred}}) \cdot \frac{1}{L(u,v)^2} \tag{4.12}$$

where $E_{\text{gt}}$ is the ground-truth edge set, $G_{\text{pred}}$ the predicted directed graph, and $L(u, v)$ the shortest path length between $u$ and $v$. Here, $\mathbb{1}(\cdot)$ denotes the *indicator function*, i.e., $\mathbb{1}[P] = 1$ if the proposition $P$ holds and 0 otherwise (in our case, whether a directed path from $u$ to $v$ exists in $G_{\text{pred}}$). The quadratic penalty $1/L(u, v)^2$ ensures full credit for exact edges ($L = 1$), moderate credit for short indirect paths ($L = 2 \Rightarrow 0.25$), and negligible credit for longer, less reliable chains. Thus, by emphasizing direct causal relations while still crediting latent mediators, CER achieves a balance between sensitivity and robustness.

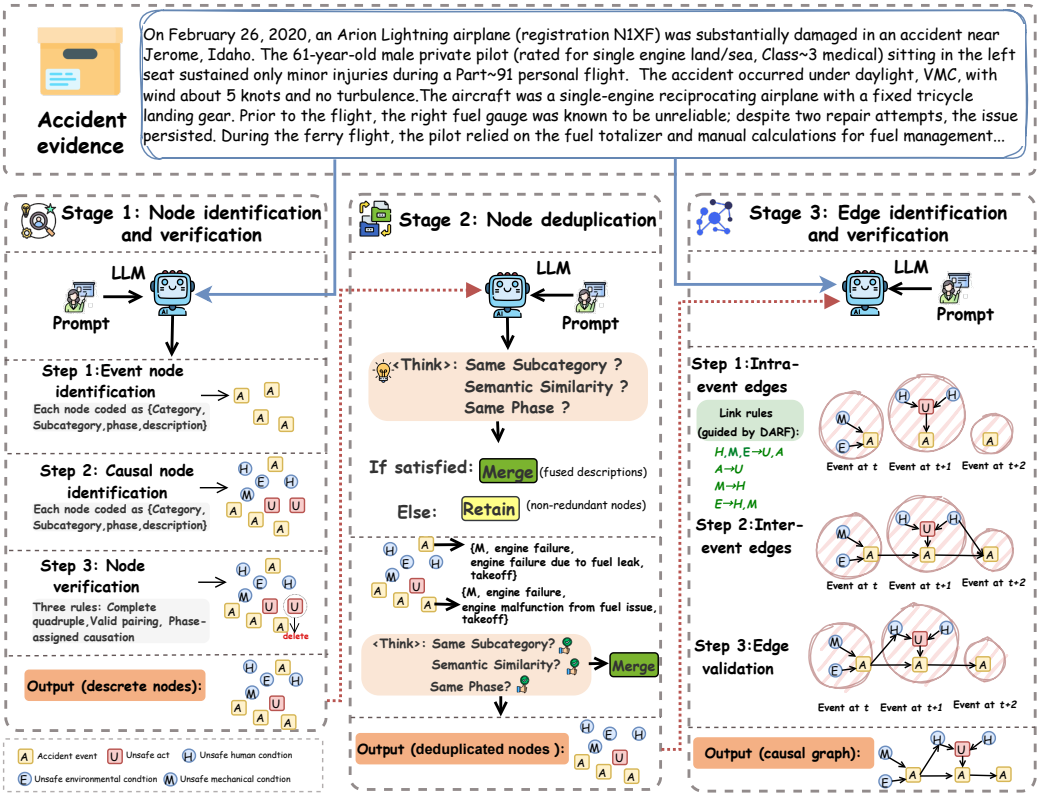



Figure 4.2: The DARF-ACR workflow

## 4.3 AeroCausal dataset construction

To evaluate causal graph generation and provide a benchmark for structured reasoning in GA accident investigation, we present **AeroCausal**, the first dataset modeling accidents as domain-informed causal graphs. Unlike prior resources limited to narratives or factor lists (e.g., NTSB findings, ASRS), AeroCausal encodes full accident progressions, with nodes representing contributing factors or events and edges $E \subseteq V \times V$ capturing causal or temporal dependencies.

The dataset is derived from the National Transportation Safety Board (NTSB) accident records (Microsoft Access MDB format), with three key sources linked

via accident identifier ('ev-id'): (1) 'narr_accp' accident narratives, (2) structured 'findings', and (3) the 'event_sequence' table. AeroCausal contains two complementary components: (i) refined accident evidence serving as model inputs, and (ii) expert-annotated causal graphs providing ground truth for evaluation. Construction proceeds in two stages: accident evidences preprocessing and causal graph annotation.

### 4.3.1 Accident evidences preprocessing

Accident evidence provides the primary factual basis for causal inference. The `narr_accp` column records details such as flight phase, aircraft state, pilot actions, weather conditions, witness accounts, and post-accident findings. However, many entries also include investigators' causal speculations (e.g., "It is likely that the pilot failed to maintain airspeed, leading to a stall"). To ensure that models perform reasoning rather than memorizing conclusions, we preprocess this column by removing inferred causes while preserving all factual observations.

The resulting evidence corpus is denoted as $\mathcal{N} = \{n_1, n_2, \ldots, n_M\}$, where each $n_i$ is a refined accident evidence record. Formally, for each raw entry $n_i^{\text{raw}}$, we construct $n_i$ as:

$$n_i = f_{\text{filter}}(n_i^{\text{raw}}), \quad f_{\text{filter}} : \text{remove inferences while retaining facts.}$$

This guarantees that AeroCausal evaluates genuine causal reasoning rather than reproduction of investigator conclusions.

### 4.3.2 Causal graph annotation

TTo construct ground-truth DAGs, we adopt a standardized taxonomy based on the human–machine–environment framework, which is widely used in accident causation analysis. Since GA accidents primarily involve individual operations without organizational layers, we focus on four node categories: (1) unsafe human conditions, (2) unsafe mechanical conditions, (3) unsafe environmental conditions, and (4) accident events. Each category is further divided into fine-grained subcategories (see Table 4.1), enabling precise and consistent causal representation.

Table 4.1: Taxonomy of contributing factors for AeroCausal dataset

| Category | Subcategory | Code |
|---|---|---|
| Accident event | Loss of control in flight | LOC-I |
| | Loss of control on ground | LOC-G |
| | Abnormal runway contact | ARC |
| | Runway excursion | RE |
| | Runway incursion | RI |
| | Undershoot/overshoot | USOS |
| | Fire/smoke post-impact | F-POST |
| | Birdstrike | BIRD |
| | Midair collision | MAC |
| | Collision with terrain/object | CTOL |
| | Off-field/emergency landing | OFF |
| | Landing gear collapse | LGC |
| | Nose over/roll over | NOR |
| | Part separation | SEP |
| | Normal return/landing | NRL |
| | Controlled flight into terrain | CFIT |
| | Airframe vibration | AV |
| | Other | OTHR |
| Human factors | Inadequate visual lookout | HF101 |
| | Inadequate preflight preparation | HF102 |
| | In-flight fuel mismanagement | HF103 |
| | Loss of aircraft control | HF104 |
| | Improper maneuvering | HF105 |
| | Improper decision-making | HF106 |
| | Aircraft configuration errors | HF107 |
| | Critical equipment non-activation | HF108 |
| | Attention failure | HF109 |
| | Psychological state impairment | HF110 |
| | Physiological state impairment | HF111 |
| | Crew communication failures | HF112 |
| | Training deficiency | HF113 |
| Mechanical factors | System/component malfunction (powerplant) | MF101 |
| | System/component malfunction (non-powerplant) | MF102 |
| | Fuel-related issues | MF103 |
| | Loss of engine power | MF104 |
| | Instrument reading error | MF105 |
| | Aircraft maintenance issues | MF106 |
| | Manufacturing defects | MF107 |
| | Installation errors | MF108 |
| | Aircraft overweight before takeoff | MF109 |
| Environment factors | Weather-related hazards | EF101 |
| | Obstructing objects | EF102 |
| | Cockpit contaminants | EF103 |
| | Noise interference | EF104 |
| | Wildlife or external intrusions | EF105 |
| | Runway or terrain challenges | EF106 |
| | Infrastructure-related hazards | EF107 |
| | Fire/smoke (non-impact) | EF108 |

In practice, nodes are drawn from three sources: events from the 'event_sequence' table, contributing factors from the 'findings' table (mapped to our taxonomy), and contextual details from the 'narr_accp' and 'narr_cause' fields. Annotators then integrate causal and temporal cues from these sources to form DAGs. In addition, a common challenge is that the 'event_sequence' table sometimes records only initial triggers (e.g., LOC-I), while downstream developments such as OFF $\rightarrow$ CTOL $\rightarrow$ SEP $\rightarrow$ F-POST are omitted. To address this, annotators cross-reference evidence and cause summaries to recover missing steps, ensuring that each causal graph captures both triggering events and subsequent patterns.

The resulting **AeroCausal** dataset covers 406 GA accidents with 2,009 nodes and 1,701 directed edges. Graph sizes range from 2-node structures to chains exceeding 10 nodes, averaging 4.9 nodes and 4.2 edges per DAG. To enable evaluation under varying reasoning difficulty, DAGs are stratified into four tiers by edge count: 1–2, 3–4, 5–6, and $>$6 edges (see Table.4.2).

Table 4.2: Dataset statistics across complexity tiers

| **Complexity** | **Cases** | **Edges** | **Nodes** |
|---|---|---|---|
| Simple | 50 | 93 | 143 |
| Moderate | 208 | 727 | 922 |
| Complex | 116 | 626 | 698 |
| Very complex | 32 | 255 | 246 |
| Total | 406 | 1701 | 2009 |

Beyond scale, AeroCausal reveals domain-specific patterns. Human-related factors dominate ($\sim$60% of DAGs), underscoring the pilot-centric nature of GA operations, while mechanical and environmental factors appear less frequently but remain critical. Accident event nodes typically terminate the DAGs, reflecting how upstream conditions escalate into observable outcomes.

In summary, AeroCausal establishes the first publicly available dataset that

systematically models GA accident causality as domain-informed DAGs. It provides not only a benchmark for our LLM-based framework but also a resource for broader tasks such as multi-hop causal reasoning, narrative-to-graph translation, and interpretable safety-critical inference. The dataset will be released upon acceptance to facilitate reproducibility and foster future research.

## 4.4 Case study

To evaluate the effectiveness of the proposed DARF-ACR workflow in GA accident causal reconstructing, we conduct comprehensive experiments on the **AeroCausal** dataset. We primarily adopt grok-3-beta as the base LLM owing to its strong performance across diverse reasoning and code-generation benchmarks. To assess generalizability, we further benchmark against a range of advanced models, including GPT-4.1, GPT-3.5-turbo, DeepSeek-V3, and Qwen-3-235b. All models are accessed via API calls.

Our evaluation centers on edge-level performance, as accurate causal identification is essential for accident causal reconstruction, while node-level metrics serve as auxiliary indicators. We use standard evaluation metrics—Recall, Precision, and F1 Score—in macro and micro variants. Macro metrics aggregate results across the dataset, capturing the model's ability to detect causal relations. Micro metrics, computed per accident and averaged, assess consistency and robustness across cases. We emphasize recall, since missing causal edges (false negatives) can obscure key failure mechanisms and cause incomplete reconstructions. We also categorize recall performance into three groups: Perfect (Recall = 1), High ($0.8 \leq Recall < 1$), and Missed (Recall = 0), to analyze the model's ability to

capture causal edges.

### 4.4.1 Baselines

We benchmark DARF-ACR against six tailored baselines for GA accident causal reconstruction, grouped into (1) classic accident investigation frameworks and (2) unconstrained LLM strategies.

The classic frameworks include HFACS, Domino, and AcciMap, each imposing structured causal pathways. (1) HFACS restricts causation to a linear sequence from preconditions (unsafe human states or environment) to unsafe acts and finally to a single accident event, omitting mechanical factors; (2) Domino models accidents as a chain from personal fault to unsafe act or mechanical condition to a single event, neglecting environmental influences; (3) AcciMap adopts a layered structure spanning environment, mechanical issues, human unsafe acts/states, and accident events. It permits top-down, cross-layer, and intra-layer connections but prohibits bottom-up influences, which prevents it from capturing cross-event causal relations; at most, events can be linked to each other but not back to contributing factors.

Unconstrained LLM strategies rely on the intrinsic reasoning ability of LLMs without imposing predefined causal constraints, and allow flexible node and edge selection in causal graphs. We consider three variants: (1) Standard Prompt, which provides task descriptions and node categories for free-form graph generation; (2) Chain-of-Thought (CoT), which extends the Standard Prompt with step-by-step reasoning; and (3) Node-Edge Strategy, which employs a two-stage workflow by first identifying nodes and then inferring edges through sequential LLM API call.

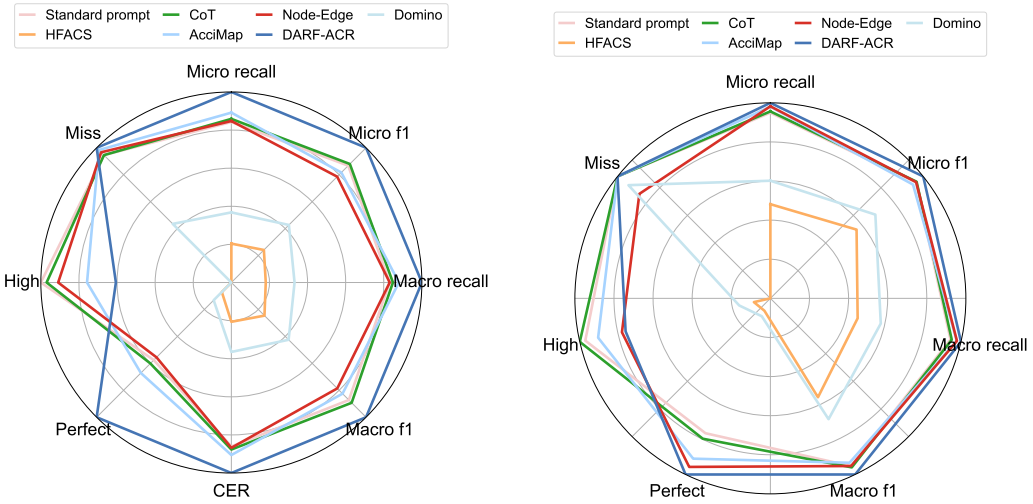
Further details are provided in Appendix C .

### 4.4.2 Overall performance

Figure 4.3 and Table 4.3 report the accident causal reconstruction performance of DARF-ACR and six baselines on grok-3-beta. DARF-ACR outperforms all baselines at the edge level, improving micro-recall by over 8% compared to AcciMap, with consistent gains in F1 and CER. The gap is smaller at the node level, where most LLM-driven strategies already exceed 85% recall, indicating that the main difficulty lies in reconstructing causal structures rather than identifying nodes. Classic frameworks such as HFACS and Domino collapse in this regard, with recalls below 30% and hundreds of missed cases, underscoring the limitations of rigid causal assumptions. AcciMap performs better through its layered representation, yet it remains confined to event-to-event links, leaving event-to-factor reasoning unaddressed. Unconstrained LLM strategies (Standard Prompt, CoT, Node-Edge) show moderate gains over the traditional frameworks, but their edge-level recall remains around 65%. It's interesting to fin that CoT in particular offers only marginal gains over simple prompting, indicating that CoT alone is inadequate for complex causal inference tasks. In contrast, DARF-ACR reconstructs 198 accidents perfectly and misses only 15, which highlights not only its higher accuracy but also its robustness. Overall, the superiority of DARF-ACR highlights a key insight: effective causal reconstruction requires balancing structure and flexibility by embedding appropriate domain knowledge into task-specific LLM reasoning. Classic frameworks fail under the multi-event, multi-factor complexity of GA accidents, while unconstrained LLMs yield unreliable results.

Table 4.3: Performance comparison across methods with highlight of max (bold) and second max (underlined) per column; inverted rule for Missed

| | Method | Micro recall | Micro f1 | Macro recall | Macro f1 | CER | Perfect | High | Missed |
|---|---|---|---|---|---|---|---|---|---|
| Edge-level | Standard prompt | 63.47 | 62.30 | 61.52 | 61.68 | 65.75 | 114 | **33** | 28 |
| | CoT | 64.69 | <u>63.18</u> | 62.82 | <u>63.15</u> | 66.97 | 119 | <u>32</u> | 28 |
| | Node-Edge | 63.85 | 56.42 | 61.57 | 55.60 | 66.28 | 110 | 30 | 23 |
| | Domino | 27.72 | 30.67 | 24.57 | 30.06 | 27.85 | 26 | 0 | 150 |
| | HFACS | 15.37 | 17.34 | 13.38 | 17.38 | 15.76 | 14 | 0 | 254 |
| | Accimap | <u>67.16</u> | 58.63 | <u>65.13</u> | 58.36 | <u>69.17</u> | <u>133</u> | 25 | <u>19</u> |
| | DARF-ACR(ours) | **75.33** | **71.82** | **74.07** | **70.72** | **76.39** | **198** | 20 | **15** |
| Node-level | Standard prompt | 86.42 | 84.77 | 85.68 | 84.93 | - | 196 | <u>101</u> | **0** |
| | CoT | 87.04 | <u>85.20</u> | 86.28 | <u>85.55</u> | - | 204 | **104** | **0** |
| | Node-Edge | <u>89.40</u> | 84.93 | <u>88.74</u> | 84.86 | - | <u>245</u> | 81 | 2 |
| | Domino | 54.75 | 61.17 | 52.51 | 61.15 | - | 26 | 17 | <u>1</u> |
| | HFACS | 43.84 | 50.22 | 41.51 | 50.05 | - | 18 | 9 | 14 |
| | Accimap | 89.23 | 83.11 | 88.64 | 83.21 | - | 233 | 94 | **0** |
| | DARF-ACR(ours) | **91.01** | **89.05** | **90.65** | **89.15** | - | **256** | 79 | **0** |



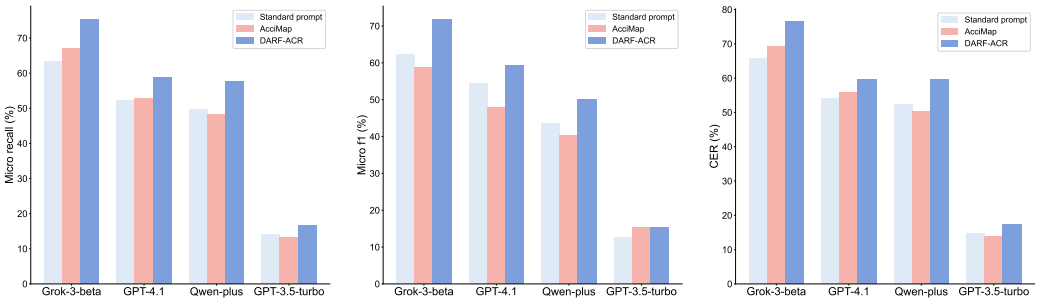

(a) Edge-level performance

(b) Node-level Performance

Figure 4.3: Radar charts of accident causal reconstruction using grok-3-beta

### 4.4.3 Performance across LLMs

To assess the generalizability of our approach, we compare DARF-ACR with two representative baselines (Standard Prompt and AcciMap) across four LLMs: Grok-3-beta, GPT-4o, Qwen-Plus, and GPT-3.5-turbo, on edge-level Micro recall, Micro f1, and CER. As shown in Figure 4.4, DARF-ACR consistently achieves the best results. It outperforms both baselines on the strongest model (Grok-3-beta), further extends the margin on mid-tier models (GPT-4o and Qwen-Plus), and still improves Recall and CER even on the weakest (GPT-3.5-turbo). These findings highlight the robustness of DARF-ACR, which serves as a performance amplifier for strong LLMs and a stabilizer for weaker ones, underscoring the importance of structured domain guidance for reliable causal reconstruction.



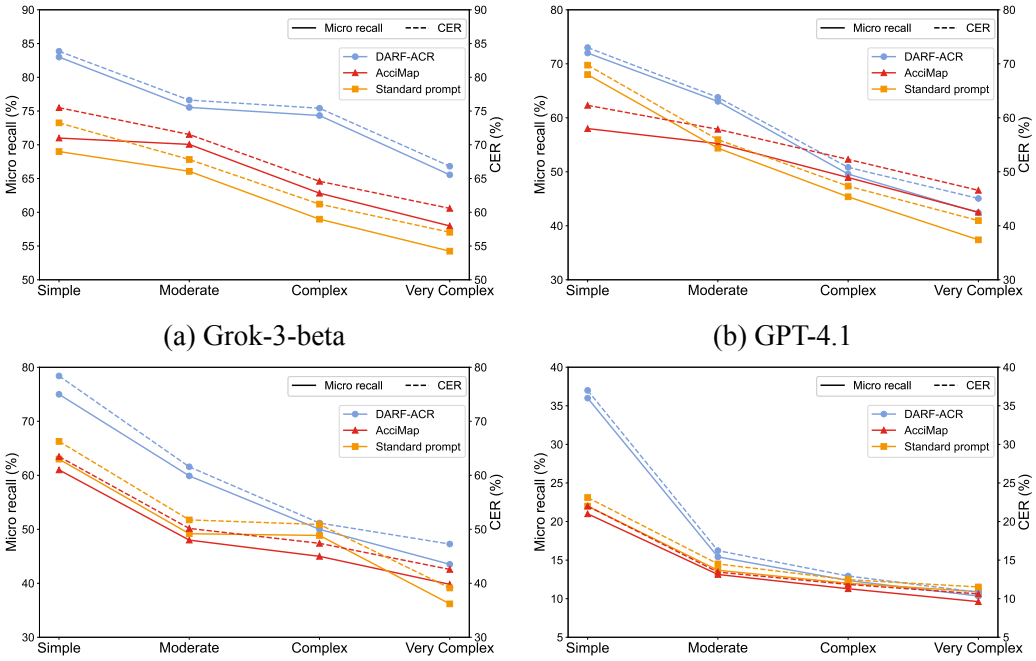

(a) Micro recall (b) Micro f1 (c) CER

Figure 4.4: DARF-ACR vs. baselines (Standard Prompt and AcciMap) across four LLMs

### 4.4.4 Performance comparison across complexity

We examine how task complexity, defined by the number of edges in the ground-truth causal graphs, affects model performance. As shown in Figure 4.5, results across four LLMs and two metrics (Micro Recall and CER) reveal two consistent

patterns. First, performance degrades as causal chains become more complex, reflecting the inherent difficulty of reconstructing long and entangled accident sequences. Nevertheless, DARF-ACR consistently outperforms both baselines (AcciMap and Standard Prompt) across all models and complexity levels. The advantage is most pronounced in simple scenarios, where DARF-ACR achieves an average gain of 13.75% in Micro Recall across models, while the margin narrows under very Complex settings. Stronger models such as grok-3-beta amplify the benefit of DARF-ACR, whereas weaker models such as gpt-3.5-turbo exhibit uniformly low performance (10.34%–36.00%), underscoring that advanced LLMs are better equipped for complex causal reasoning.




(a) Grok-3-beta (b) GPT-4.1

(c) Qwen-plus (d) GPT-3.5-turbo

Figure 4.5: Performance comparison of Micro recall and CER across LLMs under different causal complexity levels

In addition, across all models CER is higher than Micro recall because of its

Table 4.4: Cer–Micro recall differences across models and complexity levels

| Model | Mode | Simple | Moderate | Complex | Very Complex |
|---|---|---|---|---|---|
| Grok-3-beta | DARF-ACR | 0.86 | 1.08 | 1.08 | 1.27 |
| | AcciMap | 4.50 | 1.47 | 1.76 | 2.60 |
| | Standard prompt | 4.25 | 1.75 | 2.23 | 2.80 |
| GPT-4.1 | DARF-ACR | 1.00 | 0.74 | 1.24 | 2.64 |
| | AcciMap | 4.31 | 2.64 | 3.36 | 4.05 |
| | Standard prompt | 1.75 | 1.58 | 1.98 | 3.56 |
| Qwen-plus | DARF-ACR | 3.40 | 1.68 | 1.17 | 3.75 |
| | AcciMap | 2.47 | 2.16 | 2.41 | 2.77 |
| | Standard prompt | 3.28 | 2.56 | 2.03 | 2.94 |
| GPT-3.5-turbo | DARF-ACR | 1.00 | 0.81 | 0.58 | 0.49 |
| | AcciMap | 1.00 | 0.31 | 0.56 | 1.00 |
| | Standard prompt | 1.11 | 0.82 | 0.46 | 0.61 |

relaxed criteria that reward indirect causal paths. Table 4.4 shows that within Grok-3-beta, DARF-ACR achieves the smallest gap (0.86–1.27), indicating close alignment between recovered edges, plausible paths, and the ground truth. In contrast, AcciMap and Standard Prompt produce larger gaps (up to 4.5), reflecting reliance on indirect links. For GPT-3.5-turbo, differences remain small across methods, consistent with its limited reasoning that is constrained to direct edges, while Qwen-3 tends to generate longer but less precise paths. Interestingly, a U-shaped pattern also appears: gaps peak in simple and very Complex cases, suggesting that models add extra nodes when the ground truth is either simplified or highly entangled. Overall, DARF-ACR shows better balance between relaxed and strict evaluation, yielding both coherent and accurate causal structures.

### 4.4.5 Conditional edge reliability

To evaluate the edge identification capability of our proposed DARF-ACR workflow, we compute edge recall per node recall (E/N) and causal edge recall per node recall (CE/N). This normalization controls for node accuracy, isolating the perfor-

mance of edge prediction once nodes are correctly identified. Table 4.5 reports results across models and prompts. DARF-ACR achieves the highest E/N, while Accimap and Standard prompts perform lower. Across settings, CE/N is consistently higher than E/N, showing that causal edges are more reliably captured than general connections. These results indicate that DARF-ACR improves edge consistency given correct nodes, providing a stronger basis for causal reconstruction in accident analysis.

Table 4.5: Edge-to-node recall (E/N) and causal edge-to-node recall (CE/N) across models and strategies

| Model | DARF-ACR | | AcciMap | | Standard prompt | |
|---|---|---|---|---|---|---|
| | E/N | CE/N | E/N | CE/N | E/N | CE/N |
| Grok-3-beta | 80.15% | 81.36% | 72.77% | 74.98% | 70.74% | 73.26% |
| GPT-4.1 | 71.92% | 73.17% | 59.95% | 63.57% | 63.66% | 65.63% |
| Qwen-plus | 67.26% | 69.42% | 56.28% | 58.94% | 62.44% | 65.53% |
| GPT-3.5-turbo | 26.88% | 28.11% | 24.26% | 25.26% | 25.92% | 27.23% |

### 4.4.6 Qualitative analysis

We present the qualitative analysis on a multi-event GA accident to illustrate the DARF-ACR workflow with grok-3-beta. The accident evidence includes adverse wind, inadequate control inputs, a misaligned touchdown, and subsequent ground run-off with wing strike. The ground-truth causal chain is EF101 $\rightarrow$ LOC-I $\rightarrow$ HF104 $\rightarrow$ LOC-G $\rightarrow$ RE $\rightarrow$ CTOL, with an additional link HF113 $\rightarrow$ HF104. In essence, weather hazards trigger in-flight loss of control, training deficiencies contribute to handling errors, and the sequence escalates to ground loss of control, runway excursion, and collision.

DARF-ACR reconstructs this structure with high fidelity. It recovers the correct nodes and correct edges, including EF101$\rightarrow$LOC-I,LOC-I$\rightarrow$HF104,HF104$\rightarrow$LOC-

G,LOC-G→RE,and RE→CTOL. The results show that the workflow captures multi-event progression, transforming unstructured evidence into coherent causal graphs with accurate edge coverage.

**Ground truth:** [(EF101, LOC-I), (LOC-I, HF104), (HF113, HF104), (HF104, LOC-G),(LOC-G, RE), (RE, CTOL)]

**Accident evidence**: *The pilot checked the weather at his planned destinationairport during the return leg of a cross-countyflight, The wind speed was higherthan his personal minimums, so he diverted to a nearby airport with reported calmwind. He entered the pattern for runway ll and noticed that the wind was from thesouth. Prior to touchdown, the wind decreased and the airplane's nose tracked tothe lefi. The pilot added power for a go-around but did not compensate withenough right rudder which aggravated the airplane's misalignment. The airplanetouched down and exited the side ofthe runway, The airplane's lefi wing struck ataxiway sign resulting in substantial damage. The pilot reported no mechanicalmalfunctions with the airplane that could have contributed to the accident.*

**DARF-ACR output**:

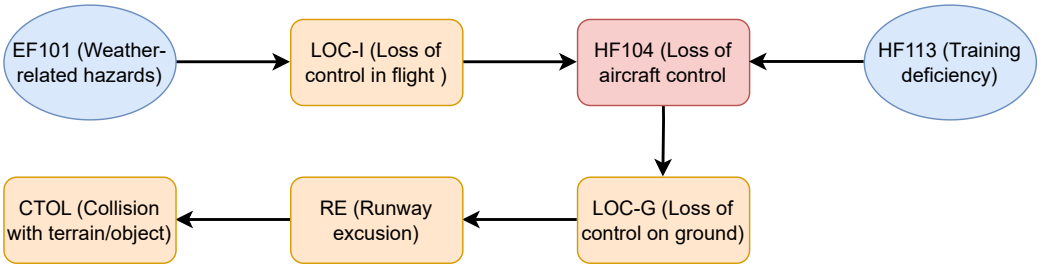



Figure 4.6: Case study: multi-event GA accident with DARF-ACR causal graph from accident evidence.

## 4.5 Concluding remarks

We proposed the **Dynamic Aviation Risk Field (DARF)** framework and the DARF-ACR workflow, a field-theory–driven approach to causal reconstruction in general aviation accident investigation. DARF provides an elegant ontology that unifies human, mechanical, and environmental risk fields with unsafe acts and accident events, enabling LLMs to generate interpretable and domain-consistent causal graphs. Together with the AeroCausal dataset and the causal edge recall (CER) metric, our results demonstrate that DARF-ACR consistently outperforms classi-

cal frameworks and unconstrained prompting across multiple models and levels of causal complexity. These findings establish DARF as a principled foundation for applying LLMs to safety-critical causal reasoning, with implications beyond aviation to domains such as autonomous driving and healthcare. In summary, our contributions are as follows:

1. We propose the **Dynamic Aviation Risk Field (DARF) Framework**, a field-theory–inspired model that represents human, mechanical, and environmental hazards as interacting causal factors, providing structured guidance for LLM-based accident investigation.

2. We design the **DARF-guided Accident Causation Reconstruction (DARF-ACR)** workflow, an event-centric, three-stage process that identifies event-level nodes, applies phase- and narrative-aware deduplication, and establishes causal links within and across events.

3. We construct **AeroCausal**, a curated GA accident causation DAG dataset derived from NTSB narratives, enabling reproducible evaluation of causation reconstruction methods.

4. We introduce a **bridge-tolerant evaluation metric** that credits legitimate intermediate expansions while penalizing causal errors, addressing the shortcomings of conventional edge- and path-recall measures for LLM-generated graphs.

Our study also reveals several limitations. Firstly, DARF-ACR currently focuses on individual-level GA accidents and does not incorporate organizational factors, limiting its coverage in more complex operational contexts. Secondly,

the LLMs used were not fine-tuned on aviation-specific knowledge, which constrains their reasoning fidelity. Thirdly, our analysis emphasizes explicit causal chains while underrepresenting latent but correlated risk factors such as pilot demographics or aircraft configurations. Addressing these limitations opens multiple research directions: extending DARF with organizational layers, enriching LLMs via domain-specific training or retrieval augmentation, and integrating statistical modeling to capture implicit risk indicators.

# Chapter 5

# Study 3: Exogenous factor for post investigation

This chapter presents a novel framework to enhance post-accident investigation in GA by addressing the challenge of incorporating exogenous risk factors into causal reconstructions. The proposed approach integrates econometric modeling with LLMs to identify statistically significant risk factors, construct a structured knowledge base, and augment causal graphs for comprehensive accident analysis. This chapter is organized as follows: Section 5.1 introduces the background and identifies research gaps in GA accident investigation. Section 5.2 describes the dataset used for risk factor analysis. Section 5.3 details the econometric modeling–LLM integrated framework, including risk factor identification, knowledge base construction, and causal graph augmentation. Section 5.4 presents the results of risk factor identification. Section 5.5 provides a case study validating the proposed framework. Finally, Section 5.6 offers concluding remarks, discussing implications, limitations, and future directions.

## 5.1 Background

In the Study 3 and Study 4, we employed LLMs for GA accident causal reconstruction, generating causal graph that explain how a specific accident unfolded. However, these reconstructions mainly emphasize explicit causes and often overlook exogenous risk factors such as pilot demographics, aircraft type, operational mode, and environmental context. Although not direct triggers, these exogenous factors statistically increase accident likelihood. To address this limitation, this study aims to augment the causal graphs generated in 4 with identified risk factors derived from historical data modeling. Specifically, we extract factors that exhibit statistically significant associations with accident severity to construct a GA risk knowledge base, which the LLM can retrieve and integrate into the causal graphs. By embedding these factors, the enhanced reconstructions are expected to explain not only *why the accident occurred* but also *why it was more likely under certain conditions*, thus providing more comprehensive insights for safety management and policy design.

Accurately identifying risk factors is a prerequisite for constructing the risk knowledge base and subsequently augmenting causal graphs. Characterized by low-altitude flying, diverse flight missions, and uncontrolled airspace, risk factors for GA safety may be unique. Numerous studies have focused on these determinants, consistently identifying human factors as the most significant [11, 15, 16, 22]. Environmental and technological factors, such as lighting, terrain, wind, and electronic flight displays, also play an important role in shaping GA accident outcomes [28, 43, 105].

While existing studies provide a strong foundation for understanding risk fac-

tors in GA, they often treat these accidents as a homogeneous group and overlook phase-specific variations. As shown in Fig. 5.1, GA accidents in the U.S. between 2008 and 2022 exhibit significant differences between flight phases. For example, nearly one-third of accidents occurred during the landing phase, with the lowest fatalities and aircraft destruction rates. In contrast, the maneuvering phase, though less frequent, has the highest fatality and destruction rates, making it the most hazardous. The standing phase, although experiencing the fewest accidents, has a notably high fatality rate. Accidents in other phases are less frequent and generally result in fewer fatalities. These distinctions suggest that risk factors vary by flight phase, and treating them uniformly can obscure important differences in accident causation.

In the work of [14], flight phases were treated as independent variables and found to be significantly associated with accident severity. Some studies have focused on risk factors in specific GA flight phases, especially takeoff and landing. [39] reported that fire, explosions, and the failure to use both lap belts and shoulder harnesses significantly increased the risk of pilot fatalities in crash landings. Similarly, [106] investigated landing airspeed mismanagement and concluded that high-speed landing accidents were more likely to result in severe injuries. During the go-around phase, [27] identified fatalities as most frequent under instrument meteorological conditions, with turbine engines, and among pilots with over 500 flight hours. For takeoff accidents, higher fatality risks have been linked to instrument meteorological conditions, twin-engine aircraft, and spring or summer operations, while no association was found with pilots' personality traits or demographics [40, 107]. We can find that most of the existing studies focus on individual flight phases. This gap limits the accuracy and granularity of GA risk analysis

and constrains the development of a more precise risk knowledge base.

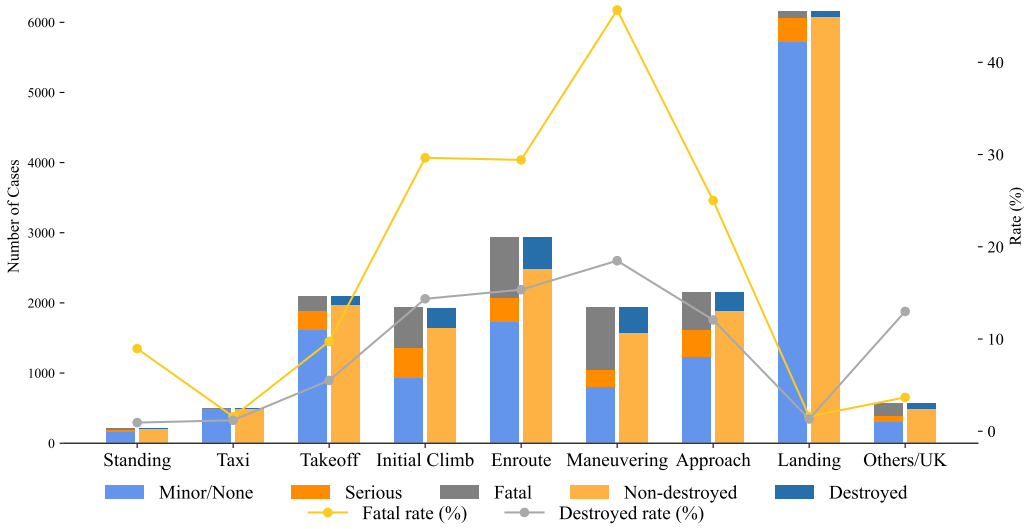



Figure 5.1: Pilot injury severity and aircraft damage across flight phases in U.S. GA accidents:2008-2022 (UK means unknown).

Besides phase transferability, three further considerations are critical for accurate risk factor identification. The first is the temporal instability of the factors' effects. At the macro level, shifts in economic conditions (e.g., financial crises), policy environments, and technological advances can alter accident characteristics, while at the micro level, pilot cognition, decision-making, and information processing also evolve accordingly [108]. Without distinguishing temporal phases, pooling data across many years risks treating time-sensitive factors as random effects, which reduces accuracy. Moreover, factors significant in specific periods may also be diluted or rendered insignificant in aggregate analyses. Therefore, we adopt a conservative principle of "inclusion over omission." That is, any factor that is significant within a given temporal window is retained in the knowledge base.

Secondly, GA accidents often result in two distinct outcomes simultaneously: personnel injuries (typically pilot injuries) and aircraft damage, which may have correlated disturbances because of the possible shared unobserved variations [109]. The Fig.5.1 also shows that the trends in the fatality rate and destroyed rate are similar over time. Modeling the two outcomes separately with single-equation probit models ignores their correlation, leading to efficiency loss and potential bias in coefficient estimates. Using one outcome as a predictor of the other introduces endogeneity issues [110], while collapsing them into a single indicator, such as accident severity, oversimplifies their relationship and masks the heterogeneous effects of determinants. In fact, several studies have attempted to jointly model dual accident outcomes in road traffic research, including the injury severity of leading and following drivers in rear-end collisions [111], injury severity of both drivers involved in truck-car crashes [112], as well as crash types and injury severity in highway ramp areas [113], thus providing more accurate and robust results.

The third issue to address is the unobserved heterogeneity in GA accident models. The unobserved heterogeneity often arises from variables that are difficult for onsite investigators to record but significantly impact model outcomes (e.g., pilot heart rate, blood oxygen levels, and cognitive state during the accident). Such heterogeneity can cause variations in the coefficients of the same influencing factors across observations, potentially biasing results if not properly addressed [114, 115].

Considering phase transferability, temporal validity, joint outcome effects and unobserved heterogeneity, we employ random parameter bivariate probit models with heterogeneity in means to improve the accuracy of risk factor identification. These models not only enable the simultaneous analysis of pilot injury severity

and aircraft damage but also represent one of the most advanced approaches for capturing heterogeneity [116].

After identifying significant risk factors using econometric modeling, we construct a structured GA risk knowledge base. Since accident evidence are largely unstructured narratives that often contain synonyms, abbreviations, negations, and implicit conditions, LLMs are thus employed for their natural language processing strengths in semantic understanding and contextual reasoning. Specifically, they are used to retrieve and align relevant evidence from accident evidences with the risk knowledge base and with the causal graphs generated in Study 4.

## 5.2 Data description

The accident data used in this study was sourced from the NTSB, which has provided publicly accessible records of general aviation accidents in the U.S. since 1982. Due to differences in the accident coding system before and after 2008 as well as to avoid potential confounding factors introduced by the COVID-19 pandemic, this research focused on accidents that occurred between January 2008 and December 2019. The dataset spans 12 years and is divided into three periods to address the challenge of small annual sample sizes, but more importantly, to account for significant economic and policy shifts that have impacted the general aviation industry: 2008-2011, 2012-2015, and 2016-2019. For example, 2008–2011 was marked by a slow recovery from the financial crisis and a significant decline in general aviation aircraft shipments [117]; 2012-2015 saw economic recovery accompanied by favorable policies such as the FAA Modernization and Reform Act of 2012 [118], Integration of Civil Unmanned Aircraft Systems (UAS) [119] and

the Small Airplane Revitalization Act [120]; 2016-2019 featured strong economic performance, ongoing diversification in general aviation, and a gradual recovery in the number of pilots starting in 2016 [121].

Moreover, it is important to note that in this research, we focused on flight phases with significant accident frequencies or severe outcomes to streamline the analysis while maintaining representativeness. These phases were categorized into four stages based on their relationship to ground and air operations (see Fig. 5.2): departure (from ground to air, including standing, taxi, takeoff, and initial climb), enroute (in-air operations), maneuvering (in-air operations) [1], and arrival (from air to ground, including approach and landing). Consequently, a series of random parameter bivariate probit models with heterogeneity in means was employed to explore the dynamic changes in general aviation accident risks across four phases and over time.

The raw data includes multiple related tables, such as aircraft, engine, event, flight crew, and findings, all of which were linked using the "ev-id" identifier. Accidents involving multiple aircraft, multiple pilots, or passengers were excluded to reduce the influence of extraneous variables. Additionally, incomplete records, such as those lacking information on injury severity, aircraft damage, or pilot demographics were removed, resulting in a filtered dataset of 11,749 single-aircraft, single-pilot general aviation accidents for analysis.

In the filtered dataset, injury severity was categorized into four levels following NTSB regulations [5]: fatal (17.29%, referring to injuries resulting in death

[1]The maneuvering phase, unique to general aviation, involves complex maneuvers like significant course and altitude adjustments, sharp turns, and aerobatics. Despite both enroute and maneuvering phases occurring in the air, their distinct operational characteristics warrant separate analysis in this study.

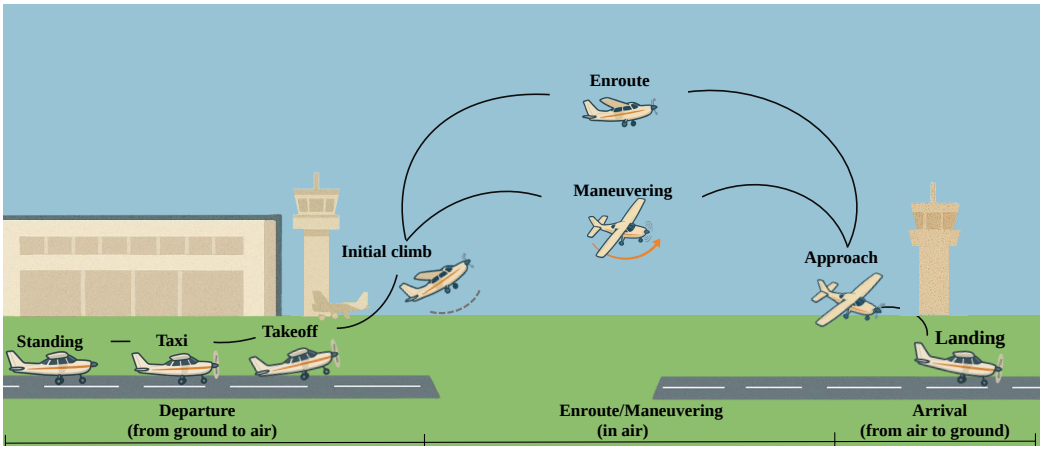



Figure 5.2: Flight phase grouping in GA by ground-air relationships.

within 30 days of the accident), severe (11.73%, representing injuries that require hospitalization for more than 48 hours), minor (15.64%, encompassing less severe injuries like superficial abrasions or lacerations), and none (55.35%, where no injury was sustained). Similarly, aircraft damage was divided into four levels: destroyed (7.8%, indicating irreparable damage or when repair costs exceed 50% of the aircraft's value, or if the aircraft is missing or inaccessible), substantial damage (90.32%, involving damage that compromises the aircraft's structural integrity or performance and requires significant repairs), minor (1.17%, representing damage that does not affect the aircraft's structural integrity), and none (0.70%, where no damage occurred). To address potential inaccuracies from imbalanced proportions, fatal and severe injuries were combined into a single category, while minor and no injuries were grouped as minor or none. Similarly, aircraft damage was reclassified into a binary variable: non-destroyed (combining minor, none, and substantial damage) and destroyed. This reclassification enables a two-by-two bi-

Table 5.1: Pilot injury and aircraft damage across different flight phases and time periods.

| Flight phrase | Period | Pilot injury | | Aircraft damage | |
|---|---|---|---|---|---|
| | | Fatal or serious | Minor or none | Destroy | Non-Destroy |
| **Departure** | 2008-2011 | 376 (31.76%) | 808 (68.24%) | 53 (4.48%) | 1131 (95.52%) |
| | 2012-2015 | 318 (34.23%) | 611 (65.77%) | 76 (8.18%) | 853 (91.82%) |
| | 2016-2019 | 252 (31.86%) | 539 (68.14%) | 92 (11.63%) | 699 (88.37%) |
| **Enroute** | 2008-2011 | 315 (41.07%) | 452 (58.93%) | 64 (8.34%) | 703 (91.66%) |
| | 2012-2015 | 262 (39.82%) | 396 (60.18%) | 84 (12.77%) | 574 (87.23%) |
| | 2016-2019 | 235 (39.97%) | 353 (60.03%) | 120 (20.41%) | 468 (79.59%) |
| **Maneuvering** | 2008-2011 | 315 (62.50%) | 189 (37.50%) | 60 (11.90%) | 444 (88.10%) |
| | 2012-2015 | 254 (58.26%) | 182 (41.74%) | 80 (18.35%) | 356 (81.65%) |
| | 2016-2019 | 172 (56.77%) | 131 (43.23%) | 78 (25.74%) | 225 (74.26%) |
| **Arrival** | 2008-2011 | 327 (15.46%) | 1788 (84.54%) | 55 (2.60%) | 2060 (97.40%) |
| | 2012-2015 | 314 (18.58%) | 1376 (81.42%) | 66 (3.91%) | 1624 (96.09%) |
| | 2016-2019 | 237 (14.38%) | 1411 (85.62%) | 87 (5.28%) | 1561 (94.72%) |
| **Total** | 2008-2019 | 3377 (29.08%) | 8236 (70.92%) | 915 (7.88%) | 10698 (92.12%) |

nary outcome model, categorizing injury severity (fatal/severe vs. minor/none) and aircraft damage (destroyed vs. non-destroyed) finally. More details about pilot injury and aircraft damage across different flight phases and periods can be found in Table 6.3.

Based on the data and research objectives, the variables were grouped into four categories: pilot characteristics (e.g., demographics, medical certification, seat occupied, human errors[2]), aircraft characteristics (e.g., aircraft category, homebuilt status, number of engines, type of landing gear), environmental factors (e.g., light conditions, sky conditions, wind conditions, basic weather), and crash characteristics (e.g., flight plan filed, mission type). The distribution of significant variables across the four flight phases and three time periods is presented in Fig. 5.3.

[2]The human error categories (e.g., skill error, perception error) are re-coded from the original accident causes using the Human Factors Analysis and Classification System (HFACS). For example, 'Personnel issues-Action/decision-Action-Forgotten action/omission-Pilot - F' is re-coded as a 'Skill error' to eliminate redundancy and simplify the original overlapping codes. For a detailed description of the HFACS coding process, please refer to [100].

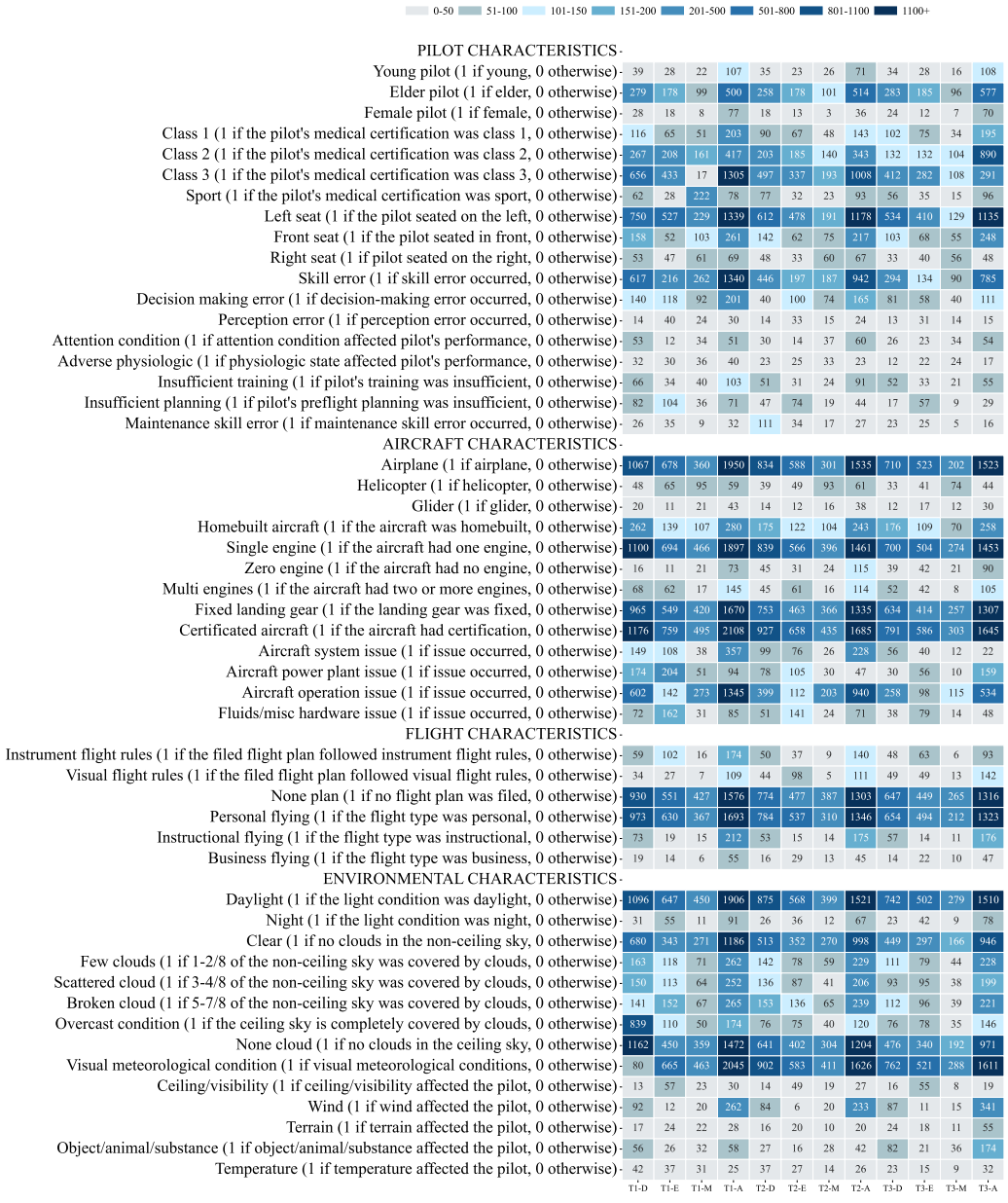



Figure 5.3: Distribution of significant variables across flight phases and time periods (D: departure, E: enroute, M: maneuvering, A: arrival) and periods (T1: 2008–2011, T2: 2012–2015, T3: 2016–2019).

## 5.3 Econometric modeling–LLM integrated framework

### 5.3.1 Risk factors identification

To accurately capture the determinants of GA accident outcomes, this study employs econometric modeling that jointly considers pilot injury severity and aircraft damage, both treated as binary variables. A random parameter bivariate probit approach with heterogeneity in means accounts for the inter-relationship between the two outcome variables while measuring the effects of influencing factors. Individual GA accident observations on $y_1$ and $y_2$ are available for all $i$. The general specification of a bivariate probit model can be specific as follows [122, 123]:

$$\begin{aligned} y_{i1}^* &= \beta_1' X_{i1} + \epsilon_{i1}, \quad y_{i1} = \begin{cases} 1, & \text{if } y_{i1}^* > 0 \\ 0, & \text{otherwise} \end{cases} \\ y_{i2}^* &= \beta_2' X_{i2} + \epsilon_{i2}, \quad y_{i2} = \begin{cases} 1, & \text{if } y_{i2}^* > 0 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{5.1}$$

where $y_{i1}$ is a binary indicator of pilot injury severity, and $y_{i2}$ is a binary variable indicating aircraft damage, while $y_{i1}^*$ and $y_{i2}^*$ denote the latent variables. $X$ is column vectors of explanatory variables, $\beta_1'$ and $\beta_2'$ represent the corresponding row vectors of estimated parameters, $\epsilon_{i1}$ and $\epsilon_{i2}$ are the error terms.

The joint probability $y_{i1} = 1, y_{i2} = 1$ can be written as:

$$\text{Prob}(y_{i1} = 1, y_{i2} = 1 \mid X_{i1}, X_{i2}) = \text{Prob}(y_{i1}^* > 0, y_{i2}^* > 0) \tag{5.2}$$

Consider that $\epsilon_{i1}$ and $\epsilon_{i2}$ are distributed as bivariate standard normal with a correlation coefficient $\rho$,

$$\begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right) \tag{5.3}$$

the joint probability density function of the bivariate probit model for $y_{i1} = 1, y_{i2} = 1$ could be expressed as [123]:

$$\begin{aligned} \text{Prob}(y_{i1} = 1, y_{i2} = 1 \mid X_{i1}, X_{i2}) &= \text{Prob}(\epsilon_{i1} < \beta_1' X_{i1}, \epsilon_{i2} < \beta_2' X_{i2}) \\ &= \Phi_2(\beta_1' X_{i1}, \beta_2' X_{i2}, \rho) \end{aligned} \tag{5.4}$$

Additionally, the other three joint probabilities are:

$$\begin{aligned} &\text{Prob}(y_{i1} = 1, y_{i2} = 0 \mid X_{i1}, X_{i2}) = \Phi_2(\beta_1' X_{i1}, -\beta_2' X_{i2}, -\rho) \\ &\text{Prob}(y_{i1} = 0, y_{i2} = 1 \mid X_{i1}, X_{i2}) = \Phi_2(-\beta_1' X_{i1}, \beta_2' X_{i2}, -\rho) \\ &\text{Prob}(y_{i1} = 0, y_{i2} = 0 \mid X_{i1}, X_{i2}) = \Phi_2(-\beta_1' X_{i1}, -\beta_2' X_{i2}, \rho) \end{aligned} \tag{5.5}$$

Thus, the log-likelihood function for this model is,

$$\ln L = \sum_{i=1}^{n} \ln \Phi_2(q_{y_{i1}} \beta_1' X_{i1}, q_{y_{i2}} \beta_2' X_{i2}, q_{y_{i1}} q_{y_{i2}} \rho) \tag{5.6}$$

where $q_{y_{i1}} = 2y_{i1} - 1$ and $q_{y_{i2}} = 2y_{i2} - 1$, and $\Phi_2(.,.,.)$ denotes the bivariate standard normal cumulative distribution function.

Moreover, random parameters with heterogeneity in means are incorporated into the model structure, allowing parameter estimates to vary across observations

to account for unobserved heterogeneity, expressed as [114]:

$$\beta_i = \beta + \Theta z_i + \Sigma_i^{1/2} v_i \tag{5.7}$$

Let $\Sigma_i^{1/2} = \text{Diag}[\sigma_{i1}, \sigma_{i2}, \ldots, \sigma_{ik}]$, then

$$\sigma_{ik} = \sigma_k \times \exp(\psi_k' w_i) \tag{5.8}$$

where $\beta$ represents the vector of mean coefficients for all observations as predefined, $\Theta$ donates the matrix of estimated parameters, $z_i$ is a vector of explanatory variables that capture heterogeneity in means of random parameters. The scale factor $\sigma_{ik}$, which provides the standard deviation of the $k$th random parameter, is placed on the diagonal of the diagonal matrix $\Sigma_i^{1/2}$. The $w_i$ is a vector of explanatory variables that capture heterogeneity in the variances, $\psi_k'$ is the $k_t h$ row elements of the matrix of estimated parameters $\Psi$, and $v_i$ is a primitive random vector, which follows the standard normal distribution.

The parameters were estimated using a simulation-based maximum likelihood approach with 1,000 Halton draws. The normal distribution was used to estimate random parameters, as prior studies consistently found no alternative distribution to be significantly superior [114, 124]. Additionally, marginal effects $ME$ were computed to provide further insights into the impact of a unit change in one accident outcome while holding the other outcome constant. This can be expressed as follows [111, 122, 125]:

$$
\begin{aligned}
ME_{y_{i1}} &= \text{Prob}(y_{i1}, \overline{y_{i2}} \mid \mathbf{X}_{i,1} = 1) - \text{Prob}(y_{i1}, \overline{y_{i2}} \mid \mathbf{X}_{i,1} = 0), \\
ME_{y_{i2}} &= \text{Prob}(\overline{y_{i1}}, y_{i2} \mid \mathbf{X}_{i,2} = 1) - \text{Prob}(\overline{y_{i1}}, y_{i2} \mid \mathbf{X}_{i,2} = 0).
\end{aligned} \tag{5.9}
$$

### 5.3.2 Risk knowledge base construction

After identifying significant risk factors, we systematize the results into a structured GA risk knowledge base to convert scattered statistical findings into machine–readable and searchable entries that provide reliable background support for LLMs in accident analysis. All factors are encoded with a uniform template to ensure standardization and semantic interpretability. Formally, each risk factor is represented as a tuple:

$$
R_i = \big(f_i,\ p_i,\ t_i,\ (d_i^{(1)}, d_i^{(2)}),\ (z_i^{(1)}, z_i^{(2)}),\ (m_i^{(1)}, m_i^{(2)}),\ fixed_i,\ hetero_i\big),
$$

where $f_i$ denotes the factor name, $p_i$ the corresponding flight phase, $t_i$ the temporal period during which the factor is statistically significant, $(d_i^{(1)}, d_i^{(2)})$ the directional effects on pilot fatal/serious injury and aircraft destruction respectively (each $\in \{-1, 0, 1\}$ but excluding $(0, 0)$), $(z_i^{(1)}, z_i^{(2)})$ the estimated $z$–statistics, $(m_i^{(1)}, m_i^{(2)})$ the corresponding marginal effects, $fixed_i \in \{0, 1\}$, where $0$ denotes a fixed parameter and $1$ a random parameter, and $hetero_i$ records the covariates influencing the random mean effect (if any, otherwise null). For retrieval, only $(p_i, (d_i^{(1)}, d_i^{(2)}))$ are used as the core matching keys, whereas the remaining elements provide auxiliary statistical and contextual information for interpretation.

All such tuples are collected into the GA risk knowledge base, denoted as $\mathcal{K}$, and each record is encoded in JSON format. As illustrative examples, Fig. 5.4 shows a fixed parameter (*Elderly pilot*) in (a) and a random parameter (*Helicopter*) in (b).


```
{
  "factor_name": "Elderly pilot",
  "flight_phase": "Enroute",
  "temporal_period": "2008-2011",
  "directions": { "injury": 1, "destroy": 0 },
  "z_stats": { "injury": 2.71, "destroy": null },
  "marginal_effects": { "injury": 0.05, "destroy": null },
  "fixed": 0,
  "hetero": null
}
```


(a) An example of fixed risk factor


```
{
  "factor_name": "Helicopter",
  "flight_phase": "Takeoff",
  "temporal_period": "2016-2019",
  "directions": { "injury": -1, "destroy": 0 },
  "z_stats": { "injury": -2.35, "destroy": null },
  "marginal_effects": { "injury": 0.08, "destroy": null },
  "fixed": 1,
  "hetero": [
    { "driver": "Homebuilt", "sign": "+", "outcome": "injury" }
  ]
}
```


(b) An example of random risk factor

Figure 5.4: Illustrative JSON entry of the GA risk knowledge base

Once entries are structured, the knowledge base supports efficient matching based on flight phase and accident outcomes, which lays the groundwork for LLM–based retrieval and subsequent causal graph augmentation.

### 5.3.3 LLM-based retrieval and causal graph augmentation

Given a target accident with narratives $\mathcal{N}_x$ and attributes $(\phi_x, (y_x^{(1)}, y_x^{(2)}))$—flight phase $\phi_x$, pilot fatal or serious injury $y_x^{(1)} \in \{0, 1\}$, and aircraft destruction $y_x^{(2)} \in \{0, 1\}$—together with an initial causal graph $G_0$ (from Study 4 ), we use an LLM to retrieve statistically validated risk factors from the GA risk knowledge base and augment the initial graph with contextual nodes.

Step 1 (structured filtering): We first filter the knowledge base $\mathcal{K}$ using the attributes $(\phi_x, (y_x^{(1)}, y_x^{(2)}))$. We map the outcomes to directions by $s(1) = +1$ and $s(0) = -1$. A factor $R_i = (f_i, p_i, t_i, (d_i^{(1)}, d_i^{(2)}), \dots)$ is retained if (i) its flight phase matches the target accident ($p_i = \phi_x$,); (ii) each effect $d_i^{(j)}$ is either the same

as the mapped outcome $s(y_x^{(j)})$ or zero; and (iii) at least one dimension matches exactly. This rule ensures that retrieved risk factors are always consistent with the observed accident outcome—they can support it or be neutral, but never contradict it. Formally,

$$\mathcal{C} = \Big\{ R_i \in \mathcal{K} \;\Big|\; p_i = \phi_x,\;\; d_i^{(1)} \in \{s(y_x^{(1)}), 0\},\;\; d_i^{(2)} \in \{s(y_x^{(2)}), 0\},\;\; (d_i^{(1)} = s(y_x^{(1)}) \vee d_i^{(2)} = s(y_x^{(2)})) \Big\}. \tag{5.10}$$

For example, if the observed outcome is $(y_x^{(1)}, y_x^{(2)}) = (1, 0)$, then $s(y_x^{(1)}) = +1$ and $s(y_x^{(2)}) = -1$, so admissible risk-factor directions are $(+1, -1)$, $(+1, 0)$, and $(0, -1)$.

This produces a candidate set $\mathcal{C}$ that is coherent with the phase and consequences evidenced in the accident.

Step 2 (narrative alignment): For each candidate $R_i = (f_i, \dots)$ from Step 1, we apply the LLM to check whether the factor name $f_i$ is explicitly mentioned in the accident narrative $\mathcal{N}_x$ and not negated. This yields the mention-filtered set $\tilde{\mathcal{C}}$:

$$\tilde{\mathcal{C}} = \text{LLM}_{\text{align}}(\mathcal{C}, \mathcal{N}_x). \tag{5.11}$$

where $\text{LLM}_{\text{align}}$ denotes the LLM-based alignment function that filters $\mathcal{C}$ according to textual evidence.

Step 3 (factor de-duplication): To avoid redundancy at the node level, we collapse directional candidates that share the same factor name into a single node representation (i.e., if a factor appears with multiple directional effects, it is counted once). In addition, in cases where both fixed and random versions of the same factor appear, we retain the fixed specification, assuming its influence on the accident

outcome is stable rather than heterogeneous. We then remove any factor already present in the initial accident graph $G_0$. Let $V(G_0)$ denote the node set of $G_0$. The de-duplicated candidate set is

$$\mathcal{C}^\star \;=\; \big\{\, f_i \;:\; R_i \in \tilde{\mathcal{C}},\; f_i \notin V(G_0) \,\big\}. \tag{5.12}$$

Thus, $\mathcal{C}^\star$ contains only *unique factor names* consistent with the evidence and not yet represented in $G_0$.

Step 4 (Graph augmentation): In this step, the LLMs takes as input the de-duplicated candidate set $\mathcal{C}^\star$, the initial accident graph $G_0$, and the accident narrative $\mathcal{N}_x$. It outputs the augmented graph $G_{\text{augment}}$ by adding each factor $f_i \in \mathcal{C}^\star$ as a dashed-box node and attaching it to an anchor node $v^\star \in V(G_0)$, with fixed and random parameters are distinguished by different colors. If an edge would introduce a cycle, the model omits or adjusts it to preserve acyclicity. Formally:

$$G_{\text{augment}} \;=\; \text{LLM}_{\text{augment}}(G_0, \mathcal{C}^\star, \mathcal{N}_x). \tag{5.13}$$

The resulting $G_{\text{augment}}$ incorporates new contextual risk factors explicitly supported by the evidence, while also clearly showing how they are integrated into the existing causal structure.

## 5.4 Risk factor identification results

### 5.4.1 Flight phase transferability and temporal stability tests

Before presenting the formal model results, likelihood ratio tests were conducted to assess the necessity of flight phase or time period division, as expressed below [126]:

$$\chi^2 = -2\left[LL\left(\beta_{mn}\right) - LL\left(\beta_m\right)\right] \tag{5.14}$$

where $LL\left(\beta_{mn}\right)$ refers to the log-likelihood at convergence of the model with parameters from $n$ on data from $m$ ( where $m \neq n$.), while $LL\left(\beta_m\right)$ denotes the log-likelihood at convergence of the model with both parameters and data from $m$. The value of $\chi^2$ follows a $\chi^2$ distribution, with degrees of freedom equal to the number of estimated parameters in the model $\beta_{mn}$. This test is two-sided by considering the transformation of $nm$ and $n$.

Table 5.2 presents the likelihood ratio test results for assessing the phase transferability of determinants in individual time periods. Notably, applying parameters from arrival (2016–2019) to enroute (2016–2019) or maneuvering (2016–2019) data yielded low $\chi^2$ values, whereas the reverse yielded high confidence in rejecting the null hypothesis of parameter equality between these phases. This suggests that, despite some shared parameters, the arrival model includes additional significant factors, highlighting phase non-transferability. Overall, in 21 of 24 models, the null hypothesis of uniform parameters across flight phases was rejected at 95% confidence, strongly supporting distinct risk factor variations across phases.

Table 5.2: Likelihood ratio test results for GA accidents across different flight phases

| Subgroup | m | n | | | |
|---|---|---|---|---|---|
| | | **Departure** | **Enroute** | **Maneuvering** | **Arrival** |
| **2008-2011** | Departure | - | 82.42(32)[>99.99%] | 177.69(21)[>99.99%] | 83.23(32)[>99.99%] |
| | Enroute | 88.97(29)[>99.99%] | - | 139.49(21)[>99.99%] | 135.18(32)[>99.99%] |
| | Maneuvering | 62.81(29)[99.97%] | 77.96(32)[>99.99%] | - | 51.27(32)[98.32%] |
| | Arrival | 136.85(29)[>99.99%] | 196.35(32)[>99.99%] | 276.85(21)[>99.99%] | - |
| **2012-2015** | Departure | - | 117.61(23)[>99.99%] | 124.81(23)[>99.99%] | 77.19(32)[>99.99%] |
| | Enroute | 66.37(30)[>99.99%] | - | 84.59(23)[>99.99%] | 66.48(32)[99.97%] |
| | Maneuvering | 57.78(30)[99.83%] | 71.34(23)[>99.99%] | - | 60.59(32)[99.83%] |
| | Arrival | 151.11(30)[>99.99%] | 269.02(23)[>99.99%] | 270.55(23)[>99.99%] | - |
| **2016-2019** | Departure | - | 98.38(19)[>99.99%] | 161.23(13)[>99.99%] | 50.18(34)[96.36%] |
| | Enroute | 104.52(20)[>99.99%] | - | 184.72(13)[>99.99%] | 38.27(34)[71.83%] |
| | Maneuvering | 25.23(20)[80.71%] | 33.13(19)[97.68%] | - | 9.66(34)[0.00%] |
| | Arrival | 167.92(20)[>99.99%] | 231.35(19)[>99.99%] | 499.84(13)[>99.99%] | - |

The degree of freedom is in round brackets, and confidence level is in square brackets.

Similarly, as Table 5.3 shows, 20 out of 24 models rejected the null hypothesis of parameter equality across periods with over 99% confidence, underscoring the temporal instability of factors influencing pilot injuries and aircraft damage in GA accidents.

Table 5.3: Likelihood ratio test results for GA accidents across different time periods

| Subgroup | m | n | | |
|---|---|---|---|---|
| | | **2008-2011** | **2012-2015** | **2016-2019** |
| **Departure** | 2008-2011 | - | 92.08(30)[>99.99%] | 171.27(20)[>99.99%] |
| | 2012-2015 | 93.59(29)[>99.99%] | - | 159.98(20)[>99.99%] |
| | 2016-2019 | 55.71(29) [99.80%] | 79.01(30)[>99.99%] | - |
| **Enroute** | 2008-2011 | - | 117.31(23)[>99.99%] | 154.41(19)[>99.99%] |
| | 2012-2015 | 39.3(32)[82.45%] | - | 91.86(19)[>99.99%] |
| | 2016-2019 | 6.96(32)[0.00%] | 67.72(23)[>99.99%] | - |
| **Maneuvering** | 2008-2011 | - | 46.44(23)[99.74%] | 84.34(13)[>99.99%] |
| | 2012-2015 | 74.43(21)[>99.99%] | - | 104.03(13)[>99.99%] |
| | 2016-2019 | 13.32 (21)[10.31%] | 2.96(23)[0.00%] | - |
| **Arrival** | 2008-2011 | - | 134.59(32)[>99.99%] | 160(34)[>99.99%] |
| | 2012-2015 | 150.27(32)[>99.99%] | - | 179.10(34)[>99.99%] |
| | 2016-2019 | 108.90(32) [>99.99%] | 141.01(32)[>99.99%] | - |

The degree of freedom is in round brackets, and confidence level is in square brackets.

### 5.4.2 Model results and discussion

Table 5.4 presents the estimation results for GA accidents occurring during four specific flight phases: departure, enroute, maneuvering, and arrival, using the random parameters bivariate probit model with heterogeneity in means. Notably, all 12 models showed a correlation coefficient $\rho$ greater than 0.500, ranging from 0.582 to 0.877, with a confidence level greater than 99.99%. This strong positive correlation suggests that unobserved factors jointly influence personnel injury and aircraft damage. All the models showed a good overall statistical fit, with $\rho^2$ values between 0.116 and 0.232, and corrected $\rho^2$ values ranging from 0.079 to 0.186. Furthermore, constant terms affecting injury severity and aircraft damage were identified in some models. Interestingly, the coefficients of constants for injury severity were always positive, while those for aircraft damage were negative. This contrasts with the positive correlation between the two outcomes, suggesting that although unobserved factors may influence both, their baseline probabilities differ. On the other hand, pilots may be more likely to sustain severe or fatal injuries than aircraft to sustain destruction in an accident. A more in-depth discussion of the determinants will be provided in the subsequent sections.

#### 5.4.2.1 The random parameters and heterogeneity in means

As represented in Table 5.4, fourteen variables resulted in significant random parameters for the four flight phase models. Eleven of these parameters were associated with pilot fatal or severe injuries and three with aircraft destruction. Fig. 5.5a -Fig. 5.5d showed the probability density diagrams of these random parameters.

Starting from the departure models, the indicators for homebuilt aircraft during

the periods of 2008–2011 and 2012–2015 were identified as significant random parameters. Both parameters followed a normal distribution: the former, with a mean of -1.991 and standard deviation of 2.269, suggests that 80.99% of homebuilt aircraft are less prone to destruction in accidents. In contrast, the latter indicates 76.02% of pilots faced higher risks of fatal or severe injuries during 2012–2015 (see Fig. 5.5a). In addition, helicopter increased the mean for homebuilt aircraft accidents occurred in 2008-2011, shifting the mean of this parameter distribution to the right, thereby increasing the pilot injury severity. In 2012–2015, pilot decision errors further elevated the mean, heightening the probability of severe or fatal outcomes.

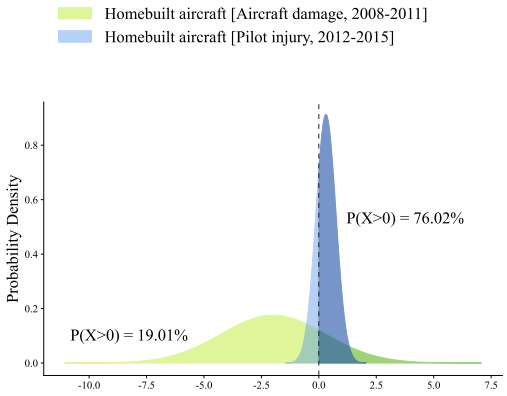
As for the enroute group models, four variables were observed to generate random parameters. The indicator for fluids/misc hardware issues during 2008-2011 was identified as a random parameter, suggesting that 75.96% of enroute accidents involving such issues are less likely to result in fatal or severe pilot injuries. When the flight plan followed instrument flight rules, the mean of this indicator increased. The home-built aircraft showed a random parameter during 2008-2011, with a decreasing mean when there were no clouds and an increasing mean when the pilot experienced a perception error. The fixed landing gear indicator generated a random parameter during 2012-2015. Lastly, the elder pilot indicator for 2012-2015 captured a random parameter, with most pilots suffering more severe injuries. The elder operator indicator is often considered a random parameter in accident analysis [115, 127, 128], reflecting substantial inter-individual variations due to declines in cognitive and physical capacities, along with changes in accumulated work experience as age increases.

For the maneuvering model, three variables resulted in significant random pa-

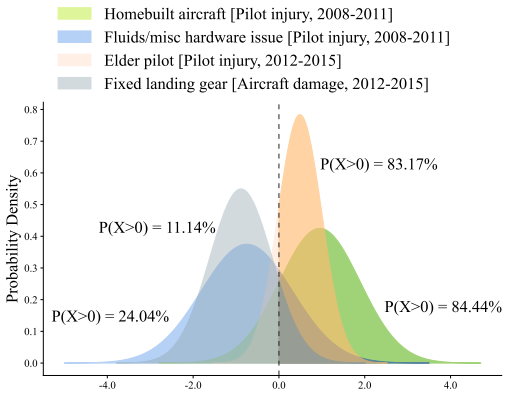
rameters. Notably, the right seat indicator was identified as a random parameter in 2008-2011, with most observations leading to fatal or severe injuries. Regarding the helicopter indicator in 2012-2015, its mean increased when the pilot was seated in the left seat or exhibited signs of inattention such as distraction, and decreased if the pilot held a class 3 medical certification. Moreover, the indicator for decision-making errors during the 2012-2015 period resulted in a random parameter, with 86.6% of the observations indicating an increased risk of severe injuries. This aligns with the view that aerobatics in GA are often considered decision errors.

In the arrival models, visual meteorological conditions and gliders were identified as significant random parameters during 2008-2011, with most observations indicating a decreased likelihood of more severe injuries. From 2012 to 2015, the adverse physiological conditions indicator captured a random parameter, with the mean decreasing if the pilot had attention issues. The aircraft system issues indicator was also a random parameter, with the mean increasing if the accident involved a helicopter. The glider indicator was captured as a random parameter for the 2015-2019 period, with its mean increased if the pilot was seated in the front seat. For the 2015-2019 period, the clear indicator produced a random parameter, while the risk of severe injuries under clear weather conditions decreased if there were no clouds in the ceiling sky.

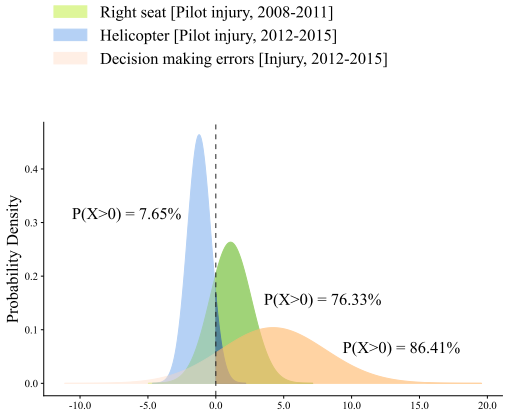
All these random parameters are recorded into the risk knowledge base as the random risk factors.



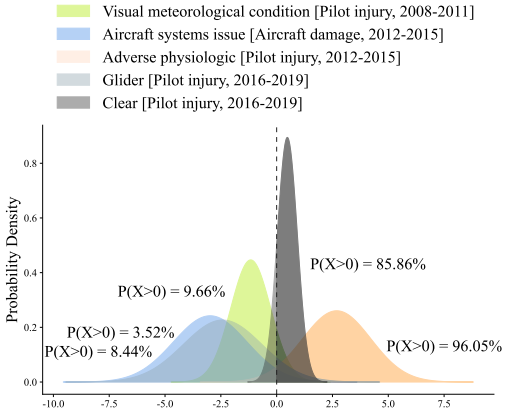

(a) Random parameters in the departure models




(b) Random parameters in the enroute models




(c) Random parameters in the maneuvering models




(d) Random parameters in the arrival models

Figure 5.5: Distribution of the random parameters.

#### 5.4.2.2 The pilot characteristics

Pilot characteristics significantly influencing injury severity and aircraft damage included demographic factors, medical classification level, seat occupied, unsafe acts, and preconditions of unsafe acts. As shown in Table 5.4, younger pilots had a lower risk of fatal or severe injuries during departure, enroute and arrival accidents for 2008–2011, a trend that persisted into the 2012–2015 period for maneuvering

accidents. In contrast, elderly pilots were more likely to sustain severe injuries and cause aircraft damage during arrival accidents in 2008–2011. Pilot age remains a critical concern in safety, particularly as the pilot population continues to age [129]. Unlike air carrier operations (e.g., Part 121), GA has no maximum age limit for pilots, raising concerns about its safety. Contradictory research findings regarding the impact of aging on flight safety persist [13–16], as declines in cognitive and physical abilities may be offset by the experience gained over time. While caution is needed, this study showed that overall, elderly pilots in GA face higher risks, highlighting the need for mandatory medical check-ups and safety training for elderly GA pilots. Additionally, female pilots were associated with decreased injury levels during arrival accidents for 2008–2011, which is consistent with [16] and [130].

In terms of pilot medical classifications, it is interesting to find that pilots with Class 1, Class 2, Class 3, and sports qualifications had a higher likelihood of causing aircraft damage compared to those with basic or no medical qualifications. In addition, overall, pilots seated in the left seat during accidents were less likely to sustain severe injuries or cause aircraft destruction in departure or enroute accidents, likely due to the greater experience of captains (who typically sit on the left) in managing emergencies. In contrast, pilots in the front or right seats had higher injury risks in maneuvering accidents during both 2008-2011 and 2016-2019, consistent with [131], but lower aircraft damage risks in maneuvering and arrival accidents in 2016-2019, likely due to improvements in training and safety technology.

Regarding the unsafe acts of pilots, pilot skill errors significantly increased the risk of fatal or severe injuries during departure, enroute, and arrival phases

in 2008-2011, with this trend continuing in 2012-2015, especially during departure and enroute accidents. Similarly, pilot decision and perception errors were consistently linked to increased injury severity across periods and flight phases. Overall, unsafe pilot behaviors, such as failing to follow procedures, misinterpreting instruments, and ignoring warnings, consistently contributed to injury severity, reflecting temporal stability and phase transferability. These behaviors have been widely recognized as leading factors in GA accidents [12, 24]. In addition to pilot errors, maintenance personnel errors also significantly impacted maneuvering accidents in 2008-2011 and departure accidents in 2012-2015.

In addition to unsafe acts, indirect human errors, particularly preconditions for unsafe acts, were significant determinants of accident outcomes. Like unsafe acts, preconditions consistently increased the risk of severe or fatal injuries across all flight phases and periods, with the strongest effect in the arrival phase, as shown in Table 6.4. Specifically, pilots experiencing attention issues had a higher risk of fatal or severe injuries during enroute and arrival accidents in 2012-2015. In the enroute phase, high automation levels may increase distractions and lapses in attention [132], while the complexity of tasks in the arrival phase could contribute to inattentional blindness [133, 134]. Pilots with suboptimal mental states or adverse physiological conditions significantly impacted accident outcomes, highlighting the need for strict medical standards and policies that encourage pilots to report such issues before flight. Additionally, consistent with [24], inadequate training and pre-flight preparation were linked to higher injury severity, especially in arrival accidents.

#### 5.4.2.3 The aircraft characteristics

The three aircraft types in the model (airplanes, helicopters, and gliders) reduced the risk of severe or fatal injuries compared to other GA types, such as gyroplanes, weight-shift aircraft, and powered-lift airplanes. Airplanes consistently reduced injury severity, especially during departure and arrival accidents, with the most notable marginal effects between 2012 and 2015 (see Table 6.4). Gliders showed lower risks of fatal or severe injuries in arrival accidents for 2012-2015, while contradictory results from [26] found higher injury risks for helicopters and gliders in Germany, likely due to differences in datasets and research focuses. For both 2008-2011 and 2012-2015, homebuilt aircraft were more likely to sustain severe damage and cause more severe injuries to pilots in departure and arrival accidents, consistent with [23]. This is unsurprising, as homebuilt aircraft, often built by enthusiasts without the commercial standards of certificated aircraft, are more prone to mechanical failures. Moreover, certificated aircraft were more likely to cause severe or fatal injuries in maneuvering accidents for 2008-2011, but showed a reduced risk in landing accidents for 2016-2019.

The number of aircraft engines is linked to flight error tolerance and operational complexity, but there is no simple linear relationship between engine count and GA safety (Table 5.4). Aircraft with no engines or multiple engines tend to result in more severe pilot injuries and greater aircraft damage. As [14] reported, while multiple engines may reduce the risk of total engine failure during cruise, a light twin-engine aircraft loses significant climb performance if one engine fails. The benefit of multi-engine aircraft is often offset by its more complex operations and harsher conditions [14, 106]. The higher risk for aircraft with no engine may

arise from their reliance on airflow, making them particularly vulnerable to control issues and accidents. Additionally, aircraft with fixed landing gear showed lower risks of severe injuries, especially in enroute accidents, and reduced damage in departure, maneuvering, and landing accidents during 2012-2015. This aligns with [14]: fixed landing gear, being simpler than retractable gear, requires no in-flight retraction, thus reducing the likelihood of errors.

Regarding aircraft-related issues affecting accidents, all factors, except for aircraft system issues, were almost positively correlated with more severe accident outcomes across various periods and flight phases. This may be due to system issues typically involving non-critical components, such as flight controls, landing gear, or autopilot systems, which have minimal impact on flight safety. Power plant issues were linked to higher injury severity in departure and arrival accidents, especially in 2008-2011 and 2016-2019, likely due to lower error tolerance during these phases. Operational issues, including power plant, ignition, and engine bleed air system problems, were positively correlated with severe outcomes, particularly in enroute and maneuvering accidents. However, fluids and miscellaneous hardware issues reduced the risk of aircraft destruction in enroute accidents during 2012-2015. Overall, these findings emphasize the need for careful attention to aircraft design, operation, and maintenance, especially for issues contributing to severe injuries and aircraft destruction.

#### 5.4.2.4 The flight characteristics

Significant flight characteristics included flight plan and mission type. It is noted that the probability of aircraft destruction increased if the filed flight plan followed instrument flight rules. Specifically, during the 2008-2011 period, instru-

ment flight rules increased the risk of destruction in both the enroute and arrival phases. In contrast, flights conducted under visual flight rules were associated with a lower chance of severe or fatal injuries in departure accidents for 2016-2019. It is not surprising because the instrument flight rule is typically used in low-visibility and adverse weather conditions, whereas the visual flight rule allows pilots to navigate visually in clear conditions, reducing both weather-related threats and reliance on instruments. Furthermore, the absence of a filed flight plan aggravated the aircraft destruction in departure accidents during the 2016-2019 period. No flight plan may indicate broader issues, such as inadequate preparation, limited resources, and experience gaps, which together increase accident risk.

Regarding flight type, both personal and business flying exhibited higher risks, associated with an increased likelihood of fatal or severe injuries during enroute and maneuvering accidents in the 2012-2015 period. This underscores the need for enhanced safety oversight in personal flying and improved flight planning in business aviation. Instructional flying was significant in both the departure and arrival models, with a decreased risk of severe or fatal injuries and aircraft destruction. Such a finding is similar to what others found [27], suggesting that while instructional flying often involves student pilots who may lack experience and knowledge, it typically occurs in a more controlled and supervised environment, thereby significantly reducing the severity of accidents.

#### 5.4.2.5 The environmental characteristics

Environmental attributes are critical factors in GA accidents. The sensitivity to these factors can vary across different flight phases due to differences in altitude, speed, and control requirements. Specifically, light conditions influence accidents

primarily through changes in visibility, pilot fatigue, pilot sightlines, and flight operations. Echoing the findings reported in previous studies [13, 14, 27, 41, 43], daylight conditions were found to provide a protective effect on GA safety, while night flights significantly increased the risk of aircraft destruction in different flight phrases and periods, except for departure accidents (2016-2019 ), likely due to improvements in aircraft design, airport lighting, and obstacle management. Similar to light conditions, cloud cover can also affect flight visibility to varying degrees. As shown in Table 6.1, for non-ceiling sky conditions, increased cloud cover generally correlates with a high likelihood of severe or fatal injuries or aircraft destruction. Interestingly, for ceiling sky conditions, both no cloud coverage and full cloud coverage (overcast) were significantly linked to increased pilot injuries and aircraft destruction, especially in arrival accidents. Although counterintuitive, in clear ceiling sky conditions, pilots may overestimate their abilities or be unprepared for weather changes, leading to accidents.

In line with the previous research [14, 27, 37, 40], our result found that the visual meteorological conditions indicator demonstrated strong temporal stability and transferability across flight phases, consistently increasing the likelihood of severe accidents. Notably, during the 2012-2015 period, this indicator elevated the risk of severe or fatal injuries in all phases except departure. It could be attributed to sample selectivity [135]. That is, compared to instrument meteorological conditions, visual meteorological conditions typically offer a clearer, high-visibility environment, which should theoretically reduce the likelihood of severe accidents.

Regarding environmental factors affecting pilot performance (as explicitly stated in accident investigation reports), most were positively associated with the risk of severe or fatal pilot injuries and aircraft destruction. Specifically, the ceiling/vis-

ibility indicator influenced both injury severity and aircraft destruction in enroute and arrival accidents during 2008-2011 and 2016-2019, and was also significant in departure and maneuvering models in certain periods. Wind factors (e.g., sudden shifts, wind shear, and downdrafts) were generally associated with less severe injuries and reduced aircraft damage. Objects, animals, or foreign substances increased the risk of severe injuries in departure accidents and aircraft destruction in arrival accidents during 2016-2019. Our results showed inconsistent effects of terrain and temperature on injuries and damage: terrain impacted survivability while causing greater aircraft damage, while temperature seemed to affect injury severity more than aircraft destruction.

Table 5.4: Model estimation results of fatal or severe injuries and destruction (with the z–statistics in round brackets)

| **Significant variables** | **Departure** | | **Enroute** | | **Maneuvering** | | **Arrival** | |
|---|---|---|---|---|---|---|---|---|
| | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** |
| Constant [2008-2011] | | -1.033 (-2.99) | | -1.310 (-6.92) | | -1.596 (-4.85) | | -1.824 (-9.75) |
| Constant [2012-2015] | 1.511 (3.61) | -1.428 (-9.39) | | | 1.029 (2.15) | | 1.794 (6.15) | |
| Constant [2016-2019] | | | | -1.382 (-6.35) | | | | -1.223 (-8.94) |
| ***Pilot characteristics*** | | | | | | | | |
| Young pilot [2008-2011] | -0.567 (-1.75) | | -0.746 (-2.17) | | | | -2.145 (-2.15) | |
| Young pilot [2012-2015] | | | | | -0.560 (-1.75) | | | |
| Elder pilot [2008-2011] | | | | | | | 0.387 (3.99) | 0.262 (1.75) |
| Elder pilot [2012-2015] | 0.209 (1.99) | -0.300 (-1.78) | | | | | | |
| Elder pilot [2016-2019] | | | | | | | 0.267 (3.11) | |
| Female pilot [2016-2019] | | | | | | | -0.589 (-2.54) | |
| Class 1 [2012-2015] | | 0.445 (2.75) | | | | | -0.685 (-3.88) | |
| Class 2 [2016-2019] | | | | 0.608 (3.21) | | 0.450 (2.49) | | |
| Class 3 [2016-2019] | | | | 0.396 (2.40) | | | | |
| Sport [2008-2011] | | | -0.690 (-1.97) | | | | | |
| Sport [2012-2015] | | | | | | | -0.613 (-3.30) | |
| Sport [2016-2019] | | | | | | | | 0.391 (2.18) |
| Left seat [2008-2011] | | | | -0.500 (-2.98) | | | | |
| Left seat [2012-2015] | -0.234 (-2.21) | | | -0.620 (-4.11) | | | | |
| Left seat [2016-2019] | -0.221 (-1.98) | | | | | | | |
| Front seat [2008-2011] | | | | | 0.420 (2.37) | | | |
| Front seat [2016-2019] | | | | | 0.355 (1.88) | | | -0.325 (-1.84) |

| Significant variables | Departure | | Enroute | | Maneuvering | | Arrival | |
|---|---|---|---|---|---|---|---|---|
| | Pilot | Aircraft | Pilot | Aircraft | Pilot | Aircraft | Pilot | Aircraft |
| Right seat [2016-2019] | | | | | 0.487 (1.82) | -0.449 (-1.81) | | |
| Skill errors [2008-2011] | 0.357 (3.91) | | 0.336 (2.58) | | | | 0.356 (3.78) | |
| Skill errors [2012-2015] | 0.217 (2.32) | | 0.297 (2.32) | | | | | |
| Skill errors [2016-2019] | | | 0.568 (3.91) | 0.320 (1.90) | 0.441 (2.70) | | | |
| Decision making errors [2008-2011] | 0.397 (3.11) | | 0.733 (4.11) | | | | | |
| Decision making errors [2012-2015] | 0.437 (2.72) | 0.457 (2.50) | 0.595 (3.76) | | | | | |
| Decision making errors [2016-2019] | | | | | | | 0.505 (3.49) | |
| Perception errors [2008-2011] | 1.638 (2.15) | 1.382 (3.11) | 1.390 (2.58) | 0.709 (2.77) | | | 1.446 (4.46) | |
| Perception errors [2012-2015] | | | 1.043 (2.63) | 1.389 (3.93) | | | | |
| Perception errors [2016-2019] | | | 0.933 (1.75) | 1.089 (3.24) | | | | |
| Maintenance skill errors [2008-2011] | | | | | -1.157 (-2.17) | | | |
| Maintenance skill errors [2012-2015] | 0.622 (2.70) | | | | | | | |
| Attention condition [2008-2011] | -0.585 (-2.08) | | | | | | | |
| Attention condition [2012-2015] | | | 0.963 (2.52) | | | | 0.965 (5.24) | 0.748 (3.43) |
| Adverse physiologic [2008-2011] | 1.971 (5.06) | | 1.287 (2.86) | | 2.413 (2.65) | 0.661 (2.91) | 2.063 (7.66) | 1.004 (3.63) |
| Adverse physiologic [2012-2015] | 2.399 (4.09) | 1.004 (3.43) | 1.890 (4.22) | | 3.041 (3.34) | | | 1.302 (3.49) |
| Adverse physiologic [2016-2019] | | | 1.790 (3.11) | 0.782 (2.59) | 1.036 (2.18) | | 1.375 (3.61) | 0.863 (2.22) |
| Inadequate training [2008-2011] | 0.486 (2.60) | | | | | | 1.006 (5.57) | 0.489 (2.13) |
| Inadequate training [2012-2015] | 0.748 (3.53) | 0.757 (3.43) | | | 0.883 (1.90) | | 0.644 (4.17) | 0.486 (2.19) |
| Inadequate training [2016-2019] | 0.497 (2.49) | | 1.171 (2.08) | | | | 0.735 (3.36) | |
| Inadequate planning [2008-2011] | 0.685 (4.51) | | 0.532 (2.46) | | | | 0.714 (3.71) | |
| Inadequate planning [2016-2019] | | | | | | | 0.636 (2.34) | |

| Significant variables | Departure | | Enroute | | Maneuvering | | Arrival | |
|---|---|---|---|---|---|---|---|---|
| | Pilot | Aircraft | Pilot | Aircraft | Pilot | Aircraft | Pilot | Aircraft |
| ***Aircraft characteristics*** | | | | | | | | |
| Airplane [2008-2011] | -0.647 (-4.80) | | -0.645 (-6.14) | | | 0.556 (2.47) | -0.605 (-3.88) | -0.386 (-1.92) |
| Airplane [2012-2015] | -0.494 (-2.62) | | | | | 0.579 (3.60) | -1.578 (-8.83) | |
| Airplane [2016-2019] | -0.290 (-2.84) | | | | | | -0.640 (-4.38) | |
| Helicopter [2008–2011] | -0.541 (-2.25) | | | | -1.365 (-5.13) | | | |
| Helicopter [2012-2015] | -0.926 (-3.02) | | | | | | -1.211 (-4.69) | 0.721 (2.74) |
| Helicopter [2016-2019] | -0.517 (-2.00) | | | | -0.655 (-2.97) | | | |
| Glider [2008-2011] | | | -1.108 (-2.45) | | | | -2.267 (-6.79) | |
| Glider [2012-2015] | | | | | | | -0.568 (-2.04) | |
| Homebuilt aircraft [2008-2011] | 0.580 (5.85) | | | | | | 0.602 (5.12) | |
| Homebuilt aircraft [2012-2015] | | 0.603 (3.75) | | | | | 0.614 (5.69) | 0.628 (3.54) |
| Certificated aircraft [2008-2011] | | | | | 1.267 (4.11) | | | |
| Certificated aircraft [2016-2019] | | | | | | 0.477 (1.66) | -1.026 (-6.10) | |
| Single engine [2008-2011] | | | | | | | -0.415 (-2.92) | |
| Single engine [2012-2015] | | | 0.346 (2.30) | | | | | -0.310 (-2.09) |
| Single engine [2016-2019] | | -0.490 (-3.36) | | -0.369 (-2.05) | | | | -0.381 (-2.50) |
| Zero engine [2008-2011] | | | | | | | 2.244 (6.85) | |
| Zero engine [2016-2019] | | | | | | | 0.833 (4.75) | |
| Multi engine[2008-2011] | | | 0.368 (1.92) | 0.404 (1.59) | | | | |
| Fixed landing gear [2008-2011] | -0.253 (-2.42) | | -0.392 (-3.83) | | | | | |
| Fixed landing gear [2012-2015] | -0.298 (-2.46) | -0.477 (-2.89) | -0.438 (-3.46) | | -0.521 (-2.25) | -0.534 (-3.09) | -0.431 (-4.23) | -0.607 (-3.68) |
| Fixed landing gear [2016-2019] | | | -0.307 (-2.76) | | | -0.741 (-4.03) | | |

| Significant variables | Departure | | Enroute | | Maneuvering | | Arrival | |
|---|---|---|---|---|---|---|---|---|
| | Pilot | Aircraft | Pilot | Aircraft | Pilot | Aircraft | Pilot | Aircraft |
| Aircraft systems issue [2008-2011] | | | | | -0.553 (-2.38) | | -0.606 (-4.10) | -0.931 (-2.00) |
| Aircraft systems issue [2012-2015] | | | | | | | -0.541 (-3.73) | |
| Aircraft power plant issue [2008-2011] | 0.456 (3.85) | | | | -0.382 (-1.75) | | 0.532 (2.71) | |
| Aircraft power plant issue [2012-2015] | 0.483 (2.94) | | | | | | | |
| Aircraft power plant issue [2016-2019] | 0.546 (2.14) | | | | | | 1.077 (3.34) | 0.813 (2.29) |
| Aircraft operation issue [2008-2011] | 0.310 (3.35) | | 0.919 (6.00) | 0.307 (1.70) | 0.776 (5.12) | 0.407 (2.52) | | |
| Aircraft operation issue [2012-2015] | 0.278 (2.92) | | 0.594 (3.44) | | 0.512 (3.37) | | 0.152 (1.85) | |
| Aircraft operation issue [2016-2019] | 0.305 (2.87) | | | 0.485 (2.78) | 0.486 (3.25) | | 0.204 (2.24) | |
| Fluids/misc hardware issue [2008-2011] | 0.560 (3.31) | | | | | | | |
| Fluids/misc hardware issue [2012-2015] | | | | -0.766 (-2.91) | | | 0.949 (5.67) | |
| Fluids/misc hardware issue [2016-2019] | | | | | | | 0.448 (2.02) | |
| ***Flight characteristics*** | | | | | | | | |
| Instrument flight rules [2008-2011] | | | | 0.543 (2.59) | | | 0.533 (3.61) | 0.831 (4.80) |
| Instrument flight rules [2012-2015] | | 0.735 (3.00) | | | | | | |
| Instrument flight rules [2016-2019] | | 1.090 (3.89) | | | | | | |
| Visual flight rules [2016-2019] | -1.216 (-2.72) | | | | | | | |
| None plan [2016-2019] | | 0.874 (3.39) | | | | | | |
| Personal flying [2012-2015] | | | | 0.448 (2.68) | 0.702 (3.89) | | | |
| Instructional flying [2008-2011] | -0.862 (-3.48) | | | | | | | |
| Instructional flying [2012-2015] | | | | | | | -0.536 (-3.07) | |
| Instructional flying [2016-2019] | -0.585 (-2.16) | | | | | | -0.533 (-2.31) | -0.675 (-1.68) |
| Business [2012-2015] | | | 0.744 (2.63) | | 1.079 (2.07) | | | |

| Significant variables | Departure | | Enroute | | Maneuvering | | Arrival | |
|---|---|---|---|---|---|---|---|---|
| | Pilot | Aircraft | Pilot | Aircraft | Pilot | Aircraft | Pilot | Aircraft |
| ***Environmental characteristics*** | | | | | | | | |
| Daylight [2008-2011] | | | | -0.415 (-2.39) | | | -0.453 (-3.48) | |
| Daylight [2012-2015] | -0.660 (-3.00) | | | | | | | |
| Daylight [2016-2019] | | -0.857 (-3.55) | | | | -0.688 (-2.63) | | |
| Night [2008-2011] | | 0.524 (1.73) | | | | 0.855 (1.99 ) | | |
| Night [2016-2019] | | -0.869 (-1.96) | | 0.461 (2.04) | | | | |
| Clear [2012-2015] | | | | -0.579 (-3.35) | | | | |
| Clear [2016-2019] | | | | | -0.295 (-2.38) | | | |
| Few clouds [2008-2011] | | | | | | | | -0.550 (-1.77) |
| Few clouds [2012-2015] | | | | | | | -0.236 (-1.76) | |
| Scattered clouds [2008-2011] | | -1.033 (-1.70) | | | | | | |
| Scattered clouds [2012-2015] | | | | -0.410 (-1.90) | | 0.504 (2.12) | | |
| Broken clouds [2012-2015] | 0.234 (2.05) | | | | | | | |
| Broken clouds [2016-2019] | | | | 0.267 (1.68) | | | | |
| Overcast condition [2008-2011] | | | 0.339 (1.88) | 0.532 (2.81) | | | | |
| Overcast condition [2012-2015] | | | | | | | 0.294 (1.84) | |
| Overcast condition [2016-2019] | | | | | | | 0.445 (2.83) | |
| None clouds [2008-2011] | 0.234 (2.40) | | | | | | 0.483 (4.48) | |
| None clouds [2016-2019] | | | | | | | 1.123 (6.28) | |
| Visual meteorological condition [2008-2011] | -0.572 (-3.37) | -0.798 (-2.25) | | | -1.209 (-3.93) | -0.401 (-1.66) | | |
| Visual meteorological condition [2012-2015] | -1.091 (-3.37) | | -0.888 (-6.23) | -0.518 (-2.68) | -1.139 (-2.83) | -1.034 (-5.33) | -1.165 (-5.52) | -1.516 (-9.90) |
| Visual meteorological condition [2016-2019] | | -0.810 (-2.96) | -0.429 (-4.29) | | | | | |

| Significant variables | Departure | | Enroute | | Maneuvering | | Arrival | |
|---|---|---|---|---|---|---|---|---|
| | Pilot | Aircraft | Pilot | Aircraft | Pilot | Aircraft | Pilot | Aircraft |
| Ceiling/visibility [2008-2011] | | | 1.004 (3.37) | 0.741 (3.32) | 1.223 (1.75) | | 0.634 (1.98) | 0.815 (3.10) |
| Ceiling/visibility [2012-2015] | | 0.577 (1.65) | | | | | | |
| Ceiling/visibility [2016-2019] | 2.121 (2.95) | | 1.715 (4.23) | 0.824 (3.60) | | | 1.168 (2.87) | 1.318 (3.82) |
| Wind [2008-2011] | -0.410 (-2.34) | | | | | | | |
| Wind [2012-2015] | | | | | -0.680 (-2.27) | | | |
| Wind [2016-2019] | -1.109 (-5.33) | -0.713 (-2.19) | | | | | -0.493 (-3.35) | -0.444 (-2.34) |
| Terrain [2008-2011] | | | | 0.917 (3.18) | | | | |
| Terrain [2016-2019] | -0.834 (-1.90) | | | | | | | |
| Object/animal/substance [2016-2019] | -0.780 (-4.04) | | | | | | | -0.707 (-2.30) |
| Temperature [2008-2011] | 0.399 (1.98) | | | | -0.731 (-2.72) | | 0.918 (2.69) | |
| ***Random parameters*** | | | | | | | | |
| Homebuilt aircraft [2008-2011] | | -1.991 (-3.07) | 0.951 (3.98) | | | | | |
| Standard deviation | | 2.269 (5.06) | 0.939 (6.06) | | | | | |
| Homebuilt aircraft [2012-2015] | 0.309 (2.39) | | | | | | | |
| Standard deviation | 0.437 (3.58) | | | | | | | |
| Fluids/misc hardware issue[2008-2011] | | | -0.750 (-3.90) | | | | | |
| Standard deviation | | | 1.064 (6.26) | | | | | |
| Right seat [2008-2011] | | | | | 1.086 (2.66) | | | |
| Standard deviation | | | | | 1.515 (4.37) | | | |
| Elder pilot [2012-2015] | | | 0.489 (4.01) | | | | | |
| Standard deviation | | | 0.509 (4.37) | | | | | |
| Fixed landing gear [2012-2015] | | | | -0.885 (-5.44) | | | | |

| Significant variables | Departure | | Enroute | | Maneuvering | | Arrival | |
|---|---|---|---|---|---|---|---|---|
| | Pilot | Aircraft | Pilot | Aircraft | Pilot | Aircraft | Pilot | Aircraft |
| Standard deviation | | | | 0.726 (6.43) | | | | |
| Helicopter [2012-2015] | | | | | -1.229 (-4.07) | | | |
| Standard deviation | | | | | 0.860 (4.10) | | | |
| Decision making errors [2012-2015] | | | | | 4.218 (3.62) | | | |
| Standard deviation | | | | | 3.839 (3.98) | | | |
| Visual meteorological condition [2008-2011] | | | | | | | -1.162 (-6.24) | |
| Standard deviation | | | | | | | 0.893 (15.72) | |
| Aircraft systems issue [2012-2015] | | | | | | | | -2.979 (-2.86) |
| Standard deviation | | | | | | | | 1.646 (3.09) |
| Adverse physiologic [2012-2015] | | | | | | | 2.690 (3.75) | |
| Standard deviation | | | | | | | 1.531 (2.41) | |
| Glider [2016-2019] | | | | | | | -2.420 (-3.14) | |
| Standard deviation | | | | | | | 1.759 (2.76) | |
| Clear [2016-2019] | | | | | | | 0.479 (1.95) | |
| Standard deviation | | | | | | | 0.446 (7.62) | |
| ***Heterogeneity in the means of random parameters*** | | | | | | | | |
| Homebuilt aircraft: Helicopter[2008-2011] | | 2.676 (2.04) | | | | | | |
| Homebuilt aircraft: Decision making errors [2012-2015] | 1.416 (2.90) | | | | | | | |
| Fluids/misc hardware issue: IFR [2008-2011] | | | 1.239 (2.62) | | | | | |
| Homebuilt aircraft: None clouds[2008-2011] | | | -0.698 (-2.47) | | | | | |
| Homebuilt aircraft: perception errors [2008-2011] | | | -2.510 (-2.17) | | | | | |
| Elder pilot: Aircraft systems issue [2012-2015] | | | -1.018 (-2.08) | | | | | |

| **Significant variables** | **Departure** | | **Enroute** | | **Maneuvering** | | **Arrival** | |
|---|---|---|---|---|---|---|---|---|
| | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** |
| Right seat: Aircraft operation issue [2008-2011] | | | | | -1.244 (-2.61) | | | |
| Helicopter: Attention condition [2012-2015] | | | | | 1.530 (2.01) | | | |
| Helicopter: Class 3 [2012-2015] | | | | | -2.094 (-2.28) | | | |
| Helicopter: Left seat [2012-2015] | | | | | 1.151 (2.88) | | | |
| Decision making errors: Left seat [2012-2015] | | | | | -2.272 (-2.81) | | | |
| Visual meteorological condition: Scattered clouds [2008-2011] | | | | | | | 0.229 (1.78) | |
| Aircraft systems issue: Helicopter [2012-2015] | | | | | | | | 3.333 (2.21) |
| Adverse physiologic: Attention condition [2012-2015] | | | | | | | -3.223 (-2.10) | |
| Glider: Front seat [2016-2019] | | | | | | | 3.365 (2.81) | |
| Clear: None clouds [2016-2019] | | | | | | | -1.484 (-5.02) | |
| ***Correlation coefficient*** | | | | | | | | |
| $\rho$ [2008-2011] | 0.585 (5.44) | | 0.751 (8.02) | | 0.582 (5.99) | | 0.648 (8.05) | |
| $\rho$ [2012-2015] | 0.859 (20.70) | | 0.877 (16.27) | | 0.793 (11.50) | | 0.814 (13.52) | |
| $\rho$ [2016-2019] | 0.673 (11.26) | | 0.770 (13.72) | | 0.685 (8.35) | | 0.735 (14.55) | |
| ***Model statistics*** | | | | | | | | |
| Number of parameters, K [2008-2011] | 29 | | 32 | | 21 | | 32 | |
| Number of observations, N [2008-2011] | 1184 | | 767 | | 504 | | 2125 | |
| Log-likelihood at convergence, LL($\beta$) [2008-2011] | -782.829 | | -534.322 | | -400.895 | | -887.979 | |
| Log-likelihood at zero, LL(0) [2008-2011] | -935.960 | | -695.612 | | -497.776 | | -1121.661 | |
| $\rho^2 = 1 - LL(\beta)/LL(0)$ [2008-2011] | 0.164 | | 0.232 | | 0.195 | | 0.208 | |
| Corrected $\rho^2 = 1 - (LL(\beta) - K)/LL(0)$ [2008-2011] | 0.133 | | 0.186 | | 0.152 | | 0.180 | |

| **Significant variables** | **Departure** | | **Enroute** | | **Maneuvering** | | **Arrival** | |
|---|---|---|---|---|---|---|---|---|
| | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** |
| Number of parameters, K [2012-2015] | 30 | | 23 | | 23 | | 32 | |
| Number of observations, N [2012-2015] | 929 | | 658 | | 436 | | 1690 | |
| Log-likelihood at convergence, LL($\beta$) [2012-2015] | -670.269 | | -504.283 | | -383.121 | | -779.847 | |
| Log-likelihood at zero, LL(0) [2012-2015] | -791.133 | | -642.183 | | -473.273 | | -1006.986 | |
| $\rho^2 = 1 - LL(\beta)/LL(0)$ [2012-2015] | 0.153 | | 0.215 | | 0.190 | | 0.226 | |
| Corrected $\rho^2 = 1 - (LL(\beta) - K)/LL(0)$ [2012-2015] | 0.115 | | 0.179 | | 0.142 | | 0.194 | |
| Number of parameters, K [2016-2019] | 20 | | 19 | | 13 | | 34 | |
| Number of observations, N [2016-2019] | 791 | | 588 | | 303 | | 1648 | |
| Log-likelihood at convergence, LL($\beta$) [2016-2019] | -624.659 | | -490.255 | | -312.614 | | -790.762 | |
| Log-likelihood at zero, LL(0) [2016-2019] | -721.315 | | -600.087 | | -353.585 | | -946.330 | |
| $\rho^2 = 1 - LL(\beta)/LL(0)$ [2016-2019] | 0.134 | | 0.183 | | 0.116 | | 0.164 | |
| Corrected $\rho^2 = 1 - (LL(\beta) - K)/LL(0)$ [2016-2019] | 0.106 | | 0.151 | | 0.079 | | 0.128 | |

Note: Significance of estimated parameters was determined at a 90% confidence level.

## 5.5 Case study

To demonstrate the application of the risk knowledge base in causal graph augmentation for accident investigation, we present a representative case study. The constructed knowledge base contains 199 statistically significant risk factors, each standardized in the format of "factor name–flight phase–outcome direction." These entries can be systematically matched with specific accident cases to support structured retrieval and subsequent graph augmentation.

**Case background.** "On February 26, 2020, an Arion Lightning airplane (registration N1XF) was substantially damaged in an accident near Jerome, Idaho. The 61-year-old male private pilot (rated for single engine land/sea, Class 3 medical) sitting in the left seat sustained only minor injuries during a Part 91 personal flight. The aircraft was a single-engine reciprocating airplane with a fixed tricycle landing gear. Prior to the flight, the right fuel gauge was known to be unreliable; despite two repair attempts, the issue persisted. During the ferry flight, the pilot relied on the fuel totalizer and manual calculations for fuel management. Approximately 25 nautical miles from the destination, the engine began to sputter and lost power. Multiple fuel selector changes did not restore engine function. After aborting an attempted road landing due to roadside fences, the pilot turned toward a nearby field, where the left wing stalled during descent, causing the airplane to impact the ground left wing first. The fuselage and wings sustained substantial damage. Post-accident inspection revealed little to no remaining fuel, and no fuel odor

was detected at the site. The accident occurred under daylight, *VMC*, with wind about 5 knots and no turbulence."

Accordingly, the attributes of this accident are annotated as $\phi_x$ = Arrival, with the outcome $(y_x^{(1)}, y_x^{(2)}) = (0, 0)$ (pilot non-fatal/non-serious injury; aircraft not destroyed).

**Step 1: Structured filtering.** Given $(\phi_x, (y_x^{(1)}, y_x^{(2)})) = (\text{Arrival}, (0, 0))$, we map the binary accident outcomes into directional values by $s(1) = +1$ and $s(0) = -1$. Thus, the outcome $(0, 0)$ corresponds to $(-1, -1)$, and the admissible directional combinations are (-1,-1), (-1,0), (0,-1).

Applying this rule to the knowledge base $\mathcal{K}$ yields 27 distinct risk factors that match the arrival phase and the specified outcome directions (see Table 5.5).

Table 5.5: The distinct candidate risk factors for the arrival phase with outcome $(0, 0)$

| Risk factor name | Fatal or serious injury direction | in-Destroy direction |
|---|---|---|
| Young pilot | -1 | 0 |
| Female pilot | -1 | 0 |
| Class 1 | -1 | 0 |
| Sport | -1 | 0 |
| Front seat | 0 | -1 |
| Airplane | -1 | -1 |
| Airplane | -1 | 0 |
| Airplane | -1 | 0 |
| Glider | -1 | 0 |
| Glider | -1 | 0 |
| Certificated aircraft | -1 | 0 |
| Single engine | -1 | 0 |
| Single engine | 0 | -1 |
| Single engine | 0 | -1 |
| Fixed landing gear | -1 | -1 |
| Aircraft systems issue | -1 | -1 |
| Aircraft systems issue | -1 | 0 |
| Aircraft systems issue | 0 | -1 |
| Instructional flying | -1 | 0 |
| Instructional flying | -1 | -1 |
| Daylight | -1 | 0 |
| Few clouds | 0 | -1 |
| Few clouds | -1 | 0 |
| Visual meteorological condition | -1 | -1 |
| Visual meteorological condition* | -1 | 0 |
| Wind | -1 | -1 |
| Object/animal/substance | 0 | -1 |

* This factor corresponds to a random parameter specification.

**Step 2: Narrative alignment.** We align the Step 1 candidates with the accident narrative using LLMs. A factor is retained if it is explicitly mentioned in the text, not negated, and its operational definition is satisfied. For example, the

factor *Wind* denotes strong wind ($\geq$15 kt); since the reported wind was only 5 kt, it is excluded.

The LLM further identified explicit mentions that the airplane was a single-engine reciprocating aircraft with a fixed tricycle landing gear, operating under daylight and *VMC* conditions. Based on this evidence, the retained set comprises **six directional candidates**—*Daylight*, *Visual meteorological condition*, *Airplane* (two directions), *Single engine*, and *Fixed landing gear* (see Table 5.6). Since *Airplane* appears with two directional effects, these six candidates correspond to five unique risk factors, which are ultimately used for augmentation. That is, the mention-filtered set $\tilde{\mathcal{C}}$ contains those tuples $R_i$ whose factor names reduce to the set $\{\,\text{Daylight},\ \text{Visual meteorological condition},\ \text{Airplane},\ \text{Single engine},\ \text{Fixed landing gear}\,\}$.

Table 5.6: Candidates retained after LLM-based narrative alignment (multiple directions may correspond to the same factor)

| Risk factor name | Fatal or serious injury direction | Destroy direction |
|---|---|---|
| Daylight | -1 | 0 |
| Visual meteorological condition | -1 | -1 |
| Visual meteorological condition* | -1 | 0 |
| Airplane | -1 | -1 |
| Airplane | -1 | 0 |
| Single engine | -1 | 0 |
| Fixed landing gear | -1 | -1 |

*This factor corresponds to a random parameter specification.

**Step 3: Graph de-duplication.** The initial accident graph $G_0$ of this accident

case identified in Study 4 defined with the following edge set:

$$
\begin{aligned}
E(G_0) = \{&(\text{Instrument Reading Error}, \text{Pilot's mismanagement of fuel in flight}),\\
&(\text{Pilot's mismanagement of fuel in flight}, \text{Fuel-Related issues}),\\
&(\text{Fuel-Related issues}, \text{Loss of engine power}),\\
&(\text{Loss of engine power}, \text{Off-field or emergency landing}),\\
&(\text{Off-field or emergency landing}, \text{Loss of control in flight}),\\
&(\text{Loss of control in flight}, \text{Nose over/Roll over}),\\
&(\text{Nose over/roll over}, \text{Collision with terr/obj (non-CFIT)})\}.
\end{aligned}
$$

and the corresponding node set:

$$
\begin{aligned}
V(G_0) = \{&\text{Collision with terr/obj (non-CFIT)}, \text{Loss of control in flight},\\
&\text{Nose over/Roll over}, \text{Off-field or emergency landing},\\
&\text{Pilot's mismanagement of fuel in flight}, \text{Fuel-Related issues},\\
&\text{Loss of engine power}, \text{Instrument Reading Error}\}.
\end{aligned}
$$

To avoid redundancy, we exclude any candidate factor from Step 2 whose name already exists in $V(G_0)$. Since the nodes of $G_0$ are coded categories (e.g., Collision with terr/obj (non-CFIT), Fuel-Related issues) and do not overlap with the Step 2 factors (*Daylight*, *Visual meteorological condition*, *Airplane*, *Single engine*, *Fixed landing gear*), no duplicates are found.

Therefore, the de-duplicated candidate set $\mathcal{C}^\star$ can be summarized, at the factor-

name level, as

$$\mathcal{C}^\star = \{\text{Daylight},\ \text{Visual meteorological condition},\ \text{Airplane},\ \text{Single engine},\ \text{Fixed landing gear}\}.$$

All five candidate risk factors are retained for subsequent graph augmentation.

**Step 4: Graph augmentation.** In this step, the LLM reasons over $\mathcal{C}^\star$, the labels and topology of $G_0$, and the narrative $\mathcal{N}_x$ to attach each retained factor to a semantically appropriate anchor node while preserving acyclicity. The attachment follows a simple principle: exogenous risk factors are directed toward the initial accident nodes that they most plausibly condition in the given case. The following examples demonstrate this attachment process.

*Example 1 (Daylight → Pilot's mismanagement of fuel in flight and loss of control in flight).* Daylight is statistically associated with a higher likelihood of minor pilot injury in Arrival. The LLM therefore links *Daylight* to *Pilot's mismanagement of fuel in flight*, reasoning that favorable visibility may reduce vigilance and contribute to inadequate fuel monitoring. Similarly, this reduced vigilance may lead to errors in aircraft controlling.

*Example 2 (Single engine → Loss of engine power).* *Single engine* captures a single-point-of-failure configuration. The LLM links it to *Loss of engine power*, since a power loss on a single-engine airplane is especially consequential; the edge flows from context to failure mode without introducing cycles.

*Example 3 (Airplane → Fuel-Related issues).* Given the narrative's emphasis on an unreliable fuel gauge and fuel-management problems, the LLM anchors the generic *Airplane* factor to *Fuel-Related issues*, where aircraft category and design

context intersect with the specific fuel failure; the orientation is context to issue, preserving acyclicity.

The remaining factors are connected analogously: *Visual meteorological condition* is anchored to *Loss of control in flight* and *Pilot's mismanagement of fuel in flight*, as VMC facilitates control failure and vigilance, while *Fixed landing gear* is linked to *Off-field or emergency landing*, reflecting its effect on touchdown dynamics and damage outcomes.

The resulting augmented graph $G_{\text{augment}}$ introduces the following edges:

$$
\begin{aligned}
E_{\text{add}} = \{&(\text{Daylight},\ \text{Pilot's mismanagement of fuel in flight}),\\
&(\text{Visual meteorological condition},\ \text{Pilot's mismanagement of fuel in flight}),\\
&(\text{Visual meteorological condition},\ \text{Loss of control in flight}),\\
&(\text{Daylight},\ \text{Loss of control in flight}),\\
&(\text{Airplane},\ \text{Fuel-Related issues}),\\
&(\text{Single engine},\ \text{Loss of engine power}),\\
&(\text{Fixed landing gear},\ \text{Off-field or emergency landing})\ \}.
\end{aligned}
$$

Finally, the augmented graph $G_{\text{augment}}$ combines the original edge set $E(G_0)$ with the augmentation edges $E_{\text{add}}$, yielding:

$$
\begin{aligned}
E(G_{\text{augment}}) = \{&(\text{Instrument Reading Error},\ \text{Pilot's mismanagement of fuel in flight}),\\
&(\text{Pilot's mismanagement of fuel in flight},\ \text{Fuel-Related issues}),\\
&(\text{Pilot's mismanagement of fuel in flight},\ \text{Loss of control in flight}),\\
&(\text{Fuel-Related issues},\ \text{Loss of engine power}),\\
&(\text{Loss of engine power},\ \text{Off-field or emergency landing}),\\
&(\text{Off-field or emergency landing},\ \text{Loss of control in flight}),\\
&(\text{Loss of control in flight},\ \text{Nose over/Roll over}),\\
&(\text{Nose over/Roll over},\ \text{Collision with terr/obj (non-CFIT)}),\\
&(\text{Daylight},\ \text{Pilot's mismanagement of fuel in flight}),\\
&(\text{Visual meteorological condition},\ \text{Fuel-Related issues}),\\
&(\text{Visual meteorological condition},\ \text{Loss of control in flight}),\\
&(\text{Airplane},\ \text{Fuel-Related issues}),\\
&(\text{Single engine},\ \text{Loss of engine power}),\\
&(\text{Fixed landing gear},\ \text{Off-field or emergency landing})\ \}.
\end{aligned}
$$

Fig.5.6 shows the initial causal graph $G_0$ augmented graph $G_{\text{augment}}$.

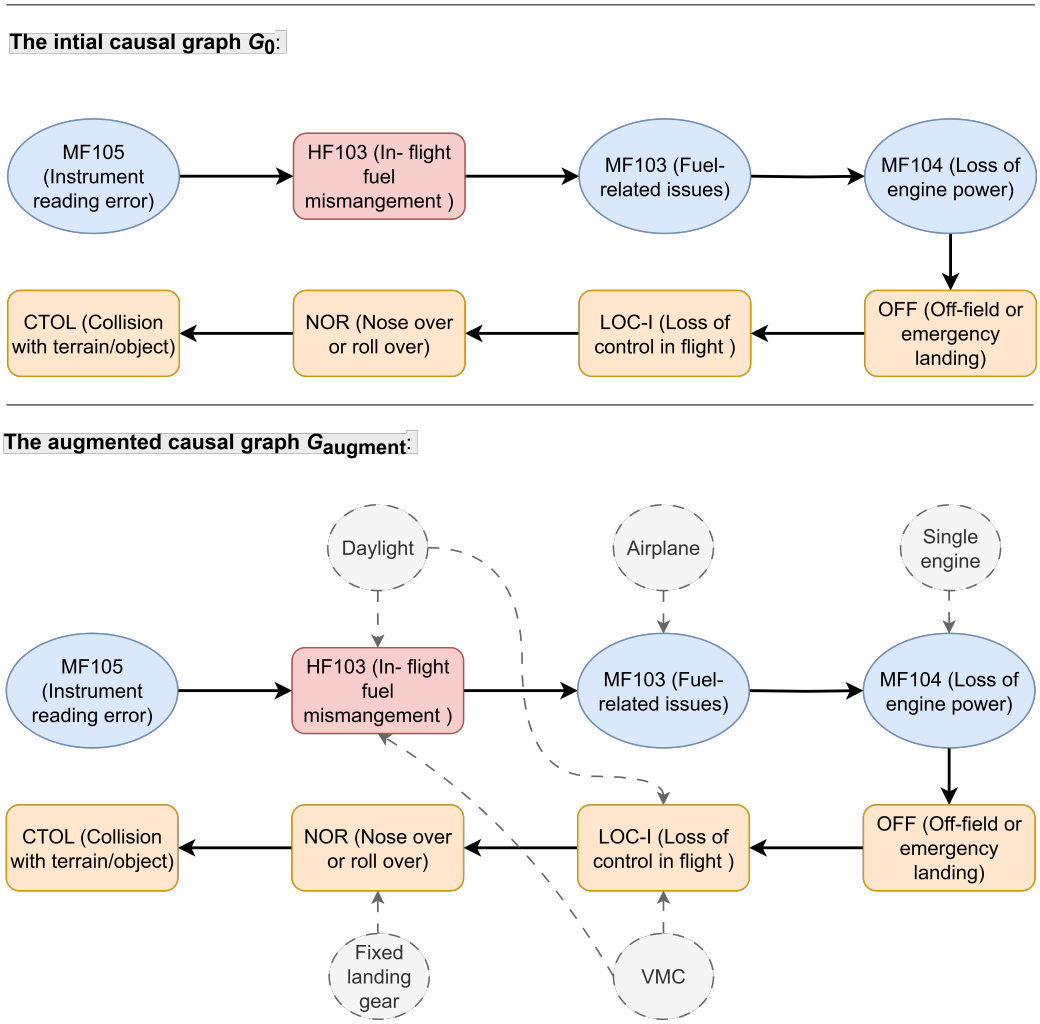



Figure 5.6: The comparison of initial graph $G_0$ and the augmented graph $G_{\text{augment}}$.

## 5.6 Concluding remarks

This study proposed a framework for GA post-accident investigation that integrates econometric modeling with LLMs to identify exogenous risk factors, build a structured knowledge base, and augment causal graphs. By jointly analyzing pilot injury severity and aircraft damage, the approach addressed phase transferabil-

ity, temporal validity, correlated outcomes, and unobserved heterogeneity, thereby improving the accuracy of risk factor identification. A case study demonstrates the framework's efficacy. Through structured filtering and narrative alignment, we retrieved several exogenous risk factors and incorporated them as dashed-box nodes into the causal graph. This augmentation revealed how these factors increased the likelihood of minor injury and substantial damage, providing a more comprehensive accident explanation.

In summary, this study offers four main contributions:

1. We integrate econometric modeling and LLMs into a unified framework, linking quantitative risk factor identification with semantic retrieval from unstructured accident narratives.

2. By employing random parameter bivariate probit models with heterogeneity in means, we account for phase transferability, temporal validity, dual accident outcomes, and unobserved heterogeneity, ensuring more accurate and robust GA risk factors.

3. We construct a structured knowledge base that systematically organizes statistically validated risk factors across flight phases, temporal windows, and accident outcomes, creating a reusable resource for accident investigation and safety research.

4. By embedding statistically validated risk factors into causal graphs, we advance accident investigation from describing how an accident unfolded to explaining why it was more likely under the prevailing conditions. This enables investigators to distinguish immediate causes from exogenous risks and supports policymakers in designing targeted safety measures.

Nevertheless, some limitations remain. First, variables influencing mean heterogeneity of random parameters were omitted from graph augmentation, which reduced granularity. Second, the knowledge base depends on the completeness of historical accident data, and missing variables such as pilot cognitive states may limit the capture of unobserved heterogeneity. Future work could incorporate heterogeneity covariates, integrate real-time flight or physiological data, and extend the framework to commercial aviation and international datasets.

# Chapter 6

# Conclusion

This thesis developed and validated an LLM-driven life-cycle framework for GA accident investigation, spanning the preliminary, detailed, and post-investigation stages. It introduced three complementary strategies: domain-informed prompting for human error identification, a dynamic reasoning workflow for multi-event causal reconstruction, and an econometric–LLM integration for risk factor augmentation. These methods not only advance human error detection, causal graph construction, and post-investigation analysis, but also reduce hallucinations and improve reasoning reliability. Collectively, they move GA accident analysis from isolated factor extraction toward comprehensive, stage-aware causal reconstruction. The remainder of this chapter summarizes the key findings and contributions in Section 6.1, and discusses limitations and future research directions in Section 6.2.

## 6.1 Contributions

This thesis addresses the problem of causal chain reconstruction in GA accident investigation and proposes a life-cycle framework driven by LLMs. The research spans the preliminary, detailed, and post-investigation stages, with systematic exploration in prompt design, workflow modeling, and statistical modeling. The main contributions are summarized as follows:

1. **Domain-informed prompt design for human error identification in preliminary investigation** This study introduces a novel prompt strategy that combines human factors knowledge with CoT reasoning to guide LLMs in stepwise identification of pilot errors from unstructured early-stage evidence such as narratives and self-reports. The approach improves the accuracy of extracting initial causal nodes related to human errors and establishes a reliable foundation for subsequent causal reconstruction.

2. **Dynamic workflow framework for causal chain reconstruction in detailed investigation** To overcome the limitations of traditional linear models that fail to capture the dynamic and multi-event evolution of GA accidents, this study develops a novel workflow framework inspired by field theory. The framework enables LLMs to reconstruct coherent and temporally consistent causal graphs by modeling both within-event and cross-event dependencies. This contribution goes beyond static root-cause analysis and provides a more explanatory representation of complex accident mechanisms.

3. **Risk factor augmentation through econometric–LLM integration in post-investigation analysis** This research is the first to integrate a random-parameter

bivariate probit model with heterogeneity in means and LLMs for the identification and validation of exogenous risk factors in GA accidents. Compared with traditional Logit or Poisson models, this econometric method jointly models pilot injury severity and aircraft damage, while accounting for phase transferability, temporal validity, correlated outcomes, and unobserved heterogeneity. The structured risk knowledge base derived from this analysis, when combined with LLM-enabled semantic retrieval and graph augmentation, extends causal graphs from explaining "how accidents occurred" to also capturing "why they were more likely under specific conditions," thereby offering more comprehensive support for investigation and safety management.

Overall, this thesis advances the academic frontier by extending the application of LLMs to aviation safety, introduces a new paradigm for accident investigation in practice, and demonstrates a methodological synthesis of domain knowledge, structured reasoning, and statistical validation that can be transferred to other safety-critical domains.

## 6.2 Limitations and future research

Despite the contributions of this thesis, several limitations should be acknowledged, which also point to directions for future work.

1. *Limited domain knowledge of LLMs.* General aviation accident investigation is a highly specialized domain, yet current large language models often lack sufficient aviation-specific knowledge. For example, an LLM may in-

correctly infer that a wind speed of 5 knots is a causal factor, although this level is far below the operational thresholds of aircraft and has negligible impact on flight performance. Addressing this limitation requires pre-training or fine-tuning LLMs on aviation-specific corpora to enhance their domain knowledge and reasoning reliability.

2. *Focus confined to causation reconstruction.* The proposed framework centers on the core task of causation reconstruction within accident investigation. However, it does not extend to upstream tasks such as data collection (e.g., LLM-based witness interviewing agents or trajectory anomaly detection) nor to downstream tasks such as automated accident report generation and safety recommendations. Future research could aim to develop an end-to-end LLM-based investigation pipeline that covers the entire investigation lifecycle, from information collection to reporting and recommendations.

3. *Data source restricted to U.S. GA accidents.* All empirical analyses in this thesis rely on accident data from the National Transportation Safety Board (NTSB), which records U.S. GA accidents. Although the U.S. has the largest volume of GA operations worldwide, accidents in other countries may exhibit unique characteristics due to different regulatory environments, operational practices, and pilot demographics. Expanding the dataset to include international GA accident records would improve the generalizability and robustness of the findings.

4. *Model improvements and their impact on results.* While this study uses state-of-the-art LLMs (GPT-4o, GPT-4.1), future improvements in LLM models (e.g., larger models or newer architectures like GPT-5) could further

enhance reasoning capabilities and thus influence the current experimental results. However, despite these advancements, the potential improvements are still limited by the model's inherent understanding of aviation-specific knowledge and causal reasoning. Even with more powerful models, LLMs will still require domain-specific pre-training, fine-tuning, and human-in-the-loop validation to ensure reliable and accurate results. Future work should explore how advancements in LLMs can be leveraged while managing their limitations in specialized domains.

# References


[1] Dambier, M., Hinkelbein, J. “Analysis of 2004 German general aviation aircraft accidents according to the HFACS model”. In: *Air medical journal* 25.6 (2006), pp. 265–269.

[2] Filho, A. P. G., Souza, C. A., Siqueira, E. L. B., Souza, M. A., Vasconcelos, T. P. “An analysis of helicopter accident reports in Brazil from a human factors perspective”. In: *Reliability Engineering and System Safety* 183 (2019), pp. 39–46. ISSN: 09518320. DOI: https://doi.org/10.1016/j.ress.2018.11.003.

[3] NTSB, *US Civil Aviation Accident Dashboard: 2008-2022*. Web Page. 2023. URL: https://www.ntsb.gov/safety/StatisticalReviews/Pages%20/CivilAviationDashboard.aspx.

[4] Johnson, C., Holloway, C. M. “A survey of logic formalisms to support mishap analysis”. In: *Reliability Engineering and System Safety* 80.3 (2003), pp. 271–291. ISSN: 09518320. DOI: https://doi.org/10.1016/s0951-8320(03)00053-x.

[5] NTSB, *49 CFR part 830*. Web Page. [accessed 18th October 2024]. 2024. URL: https://www.ecfr.gov/current/title-49/subtitle-B/chapter-VIII/part-830.

[6] Patil, A., Jadon, A. "Advancing Reasoning in Large Language Models: Promising Methods and Approaches". In: *arXiv e-prints*, arXiv:2502.03671 (Feb. 2025), arXiv:2502.03671. DOI: 10.48550/arXiv.2502.03671. arXiv: 2502.03671 [cs.CL].

[7] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., Fung, P. "Survey of Hallucination in Natural Language Generation". In: *ACM Computing Surveys* 55.12 (Mar. 2023), pp. 1–38. ISSN: 1557-7341. DOI: 10.1145/3571730. URL: http://dx.doi.org/10.1145/3571730.

[8] Chang, Y., Cao, B., Lin, L. "Monitoring Decoding: Mitigating Hallucination via Evaluating the Factuality of Partial Response during Generation". In: *ArXiv* abs/2503.03106 (2025). DOI: 10.48550/arXiv.2503.03106.

[9] Jiang, Z., Peng, H., Feng, S., Li, F., Li, D. "LLMs can Find Mathematical Reasoning Mistakes by Pedagogical Chain-of-Thought". In: *arXiv preprint arXiv:2405.06705* (2024).

[10] Li, R., Luo, Z., Du, X. "Fine-grained Hallucination Detection and Mitigation in Language Model Mathematical Reasoning". In: *ArXiv* abs/2410.06304 (2024). DOI: 10.48550/arXiv.2410.06304.

[11] Boyd, D. D. "A Review of General Aviation Safety (1984-2017)". In: *Aerospace Medicine and Human Performance* 88.7 (2017), pp. 657–664.

ISSN: 2375-6314 (Print) 2375-6314 (Linking). DOI: https://doi.org/10.3357/AMHP.4862.2017.

[12] Burns, K., Bonaceto, C. "An empirically benchmarked human reliability analysis of general aviation". In: *Reliability Engineering and System Safety* 194 (2020). ISSN: 09518320. DOI: https://doi.org/10.1016/j.ress.2018.07.028.

[13] Li, G., Baker, S. P. "Correlates of pilot fatality in general aviation crashes". In: *Aviation, Space, and Environmental Medicine* 70.4 (1999), pp. 305–309.

[14] Bazargan, M., Guzhva, V. S. "Factors contributing to fatalities in general aviation accidents". In: *World Review of Intermodal Transportation Research* 1.2 (2007), pp. 170–182.

[15] Shao, B. S., Guindani, M., Boyd, D. D. "Causes of fatal accidents for instrument-certified and non-certified private pilots". In: *Accident Analysis and Prevention* 72 (2014), pp. 370–375. ISSN: 1879-2057 (Electronic) 0001-4575 (Linking). DOI: https://doi.org/10.1016/j.aap.2014.07.013.

[16] Bazargan, M., Guzhva, V. S. "Impact of gender, age and experience of pilots on general aviation accidents". In: *Accident Analysis and Prevention* 43.3 (2011), pp. 962–970. ISSN: 0001-4575. DOI: 10.1016/j.aap.2010.11.023.

[17] Li, G., Baker, S. P., Grabowski, J. G., Rebok, G. W. "Factors associated with pilot error in aviation crashes". In: *Aviation, Space, and Environmental Medicine* 72.1 (2001), pp. 52–58.

[18] Li, G., Baker, S. P., Lamb, M. W., Qiang, Y., McCarthy, M. L. "Characteristics of alcohol-related fatal general aviation crashes". In: *Accident Analysis and Prevention* 37.1 (2005), pp. 143–8. ISSN: 0001-4575 (Print) 0001-4575 (Linking). DOI: https://doi.org/10.1016/j.aap.2004.03.005.

[19] Li, G., Baker, S. P., Qiang, Y., Grabowski, J. G., McCarthy, M. L. "Driving-while-intoxicated history as a risk marker for general aviation pilots". In: *Accident Analysis and Prevention* 37.1 (2005), pp. 179–184. ISSN: 0001-4575.

[20] Canfield, D. V., Dubowski, K. M., Chaturvedi, A. K., Whinnery, J. E. "Drugs and alcohol found in civil aviation accident pilot fatalities from 2004–2008". In: *Aviation, Space, and Environmental Medicine* 83.8 (2012), pp. 764–770. ISSN: 0095-6562.

[21] Shao, B. S., Guindani, M., Boyd, D. D. "Causes of fatal accidents for instrument-certified and non-certified private pilots". In: *Accident Analysis and Prevention* 72 (2014), pp. 370–5. ISSN: 1879-2057 (Electronic) 0001-4575 (Linking). DOI: 10.1016/j.aap.2014.07.013.

[22] Van Benthem, K., Herdman, C. M. "The importance of domain-dependent cognitive factors in GA safety: Predicting critical incidents with prospective memory, situation awareness, and pilot attributes". In: *Safety Science* 130 (2020). ISSN: 09257535. DOI: https://doi.org/10.1016/j.ssci.2020.104892.

[23] De Voogt, A. J., Doorn, R. R. “Sports aviation accidents: Fatality and aircraft specificity”. In: *Aviation, Space, and Environmental Medicine* 81.11 (2010), pp. 1033–1036.

[24] Boyd, D. D., Stolzer, A. “Accident-precipitating factors for crashes in turbine-powered general aviation aircraft”. In: *Accident Analysis and Prevention* 86 (2016), pp. 209–16. ISSN: 1879-2057 (Electronic) 0001-4575 (Linking). DOI: https://doi.org/10.1016/j.aap.2015.10.024.

[25] de Voogt, A., Lu, Y., Pan, F., Zuckerman, H. “Abrupt maneuvers in General Aviation accidents”. In: *Safety Science* 185 (2025), p. 106785. ISSN: 0925-7535. DOI: https://doi.org/10.1016/j.ssci.2025.106785.

[26] Neuhaus, C., Dambier, M., Glaser, E., Schwalbe, M., Hinkelbein, J. “Probabilities for severe and fatal injuries in general aviation accidents”. In: *Journal of Aircraft* 47.6 (2010), pp. 2017–2020.

[27] Voogt, A., Kalagher, H., Santiago, B., Lang, J. W. B. “Go-around accidents and general aviation safety”. In: *Journal of Safety Research* 82 (2022), pp. 323–328. ISSN: 1879-1247 (Electronic) 0022-4375 (Linking). DOI: 10.1016/j.jsr.2022.06.008.

[28] Aguiar, M., Stolzer, A., Boyd, D. D. “Rates and causes of accidents for general aviation aircraft operating in a mountainous and high elevation terrain environment”. In: *Accident Analysis and Prevention* 107 (2017), pp. 195–201. ISSN: 1879-2057 (Electronic) 0001-4575 (Linking). DOI: https://doi.org/10.1016/j.aap.2017.03.017.

[29] Heinrich, H. W. “Industrial Accident Prevention. A Scientific Approach.” In: (1941).

[30] Reason, J. “Human error: models and management”. In: *Bmj* 320.7237 (2000), pp. 768–770.

[31] Wiegmann, D. A., Shappell, S. A. “Human Error Analysis of Commercial Aviation Accidents Using the Human Factors Analysis and Classification System (HFACS)”. In: (DOT/FAA/AM-01/3, Feb. 1, 2001). URL: https://rosap.ntl.bts.gov/view/dot/15409.

[32] Rasmussen, J. “Risk management in a dynamic society: a modelling problem”. In: *Safety science* 27.2-3 (1997), pp. 183–213.

[33] Leveson, N. “A new accident model for engineering safer systems”. In: *Safety science* 42.4 (2004), pp. 237–270.

[34] Xue, Y., Fu, G. “A modified accident analysis and investigation model for the general aviation industry: Emphasizing on human and organizational factors”. In: *Journal of Safety Research* 67 (2018), pp. 1–15. ISSN: 1879-1247 (Electronic) 0022-4375 (Linking). DOI: https://doi.org/10.1016/j.jsr.2018.09.008.

[35] Chi, C. F., Sigmund, D., Lin, Y. C., Drury, C. G. “The development of a scenario-based human-machine-environment-procedure (HMEP) classification scheme for the root cause analysis of helicopter accidents”. In: *Applied Ergonomics* 103 (2022). ISSN: 0003-6870. DOI: https://doi.org/10.1016/j.apergo.2022.103771.

[36] Vempati, L., Woods, S., Solano, R. C. “Qualitative Analysis of General Aviation Pilots’ Aviation Safety Reporting System Incident Narratives Using the Human Factors Analysis and Classification System”. In: *The International Journal of Aerospace Psychology* 33.3 (2023), pp. 182–196. ISSN: 2472-1840 2472-1832. DOI: https://doi.org/10.1080/24721840.2023.2232387.

[37] Fultz, A. J., Ashley, W. S. “Fatal weather-related general aviation accidents in the United States”. In: *Physical Geography* 37.5 (2016), pp. 291–312.

[38] Voogt, A., Louteiro, K. “Nose-Over and Nose-Down Accidents in General Aviation: Tailwheels and Aging Airplanes”. In: *Safety* 10.2 (2024), p. 39.

[39] Rostykus, P. S., Cummings, P., Mueller, B. A. “Risk factors for pilot fatalities in general aviation airplane crash landings”. In: *JAMA* 280.11 (1998), pp. 997–999.

[40] Huang, C. “Further improving general aviation flight safety: analysis of aircraft accidents during takeoff”. In: *Collegiate Aviation Review International* 38.1 (2020).

[41] Burgess, S., Boyd, S., Boyd, D. D. “Fatal general aviation accidents in furtherance of business (1996–2015): rates, risk factors, and accident causes”. In: *Journal of Aviation Technology and Engineering* 8.1 (2018), p. 11.

[42] O'Hare, D., Owen, D. "Cross-country VFR crashes: pilot and contextual factors." In: *Aviation, Space, and Environmental Medicine* 73.4 (2002), pp. 363–366.

[43] Boyd, D. D. "Causes and risk factors for fatal accidents in non-commercial twin engine piston general aviation aircraft". In: *Accident Analysis and Prevention* 77 (2015), pp. 113–119.

[44] Hinkelbein, J., Hippler, C., Liebold, F., Schmitz, J., Rothschild, M., Schick, V. "A new scoring system to predict fatal accidents in General Aviation and to facilitate emergency control centre response". In: *Scientific Reports* 14.1 (2024), p. 27969. ISSN: 2045-2322. DOI: 10.1038/s41598-024-77994-3.

[45] Jia, Q., Fu, G., Xie, X., Xue, Y., Hu, S. "Enhancing accident cause analysis through text classification and accident causation theory: A case study of coal mine gas explosion accidents". In: *Process Safety and Environmental Protection* 185 (2024), pp. 989–1002. ISSN: 09575820. DOI: https://doi.org/10.1016/j.psep.2024.03.066.

[46] Goh, Y. M., Ubeynarayana, C. "Construction accident narrative classification: An evaluation of text mining techniques". In: *Accident Analysis and Prevention* 108 (2017), pp. 122–130.

[47] Zhang, X., Srinivasan, P., Mahadevan, S. "Sequential deep learning from NTSB reports for aviation safety prognosis". In: *Safety Science* 142 (2021). ISSN: 09257535. DOI: 10.1016/j.ssci.2021.105390.

[48] Luo, X., Li, X., Song, X., Liu, Q. "Convolutional neural network algorithm–based novel automatic text classification framework for con-

struction accident reports". In: *Journal of Construction Engineering and Management* 149.12 (2023), p. 04023128.

[49] Ahmadi, E., Muley, S., Wang, C. "Automatic Construction Accident Report Analysis Using Large Language Models (LLMs)". In: *Journal of Intelligent Construction* (2024).

[50] Chen, X., Mao, X., Guo, Q., Wang, L., Zhang, S., Chen, T. "RareBench: Can LLMs Serve as Rare Diseases Specialists?" In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2024, pp. 4850–4861.

[51] Jose, S., Nguyen, K. T., Medjaher, K., Zemouri, R., Lévesque, M., Tahan, A. "Advancing multimodal diagnostics: Integrating industrial textual data and domain knowledge with large language models". In: *Expert Systems with Applications* 255 (2024), p. 124603.

[52] Qiu, Y., Jin, Y. "ChatGPT and finetuned BERT: A comparative study for developing intelligent design support systems". In: *Intelligent Systems with Applications* 21 (2024), p. 200308.

[53] Mumtarin, M., Chowdhury, M. S., Wood, J. "Large Language Models in Analyzing Crash Narratives–A Comparative Study of ChatGPT, BARD and GPT-4". In: *arXiv preprint arXiv:2308.13563* (2023).

[54] Zheng, O., Wang, D., Wang, Z., Ding, S. "Chat-GPT Is on the Horizon: Could a Large Language Model Be Suitable for Intelligent Traffic Safety Research and Applications?" In: *arXiv preprint arXiv:2303.05382* (2023).

[55] Grigorev, A., Saleh, K., Ou, Y., Mihaita, A.-S. “Integrating Large Language Models for Severity Classification in Traffic Incident Management: A Machine Learning Approach”. In.

[56] Ren, T., Zhang, Z., Jia, B., Zhang, S. “Retrieval-Augmented Generation-aided causal identification of aviation accidents: A large language model methodology”. In: *Expert Systems with Applications* 278 (2025), p. 127306. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2025.127306. URL: https://www.sciencedirect.com/science/article/pii/S0957417425009285.

[57] Andrade, S., Walsh, H. “SafeAeroBERT: Towards a Safety-Informed Aerospace-Specific Language Model”. In: *In AIAA AVIATION 2023 Forum; American Institute of Aeronautics and Astronautics (AIAA): San Diego, CA, USA* (2023).

[58] Tikayat Ray, A., Bhat, A. P., White, R. T., Nguyen, V. M., Pinon Fischer, O. J., Mavris, D. N. “Examining the Potential of Generative Language Models for Aviation Safety Analysis: Case Study and Insights Using the Aviation Safety Reporting System (ASRS)”. In: *Aerospace* 10.9 (2023). ISSN: 2226-4310. DOI: https://doi.org/10.3390/aerospace10090770.

[59] Zhang, R., Wang, B., Zhang, J., Bian, Z., Feng, C., Ozbay, K. “When language and vision meet road safety: Leveraging multimodal large language models for video-based traffic accident analysis”. In: *Accident Analysis and Prevention* 219 (2025), p. 108077. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2025.108077. URL: https://www.sciencedirect.com/science/article/pii/S0001457525001630.

[60] Chowdhury, M. T. B. Z., Hossain, M. *Design and Application of Multi-modal Large Language Model Based System for End to End Automation of Accident Dataset Generation*. 2025. arXiv: 2505.00015 [cs.CL]. URL: https://arxiv.org/abs/2505.00015.

[61] Wang, L., Ren, Y., Jiang, H., Cai, P., Fu, D., Wang, T., Cui, Z., Yu, H., Wang, X., Zhou, H., Huang, H., Wang, Y. "AccidentGPT: Accident Analysis and Prevention from V2X Environmental Perception with Multi-modal Large Model". In: *ArXiv* abs/2312.13156 (2023).

[62] Wu, K., Li, W., Xiao, X. "AccidentGPT: Large Multi-Modal Foundation Model for Traffic Accident Analysis". In: *ArXiv* abs/2401.03040 (2024).

[63] Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Blin, K., Adauto, F. G., Kleiman-Weiner, M., Sachan, M., Schölkopf, B. *CLadder: Assessing Causal Reasoning in Language Models*. 2024. arXiv: 2312.04350 [cs.CL]. URL: https://arxiv.org/abs/2312.04350.

[64] Wu, A., Kuang, K., Zhu, M., Wang, Y., Zheng, Y., Han, K., Li, B., Chen, G.-H., Wu, F., Zhang, K. "Causality for Large Language Models". In: *ArXiv* abs/2410.15319 (2024). DOI: 10.48550/arXiv.2410.15319.

[65] Gendron, G., Rovzanec, J. M., Witbrock, M., Dobbie, G. "Counterfactual Causal Inference in Natural Language with Large Language Models". In: *ArXiv* abs/2410.06392 (2024). DOI: 10.48550/arXiv.2410.06392.

[66] Jin, Z. "Causality for Natural Language Processing". In: *ArXiv* abs/2504.14530 (2025). DOI: 10.48550/arXiv.2504.14530.

[67] Kim, Y., Abdelrahman, A. S., Abdel-Aty, M. "VRU-Accident: A Vision-Language Benchmark for Video Question Answering and Dense Captioning for Accident Scene Understanding". In: *arXiv preprint arXiv:2507.09815* (2025).

[68] Tan, F., Desai, J., Sengamedu, S. H. "Enhancing Fact Verification with Causal Knowledge Graphs and Transformer-Based Retrieval for Deductive Reasoning". In: *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)* (2024). DOI: 10.18653/v1/2024.fever-1.20.

[69] Cohrs, K.-H., Varando, G., Díaz, E., Sitokonstantinou, V., Camps-Valls, G. "Large Language Models for Constrained-Based Causal Discovery". In: *ArXiv* abs/2406.07378 (2024). DOI: 10.48550/arXiv.2406.07378.

[70] Chen, M., Ma, Y., Song, K., Cao, Y., Zhang, Y., Li, D. "Learning to teach large language models logical reasoning". In: *arXiv preprint arXiv:2310.09158* (2023).

[71] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R. "Training verifiers to solve math word problems". In: *arXiv preprint arXiv:2110.14168* (2021).

[72] Creswell, A., Shanahan, M., Higgins, I. "Selection-inference: Exploiting large language models for interpretable logical reasoning". In: *arXiv preprint arXiv:2205.09712* (2022).

[73] Jiang, X., Wu, Y., Zhang, C., Zhang, Y. "DRoC: Elevating Large Language Models for Complex Vehicle Routing via Decomposed Retrieval

of Constraints". In: *The Thirteenth International Conference on Learning Representations* (2025). URL: https://openreview.net/forum?id=s9zoyICZ4k.

[74] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[75] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y. "ReAct: Synergizing Reasoning and Acting in Language Models". In: *ArXiv* abs/2210.03629 (2022).

[76] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.

[77] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., Narasimhan, K. "Tree of thoughts: Deliberate problem solving with large language models". In: *Advances in Neural Information Processing Systems* 36 (2024).

[78] Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P. "Graph of thoughts: Solving elaborate problems with large language models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 16. 2024, pp. 17682–17690.

[79] Bi, Z., Zhang, N., Jiang, Y., Deng, S., Zheng, G., Chen, H. "When Do Program-of-Thought Works for Reasoning?" In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 16. 2024, pp. 17691–17699.

[80] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., Iwasawa, Y. "Large language models are zero-shot reasoners". In: *Advances in neural information processing systems* 35 (2022), pp. 22199–22213.

[81] Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K.-W., Lim, E.-P. "Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models". In: *arXiv preprint arXiv:2305.04091* (2023).

[82] Zhang, Z., Zhang, A., Li, M., Smola, A. "Automatic chain of thought prompting in large language models". In: *arXiv preprint arXiv:2210.03493* (2022).

[83] Sun, X., Zhang, K., Liu, Q., Bao, M., Chen, Y. "Harnessing domain insights: A prompt knowledge tuning method for aspect-based sentiment analysis". In: *Knowledge-Based Systems* 298 (2024), p. 111975.

[84] Chen, M., Zhang, H., Wan, C., Wei, Z., Xu, Y., Wang, J., Gu, X. "On the effectiveness of large language models in domain-specific code generation". In: *arXiv preprint arXiv:2312.01639* (2023).

[85] Hong, S., Zheng, X., Chen, J. P., Cheng, Y., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C. "MetaGPT: Meta Programming for Multi-Agent Collaborative Framework". In: *ArXiv* abs/2308.00352 (2023). DOI: 10.48550/arXiv.2308.00352.

[86] Moncada-Ramirez, J., Matez-Bandera, J., Gonzalez-Jimenez, J., Ruiz-Sarmiento, J. "Agentic Workflows for Improving Large Language Model Reasoning in Robotic Object-Centered Planning". In: *Robotics* (2025). DOI: 10.3390/robotics14030024.

[87] Yu, C.-E., Jalaian, B., Bastian, N. D. "Mitigating Large Vision-Language Model Hallucination at Post-hoc via Multi-agent System". In: *Proceedings of the AAAI Symposium Series* (2024). DOI: 10.1609/aaaiss.v4i1.31780.

[88] Duan, Z., Wang, J. "Enhancing Multi-Agent Consensus Through Third-Party LLM Integration: Analyzing Uncertainty and Mitigating Hallucinations in Large Language Models". In: *2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE)* (2024), pp. 2222–2227. DOI: 10.1109/ICAACE65325.2025.11019272.

[89] Yang, Y., Ma, Y., Feng, H., Cheng, Y., Han, Z. "Minimizing Hallucinations and Communication Costs: Adversarial Debate and Voting Mechanisms in LLM-Based Multi-Agents". In: *Applied Sciences* (2025). DOI: 10.3390/app15073676.

[90] Sun, Z., Zang, X., Zheng, K., Song, Y., Xu, J., Zhang, X., Yu, W., Song, Y., Li, H. *ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability*. 2025. arXiv: 2410.11414 [cs.CL]. URL: https://arxiv.org/abs/2410.11414.

[91] Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., Zhang, T. *RAGTruth: A Hallucination Corpus for Developing Trustwor-*

*thy Retrieval-Augmented Language Models*. 2024. arXiv: 2401.00396 [cs.CL]. URL: https://arxiv.org/abs/2401.00396.

[92] Murphy, A., Rizvi, M. S. Z., Haussmann, A., Nie, P., Liu, G., Gema, A. P., Minervini, P. *An Analysis of Decoding Methods for LLM-based Agents for Faithful Multi-Hop Question Answering*. 2025. arXiv: 2503.23415 [cs.CL]. URL: https://arxiv.org/abs/2503.23415.

[93] Perboli, G., Gajetti, M., Fedorov, S., Giudice, S. L. "Natural Language Processing for the identification of Human factors in aviation accidents causes: An application to the SHEL methodology". In: *Expert Systems with Applications* 186 (2021), p. 115694.

[94] Shappell, S. "The Human Factors Analysis and Classification System-HFACS". In: (2000).

[95] Defense, D. o. *HFACS - Human Factors Analysis and Classification System 8.0 Handbook*. 2022. ISBN: 978-1-68423-195-9.

[96] Ergai, A., Cohen, T., Sharp, J., Wiegmann, D., Gramopadhye, A., Shappell, S. "Assessment of the Human Factors Analysis and Classification System (HFACS): Intra-rater and inter-rater reliability". In: *Safety Science* 82 (2016), pp. 393–398. ISSN: 09257535. DOI: https://doi.org/10.1016/j.ssci.2015.09.028.

[97] Lu, M., Gao, F., Tang, X., Chen, L. "Analysis and prediction in SCR experiments using GPT-4 with an effective chain-of-thought prompting strategy". In: *Iscience* 27.4 (2024).

[98] Tikayat Ray, A., Bhat, A. P., White, R. T., Nguyen, V. M., Pinon Fischer, O. J., Mavris, D. N. "Examining the Potential of Generative Language Models for Aviation Safety Analysis: Case Study and Insights Using the Aviation Safety Reporting System (ASRS)". In: *Aerospace* 10.9 (2023). ISSN: 2226-4310. DOI: 10.3390/aerospace10090770. URL: https://www.mdpi.com/2226-4310/10/9/770.

[99] Chen, L., Xu, J., Wu, T., Liu, J. "Information Extraction of Aviation Accident Causation Knowledge Graph: An LLM-Based Approach". In: *Electronics* 13.19 (2024). ISSN: 2079-9292. DOI: 10.3390/electronics13193936. URL: https://www.mdpi.com/2079-9292/13/19/3936.

[100] Liu, Q., Li, F., Ng, K. K., Han, J., Feng, S. "Accident investigation via LLMs reasoning: HFACS-guided Chain-of-Thoughts enhance general aviation safety". In: *Expert Systems with Applications* 269 (2025), p. 126422. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2025.126422. URL: https://www.sciencedirect.com/science/article/pii/S0957417425000442.

[101] NTSB, Web Page. [accessed 8th November 2024]. 2024. URL: https://www.ntsb.gov/investigations/process/Pages/default.aspx.

[102] Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., Wang, P. "K-BERT: Enabling Language Representation with Knowledge Graph". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.03 (2020), pp. 2901–2908. DOI: 10.1609/aaai.v34i03.5681. URL: https://ojs.aaai.org/index.php/AAAI/article/view/5681.

[103] Rostamabadi, A., Jahangiri, M., Zarei, E., Kamalinia, M., Banaee, S., Samaei, M. R. "A Novel Fuzzy Bayesian Network-HFACS (FBN-HFACS) model for analyzing Human and Organization Factors (HOFs) in process accidents". In: *Process Safety and Environmental Protection* 132 (2019), pp. 59–72. ISSN: 0957-5820. DOI: https://doi.org/10.1016/j.psep.2019.08.012. URL: https://www.sciencedirect.com/science/article/pii/S0957582019306007.

[104] Lewin, K. *Principles of topological psychology*. Read Books Ltd, 2013.

[105] Ahlstrom, U., Ohneiser, O., Caddigan, E. "Portable Weather Applications for General Aviation Pilots". In: *Human Factors* 58.6 (2016), pp. 864–885. ISSN: 1547-8181 (Electronic) 0018-7208 (Linking). DOI: https://doi.org/10.1177/0018720816641783.

[106] Boyd, D. D. "Occupant injury severity in general aviation accidents involving excessive landing airspeed". In: *Aerospace Medicine and Human Performance* 90.4 (2019), pp. 355–361.

[107] Knecht, W., Harris, H., Shappell, S. "General aviation pilot takeoff into adverse weather: the "go/no-go" equation". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 48. 3. SAGE Publications Sage CA: Los Angeles, CA. 2004, pp. 340–344.

[108] Mannering, F. "Temporal instability and the analysis of highway accident data". In: *Analytic Methods in Accident Research* 17 (2018), pp. 1–13. ISSN: 22136657. DOI: 10.1016/j.amar.2017.10.002.

[109] Ahmed, S. S., Pantangi, S. S., Eker, U., Fountas, G., Still, S. E., Anastasopoulos, P. C. "Analysis of safety benefits and security concerns from

the use of autonomous vehicles: A grouped random parameters bivariate probit approach with heterogeneity in means". In: *Analytic Methods in Accident Research* 28 (2020), p. 100134.

[110] Abay, K. A., Paleti, R., Bhat, C. R. "The joint analysis of injury severity of drivers in two-vehicle crashes accommodating seat belt use endogeneity". In: *Transportation research part B: methodological* 50 (2013), pp. 74–89.

[111] Wang, C., Abdel-Aty, M., Han, L. "Effects of speed difference on injury severity of freeway rear-end crashes: Insights from correlated joint random parameters bivariate probit models and temporal instability". In: *Analytic Methods in Accident Research* 42 (2024), p. 100320.

[112] Song, D., Yang, X., Yang, Y., Cui, P., Zhu, G. "Bivariate joint analysis of injury severity of drivers in truck-car crashes accommodating multilayer unobserved heterogeneity". In: *Accident Analysis and Prevention* 190 (2023), p. 107175.

[113] Song, P., Sze, N., Chen, S., Labi, S. "Correcting for endogeneity of crash type in crash injury severity at highway ramp areas". In: *Accident Analysis and Prevention* 208 (2024), p. 107785.

[114] Mannering, F. L., Shankar, V., Bhat, C. R. "Unobserved heterogeneity and the statistical analysis of highway accident data". In: *Analytic Methods in Accident Research* 11 (2016), pp. 1–16. ISSN: 22136657. DOI: 10.1016/j.amar.2016.04.001.

[115] Liu, Q., Li, F., Ng, K. K. "Unveiling the determinants of injury severities across age groups and time: A deep dive into the unobserved heterogeneity

among pedestrian crashes". In: *Analytic Methods in Accident Research* 43 (2024), p. 100336. ISSN: 2213-6657. DOI: https://doi.org/10.1016/j.amar.2024.100336. URL: https://www.sciencedirect.com/science/article/pii/S2213665724000204.

[116] Alnawmasi, N., Mannering, F. "An analysis of day and night bicyclist injury severities in vehicle/bicycle crashes: A comparison of unconstrained and partially constrained temporal modeling approaches". In: *Analytic Methods in Accident Research* 40 (2023), p. 100301. ISSN: 22136657. DOI: 10.1016/j.amar.2023.100301.

[117] Association, G. A. M. *2019 Databook*. Web Page. [accessed 19th November 2024]. Washington, DC, 2020. URL: https://gama.aero/wp-content/uploads/GAMA_2019Databook_Final-2020-03-20.pdf.

[118] Congress, *H.R.658 - FAA Modernization and Reform Act of 2012*. [accessed 10th November 2024]. 2012. URL: https://www.congress.gov/bill/112th-congress/house-bill/658.

[119] FAA, *Integration of Civil Unmanned Aircraft Systems (UAS) in the National Airspace System [NAS) Roadmap*. Web Page. [accessed 9th February 2025]. Washington, DC, 2013. URL: https://www.faa.gov/sites/faa.gov/files/uas/resources/policy_library/2019_UAS_Civil_Integration_Roadmap_third_edition.pdf.

[120] Congress, *H.R.1848 - Small Airplane Revitalization Act of 2013*. Webpage. [accessed 9th November 2024]. 2013. URL: https://www.congress.gov/bill/113th-congress/house-bill/1848.

[121] FAA, *U.S. Civil Airmen Statistics*. Web Page. [accessed 9th November 2024]. 2024. URL: https://www.faa.gov/data_research/aviation_data_statistics/civil_airmen_statistics.

[122] Greene, W. H. *Econometric analysis*. Pretence Hall, 2003.

[123] Washington, S., Karlaftis, M. G., Mannering, F., Anastasopoulos, P. *Statistical and econometric methods for transportation data analysis*. Chapman and Hall/CRC, 2020.

[124] Liu, D., Ding, C., Bold, D., Bouvier, M., Lu, J., Shickel, B., Jabaley, C. S., Zhang, W., Park, S., Young, M. J. "Evaluation of General Large Language Models in Contextually Assessing Semantic Concepts Extracted from Adult Critical Care Electronic Health Record Notes". In: *arXiv preprint arXiv:2401.13588* (2024).

[125] Christofides, L. N., Stengos, T., Swidinsky, R. "On the calculation of marginal effects in the bivariate probit model". In: *Economics Letters* 54.3 (1997), pp. 203–208.

[126] Behnood, A., Mannering, F. L. "The temporal stability of factors affecting driver-injury severities in single-vehicle crashes: Some empirical evidence". In: *Analytic Methods in Accident Research* 8 (2015), pp. 7–32. ISSN: 22136657. DOI: 10.1016/j.amar.2015.08.001.

[127] Yan, X., He, J., Zhang, C., Wang, C., Ye, Y., Qin, P. "Temporal instability and age differences of determinants affecting injury severities in nighttime crashes". In: *Analytic Methods in Accident Research* 38 (2023), p. 100268. ISSN: 22136657. DOI: 10.1016/j.amar.2023.100268.

[128] Islam, M. "The effect of motorcyclists' age on injury severities in single-motorcycle crashes with unobserved heterogeneity". In: *Journal of Safety Research* 77 (2021), pp. 125–138.

[129] Shao, B. S., Guindani, M., Boyd, D. D. "Fatal accident rates for instrument-rated private pilots". In: *Aviation, Space, and Environmental Medicine* 85.6 (2014), pp. 631–637.

[130] Li, G., Baker, S. P., Qiang, Y., Grabowski, J. G., McCarthy, M. L. "Driving-while-intoxicated history as a risk marker for general aviation pilots". In: *Accident Analysis and Prevention* 37.1 (2005), pp. 179–184.

[131] Hinkelbein, J., Spelten, O., Neuhaus, C., Hinkelbein, M., Özgür, E., Wetsch, W. A. "Injury severity and seating position in accidents with German EMS helicopters". In: *Accident Analysis and Prevention* 59 (2013), pp. 283–288.

[132] **Lyu, M.**, Li, F., Qu, X., Li, Q. "Flashlight model: integrating attention distribution and attention resources for pilots' visual behaviour analysis and performance prediction". In: *International Journal of Industrial Ergonomics* 103 (2024), p. 103630.

[133] Silagyi, 2., Liu, D. "Prediction of severity of aviation landing accidents using support vector machine models". In: *Accident Analysis and Prevention* 187 (2023), p. 107043. ISSN: 1879-2057 (Electronic) 0001-4575 (Linking). DOI: 10.1016/j.aap.2023.107043.

[134] Li, Z., Li, F., **Lyu, M.** "Tracking the Unseen and Unaware: Deciphering Controllers' Detection Failures to Warnings Through Eye-Tracking Met-

rics". In: *International Journal of Human–Computer Interaction* (2025), pp. 1–20.

[135] Islam, M., Alogaili, A., Mannering, F., Maness, M. "Evidence of sample selectivity in highway injury-severity models: The case of risky driving during COVID-19". In: *Analytic Methods in Accident Research* 38 (2023), p. 100263. ISSN: 22136657. DOI: 10.1016/j.amar.2022.100263.

# Appendices

## Appendix I. F1 score comparison of four LLMs

Table 6.1: F1 score comparison of four LLMs

| Matrix | Prompt | AE100 | AE200 | AD000 | PC100 | PC200 | PC300 | PE100 | PE200 | PP100 | PT100 | Ave | Ave+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GPT-4o | Basic zero-shot | 88.42 | 58.57 | 6.90 | 21.88 | 28.32 | 47.24 | 56.17 | 19.18 | 31.76 | 27.32 | 38.58 | 42.10 |
| | Basic few-shot | 89.58 | 58.57 | 11.32 | 20.06 | 43.59 | 31.09 | 60.44 | 23.08 | 30.30 | 36.11 | 40.41 | 43.65 |
| | Auto-CoT | 81.69 | 59.97 | 19.51 | 18.66 | 42.55 | 32.56 | 65.10 | 23.08 | 31.45 | 37.32 | 41.19 | 43.60 |
| | Plan-and-solve | 85.84 | 57.00 | 17.65 | 17.13 | 30.14 | 51.43 | 60.92 | 19.23 | 14.29 | 23.61 | 37.72 | 39.95 |
| | HFACS-CoT | 89.03 | **62.46** | 0.00 | 27.45 | **56.41** | 58.00 | 64.33 | 32.88 | 38.04 | 36.78 | 46.54 | 51.71 |
| | HFACS-CoT+ | **89.80** | 61.95 | **30.00** | **33.02** | 47.06 | **58.82** | **73.97** | **30.36** | **40.00** | **49.18** | **51.31** | **53.68** |
| Qwen-turbo | Basic zero-shot | 90.02 | 36.13 | 0.00 | 12.90 | 0.82 | 2.25 | 49.01 | 0.00 | 1.75 | 19.39 | 21.23 | 23.59 |
| | Basic few-shot | 92.33 | 43.62 | 0.00 | 14.50 | 3.31 | 9.09 | 48.30 | 0.00 | 10.43 | 21.12 | 24.27 | 26.97 |
| | Auto-CoT | 77.08 | 37.26 | 0.00 | 16.36 | 3.03 | **17.54** | 41.85 | 0.00 | 3.28 | **27.87** | 22.43 | 24.92 |
| | Plan-and-solve | 86.46 | 43.53 | 0.00 | 19.90 | 0.00 | 9.23 | **55.63** | 0.00 | 0.00 | 27.81 | 24.26 | 26.95 |
| | HFACS-CoT | 92.39 | 45.02 | 0.00 | 16.75 | 3.08 | 7.14 | 49.23 | **2.58** | 18.73 | 24.63 | 25.96 | 28.84 |
| | HFACS-CoT+ | **92.98** | **69.63** | 0.00 | **21.65** | **6.12** | 16.13 | 55.44 | 1.92 | **22.13** | 24.10 | **31.01** | **34.46** |
| GPT-3.5-turbo | Basic zero-shot | 88.77 | 46.46 | 0.00 | 12.13 | 0.00 | 6.13 | **57.18** | 0.00 | 0.00 | 16.60 | 22.73 | 25.25 |
| | Basic few-shot | 89.21 | **49.05** | 0.00 | **14.74** | 0.00 | 11.11 | 52.11 | 0.00 | 0.00 | **19.35** | 23.81 | 26.18 |
| | Auto-CoT | **89.35** | 23.72 | 0.00 | 12.64 | **3.85** | 0.00 | 51.88 | 0.00 | 0.00 | 12.80 | 19.42 | 21.58 |
| | Plan-and-solve | 89.20 | 12.19 | 0.00 | 12.78 | 0.00 | 0.00 | 51.72 | 0.00 | 0.00 | 11.33 | 17.72 | 19.69 |
| | HFACS-CoT | 87.90 | 46.89 | 0.00 | 2.94 | 0.00 | **46.89** | 51.75 | 0.00 | 2.94 | 17.04 | **25.64** | **28.48** |
| | HFACS-CoT+ | 89.15 | 46.52 | 0.00 | 12.53 | 1.89 | 9.59 | 51.55 | **1.42** | **12.95** | 12.46 | 23.81 | 26.45 |
| Llama-3-8b | Basic zero-shot | 87.11 | 35.03 | 0.00 | 21.28 | 42.42 | 15.79 | 38.14 | 38.89 | 41.03 | 33.80 | 35.35 | 39.28 |
| | Basic few-shot | 92.02 | 39.32 | **10.53** | 23.45 | 59.26 | 13.64 | 40.98 | 46.15 | 36.84 | 29.63 | 39.18 | 42.36 |
| | Auto-CoT | 91.57 | 36.36 | 10.00 | 20.78 | 54.55 | 7.69 | 41.75 | 40.00 | **43.90** | 32.10 | 37.87 | 40.97 |
| | Plan-and-solve | 76.60 | **42.05** | 2.90 | 11.61 | 45.45 | 11.76 | 31.14 | 48.28 | 0.00 | 14.06 | 28.38 | 31.22 |
| | HFACS-CoT | 86.22 | 35.60 | 0.00 | **28.07** | **68.75** | 16.67 | **42.65** | **57.14** | 37.21 | **40.00** | **41.23** | **45.81** |
| | HFACS-CoT+ | **93.62** | 33.44 | 0.00 | 17.02 | 7.34 | **18.18** | 30.94 | 13.22 | 13.33 | 25.00 | 25.21 | 28.01 |

Note: All prompts achieve higher ave+ f1 scores on GPT-4 than on other LLMs. Our proposed prompts outperform the baselines in average f1 and average+ f1 scores for GPT-4, Qwen-Turbo, and GPT-3.5-Turbo. HFACS-CoT+ performs worst on Llama-3-8b, indicating it does not significantly enhance reasoning ability unless used with a model of sufficient scale, and multiple API calls in small-scale models may lead to context loss and reduced performance.

# Appendix II. F1 score comparison of four LLMs

Table 6.2: Recall comparison of four LLMs

| Matrix | Prompt | AE100 | AE200 | AD000 | PC100 | PC200 | PC300 | PE100 | PE200 | PP100 | PT100 | Ave | Ave+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GPT-4o | Basic zero-shot | 88.36 | 82.96 | 11.76 | 48.61 | 50.00 | 50.00 | 84.62 | 35.00 | 26.47 | 73.53 | 55.13 | 59.95 |
| | Basic few-shot | **92.85** | 91.69 | 17.65 | 47.95 | 53.13 | 50.00 | 87.18 | 45.00 | 24.51 | 57.35 | 56.73 | 61.07 |
| | Auto-CoT | 71.67 | 67.73 | **23.53** | 53.42 | 62.50 | 23.33 | 82.91 | 30.00 | 24.51 | **77.94** | 51.75 | 54.89 |
| | Plan-and-solve | 81.21 | 71.57 | 17.65 | 67.12 | 34.38 | 45.00 | 67.95 | 25.00 | 9.80 | 72.06 | 49.17 | 52.68 |
| | HFACS-CoT | 87.10 | **93.29** | 0.00 | 57.53 | 68.75 | 50.00 | **90.17** | 60.00 | 38.04 | 70.59 | 60.62 | 67.35 |
| | HFACS-CoT+ | 88.80 | 92.33 | 17.65 | **71.23** | **75.00** | **75.00** | 88.03 | **85.00** | **57.84** | 66.18 | **69.21** | **74.93** |
| Qwen-turbo | Basic zero-shot | 94.36 | 31.27 | 0.00 | 27.59 | 12.50 | 2.78 | 70.62 | 0.00 | 1.05 | 25.81 | 26.60 | 29.55 |
| | Basic few-shot | **99.11** | 66.55 | 0.00 | 75.86 | 25.00 | 16.67 | 87.68 | 0.00 | 6.32 | 54.84 | 43.20 | 48.00 |
| | Auto-CoT | 91.57 | 37.50 | 0.00 | 52.94 | **37.50** | 14.29 | 32.70 | 0.00 | 1.85 | 31.48 | 27.05 | 30.05 |
| | Plan-and-solve | 83.95 | 42.18 | 0.00 | 65.52 | 0.00 | 8.33 | 57.08 | 0.00 | 0.00 | 33.87 | 29.09 | 32.33 |
| | HFACS-CoT | **99.11** | 37.82 | 0.00 | **86.21** | 28.57 | **19.44** | 83.41 | **40.00** | 26.32 | 80.65 | 51.05 | 56.72 |
| | HFACS-CoT+ | 96.29 | **68.36** | 0.00 | 65.52 | **37.50** | 13.89 | **90.52** | **40.00** | **57.89** | **91.94** | **56.19** | **62.43** |
| GPT-3.5-turbo | Basic zero-shot | 99.11 | 68.50 | 0.00 | 55.36 | 0.00 | 11.90 | 75.95 | 0.00 | 0.00 | 70.69 | 38.15 | 42.39 |
| | Basic few-shot | **100.00** | 75.98 | 0.00 | 53.57 | 0.00 | **73.81** | 63.57 | 0.00 | 0.00 | 25.86 | 39.28 | 43.64 |
| | Auto-CoT | 98.96 | 14.57 | 0.00 | 98.21 | 12.50 | 0.00 | 78.35 | 0.00 | 0.00 | 36.21 | 33.88 | 37.64 |
| | Plan-and-solve | 98.66 | 6.69 | 0.00 | 98.21 | 0.00 | 0.00 | 77.66 | 0.00 | 0.00 | 29.31 | 31.05 | 34.51 |
| | HFACS-CoT | 95.40 | 51.97 | 0.00 | 3.57 | 0.00 | 51.97 | 91.41 | 0.00 | 1.72 | 32.76 | 32.88 | 36.53 |
| | HFACS-CoT+ | **100.00** | **100.00** | 0.00 | **100.00** | **100.00** | 9.59 | **100.00** | **100.00** | **100.00** | **94.83** | **80.44** | **89.38** |
| Llama-3-8b | Basic zero-shot | 81.68 | 53.45 | 0.00 | 50.00 | 41.18 | 10.34 | 61.67 | 35.00 | 44.44 | 52.17 | 42.99 | 47.77 |
| | Basic few-shot | 90.84 | 76.67 | **7.14** | 54.84 | 47.06 | 10.34 | 70.00 | 45.00 | 38.89 | 34.78 | **52.05** | 52.05 |
| | Auto-CoT | 88.52 | 56.67 | 6.25 | 51.61 | 52.94 | 7.14 | 74.14 | 35.00 | **54.17** | **54.17** | 51.95 | 51.95 |
| | Plan-and-solve | 66.91 | 60.66 | 6.25 | 58.82 | 58.82 | **17.24** | 44.83 | 35.00 | 0.00 | 39.13 | 35.70 | 38.97 |
| | HFACS-CoT | 80.22 | 56.67 | 0.00 | 51.61 | **64.71** | 10.34 | **75.00** | **60.00** | 44.44 | 47.83 | 49.08 | **54.54** |
| | HFACS-CoT+ | **95.54** | **86.89** | 0.00 | **62.50** | 47.06 | 13.79 | 74.14 | 40.00 | 16.67 | 39.13 | 47.57 | 52.86 |

Note:HFACS-CoT+ significantly outperforms other prompts in terms of average F1 score across various LLMs. HFACS-CoT+ on GPT-3.5-turbo achieves a recall rate of 100% on seven sub-tasks, which, when compared to its relatively low F1 score, suggests that GPT-3.5-turbo under HFACS-CoT+ tends to answer ”yes” to the majority of questions (sub-tasks).

# Appendix III. HFACS-CoT prompt

Listing 6.1: Python code snippet

```
import os
import pandas as pd
from openai import OpenAI
client = OpenAI(api_key=os.environ.get("OPENAI_API_KEY"))


def get_completion(prompt, model="gpt-4o"):
    messages = [{"role": "user", "content": prompt}]
    response = client.chat.completions.create(
```

```
        model=model,
        messages=messages,
        temperature=0.1,
        max_tokens=1024,
        n=1
    )

    return response.choices[0].message.content
text=f"""

"The pilot reported that...



 """



background_info =f"""
    UNSAFE ACTS:
1.Performance/Skill Based Errors (AE100)
AE101 Unintended Activation or Deactivation: Mistaken activation/deactivation of
    equipment without intent, incorrect action performance.
AE102 Procedure or Checklist Not Followed Correctly: Errors in following
    procedures or checklists, such as wrong execution sequence or omissions,
    incorrect action selection, incorrect action sequence, lack of action,
    forgotten action/omission, forgotten preflight inspection, incorrect fuel
    management, wrong modification/alteration/Installation/authorized
    maint/repair.
...
"""

prompt =  f'''
As an expert in general aviation accident investigations, you are to deduce the
    most likely types of unsafe acts by pilots that led to a mishap or near
    mishap, along with their preconditions from the narrative: '{text}', given
    by pilots or other witnesses following the accident.
Please analyze according to the definitions and classifications of unsafe acts
    and preconditions given: {background_info}
```

```
Your analysis should strictly rely on direct statements from pilots or
    witnesses. Avoid making speculative guesses that extend beyond the provided
    testimony.

You need to use the following steps for a detailed accident analysis and answer
    all the questions below  as sequence and answer the most applicable code if
    you answer yes:

Question A1: Whether the mishap individual do or fail to do that allowed the
    near-miss or mishap to occur? If yes, go to Question A2. If no, exit.

Question A2: Was the error a Performance Based Error? If yes, answer the most
    applicable error code, then go to Question A3. If no, go to Question A3
    directly.

Question A3: Was the error a Judgment and Decision Making Error? If yes, answer
    the most applicable error code and then start analyzing the preconditions
    associated with the individuals unsafe act. Only when you answered no to
    both A2 and A3, you need to go to Question A4, otherwise, you should start
    analyzing the preconditions associated with the individuals unsafe act
    directly.

Question A4: Was the act a Known Deviation? If yes, answer the most applicable
    code and then analyze the preconditions associated with the individual
    unsafe act.


Next, analyze the preconditions related to unsafe acts identified in A2-A4
    respectively.If you have identified one or more unsafe acts, you must
    identified at least one precondition of unsafe act(s).
If you have identified unsafe acts in both A2 and A3, you need to repeat P2-P10
    twice for unsafe act identified in A2 and A3 respectively
If you do not recognize any precondtion, report that the precondtion recognition
    result is none.

Question P1: Conclude how many unsafe acts have been identified (if you have
    identified unsafe acts in both A2 and A3, and the answer is two).
```

For the first identified unsafe act, you need consider P2-P10 in sequence to analyze its preconditions:

Question P2: Did the conditions of the mishap individual affect the actions of the mishap individual. If yes, proceed to Question P3. If no, skip to Question P6.Please note that if your answer about P2 is yes, and at least one of the answers to Question P3-P5 must be yes.

Question P3: Did a mental awareness condition of the mishap individual influence the unsafe act? If yes, answer which code is the most appropriate to support the unsafe act, then go to Question P4. If not, go to Question P4 directly.

Question P4: Did the mishap person state of mind influence the unsafe act? If yes, answer which code is the most appropriate to support the unsafe act, then go to Question P5. If not, go to Question P5 directly.

Question P5: Did the mishap person have a physical condition that negatively affected performance and influenced the unsafe act? If yes, answer which code is the most appropriate to support the unsafe act, then go to Question P6. If not, go to Question P6 directly.

Question P6: Did conditions of the operational environment affect the actions of the mishap individual? If yes for go to question P7. If no, go to question P9.Please note that if your answer about P6 is yes, and at least one of the answers to Question P7-P8 must be yes.

Question P7: Did conditions of the physical environment affect the actions of the mishap individual or team? If yes, answer which code is the most appropriate to support the unsafe act, then go to Question P8. If not, go to Question P8 directly.

Question P8: Did the technological environment (workspace design) negatively affect the performance of the mishap individual? If yes, answer which code is the most appropriate to support the unsafe act, then go to Question P9. If not, go to Question P9 directly.

```
Question P9: Did Communication/Mission Planning Conditions contribute to the
    unsafe act? If yes, answer which code is the most appropriate to support the
    unsafe act, then go to Question P10. If not, go to Question P10 directly.

Question P10: Did the certification/experience conditions of pilot contribute to
    the unsafe act? If yes, answer which code is the most appropriate to support
    the unsafe act



Note: Please answer all the questions step by step. For all your answer to the
    above questions, you need to response yes or no, and report the most
    applicable error code if your answer is yes, and need to report the reason
    or give more further information.
For the second identified unsafe act (If has), you need analyze its
    preconditions consider P2-P10 in sequence independently. If the analysis
    results are totally same as those for the first unsafe act, you just
    response simply that the answers to P2-P10 are totally same as for the first
    unsafe act. If not, you need to write down the detail result as for the
    analysis of preconditions of first unsafe act.

'''
print(get_completion(prompt))
```

# Appendix IV. HFACS-CoT+ prompt

Listing 6.2: Python code snippet

```
import os
import re
from openai import OpenAI

client = OpenAI(api_key=os.environ.get("openai_API_KEY_LI"))
```

```
# Define error codes and descriptions dictionary
AE100_dict = {
    "AE101": "Unintended Activation or Deactivation",
    "AE102": "Procedure or Checklist Not Followed Correctly",
    "AE103": "Over-Controlled/Under-Controlled Aircraft/Vehicle/Vessel or
        System",
    "AE105": "Breakdown in Visual Scan or Instrument Cross-check",
    "AE107": "Rushed or Delayed a Necessary Action",
    "AE108": "Misinterpreted/Misread Instrument",
}

AE100_info = """
Performance/Skill-Based Errors: ...
"""

AE200_dict = {
    "AE201": "Inadequate Real-Time Risk Assessment/Action",
    "AE205": "Ignored a Caution/Warning",
    "AE207": "Situational awareness/spatial disorientation",
}

AE200_info = """
Judgment and Decision-Making Errors: ...
"""

AD000_dict = {
    "AD001": "Performed Known Deviation (Work-Around)",
    "AD002": "Commits Routine/Widespread Known Deviation (Normalization of
        Deviance)",
    "AD003": "Extreme Lack of Discipline (Indiscipline)",
}

AD000_info = """
Known Deviations: ...
"""

PC100_info = """
```

```
Attention Conditions: ...
"""

PC200_info = """
State of Mind Conditions: ...
"""

PC300_info = """
Adverse Physiological Conditions: ...
"""

PE100_info = """
Physical Environment: ...
"""

PE200_info = """
Technological Environment: ...
"""

PP100_info = """
Communication/Mission Planning Conditions: ...
"""

PT100_info = """
Certification/Experience Conditions: ...
"""

# Helper functions
def answer_machine(prompt):
    chat_completion = client.chat.completions.create(
        messages=[{"role": "user", "content": prompt}],
        model="gpt-4o",
        temperature=0.1,
        max_tokens=1024,
        n=1
    )
    return chat_completion.choices[0].message.content
```

```
def extract_error_code(answer, pattern):
    match = re.search(pattern, answer)
    if match:
        first_char = match.group(1)
        error_code = match.group(3) if len(match.groups()) > 2 else None
        return first_char, error_code
    return None, None

def ask_question(text, condition_info, question, pattern):
    prompt = f"""
    The narrative given by pilots or other witnesses following the accident:
        '{text}'.
    The definitions and classifications of unsafe acts given: {condition_info}

    As an expert in general aviation accident investigation, you need to answer
        the questions below (Ensure that your analysis relies strictly on actual
        statements from pilots or witnesses and avoids speculative guessing.):
    Question: {question}
    """
    answer = answer_machine(prompt)
    print (answer)
    return extract_error_code(answer, pattern)


def ask_precondition_question(text, condition_info, question, code_list,
    combined_conditions=None):
    combined_text = ""
    if combined_conditions:
        combined_text = f"From the perspective of both {combined_conditions[0]}
            and {combined_conditions[1]}, "

    prompt = f"""
    Narrative given by pilots or other witnesses following the accident:
        '{text}'.
    The definitions and classifications of unsafe preconditions given:
        {condition_info}
```

```
    {combined_text}you need to answer the questions below (Ensure that your
        analysis relies strictly on actual statements from pilots or witnesses
        and avoids speculative guessing.):
    Question: {question}
    """
    answer = answer_machine(prompt)
    print (answer)
    code_pattern = "|".join(code_list)
    pattern = rf"(1|0)(.*?({code_pattern}))?"
    return extract_error_code(answer, pattern)

# Main program
if __name__ == "__main__":
    text = """
According to several eyewitnesses, including the current and previous owners,
    the pilot taxied out to the runway, returned to the hangar, and shut down
    the airplane's engine. The pilot reported to the previous owner that the red
    "water temp" indicator light illuminated. After discussing the need to open
    the cowling louvers, the pilot taxied to the end of the runway and began the
    takeoff roll. The airplane became airborne, immediately banked to the left,
    and then climbed to about 100 feet above ground level. The airplane
    continued into a 360-degree left turn and then nosed into the ground. The
    witnesses further reported that they heard the engine operating at or near
    full power until the airplane impacted the ground, with one describing the
    airplane as looking like a "lawn dart" until impact. Two other eyewitnesses
    reported that during taxi, the nose landing gear kept rising off the ground
    and that the pilot had difficulty maintaining ground contact with the nose
    landing gear.
"""

    q1_question = "Whether the mishap individual do or fail to do that allowed
        the near-miss or mishap to occur? If yes, please reply 1, if no or not
        sure, please reply 0. Then, please give the necessary analysis or
        reasoning for your choices."
    q1_pattern = r"(1|0)"
    q1_bool, _ = ask_question(text, AE100_info, q1_question, q1_pattern)
```

```
print(f"Q1 Answer is: {q1_bool}")

if q1_bool == "0":
    return "No further questions needed."

q2_question = "Was the error a Performance/Skill Based Error? If yes, please
    reply 1 and the most applicable error code in [AE101, AE102, AE103,
    AE105, AE107, AE108], if no or not sure, please reply 0. Then, please
    give the necessary analysis or reasoning for your choices."
q2_pattern = r"(1|0)(.*?(AE10[1-9]))?"
q2_bool, q2_code = ask_question(text, AE100_info, q2_question, q2_pattern)
print(f"Q2 Answer is: {q2_bool}, Q2 Code is: {q2_code}")
q3_question = "Was the error a Judgment and Decision Making Error? If yes,
    please reply 1 and provide the most applicable error code from [AE201,
    AE202, AE205, AE207]. If no or not sure, please reply 0. Then, please
    give the necessary analysis or reasoning for your choices. If the pilot
    deliberately performed an incorrect operation, it should be considered
    an AD000 rather than a Judgment and Decision Making Error."
q3_pattern = r"(1|0)(.*?(AE20[1-7]))?"
q3_bool, q3_code = ask_question(text, AE200_info, q3_question, q3_pattern)
print(f"Q3 Answer is: {q3_bool}, Q3 Code is: {q3_code}")

if q2_bool == "0" and q3_bool == "0":
    q4_question = "Was the act a Known Deviation? If yes, please reply 1 and
        the most applicable error code in [AD001, AD002, AD003], if no or
        not sure, please reply 0. Then, please give the necessary analysis
        or reasoning for your choices."
    q4_pattern = r"(1|0)(.*?(AD00[1-3]))?"
    q4_bool, q4_code = ask_question(text, AD000_info, q4_question,
        q4_pattern)
    print(f"Q4 Answer is: {q4_bool}, Q4 Code is: {q4_code}")
else:
    print("Q4 is not needed.")

print("Start analyzing the preconditions!")

if q2_bool == "1" and q3_bool == "1":
```

```
        unsafe_act0 = AE100_dict[q2_code]
        unsafe_act1 = AE200_dict[q3_code]
        combined_conditions = (unsafe_act0, unsafe_act1)
        print("Two unsafe acts have been identified. Both Performance/Skill
            Based Errors and Judgment and Decision-Making Errors")
    elif q2_bool == "1":
        unsafe_act0 = AE100_dict[q2_code]
        combined_conditions = (unsafe_act0, None)
        print("One unsafe act has been identified: Performance/Skill Based
            Error")
    elif q3_bool == "1":
        unsafe_act0 = AE200_dict[q3_code]
        combined_conditions = (unsafe_act0, None)
        print("One unsafe act has been identified: Judgment and Decision-Making
            Error")
    elif q4_bool == "1":
        unsafe_act0 = AD000_dict[q4_code]
        combined_conditions = (unsafe_act0, None)
        print("One unsafe act has been identified: Violation")


    print(f"Condition 1: {unsafe_act0}")
    if combined_conditions[1]:
        print(f"Condition 2: {combined_conditions[1]}")
    else:
        print("No Condition 2")

    # Define P-related questions and code lists
    p_questions = [
        ("P3", "Did a mental awareness condition of the mishap individual
            influence the unsafe act? If yes, please reply 1 and the most
            applicable error code; if no,  please only reply 0.Then,please give
            the necessary analysis or reasoning for your choices.", PC100_info,
            ["PC101", "PC102", "PC103", "PC104", "PC105", "PC107"]),
        ("P4", "Did a state of mind condition of the mishap individual influence
            the unsafe act? If yes, please reply 1 and the most applicable error
            code; if no, please only reply 0. Then,please give the necessary
```

```
            analysis or reasoning for your choices.", PC200_info, ["PC202",
            "PC203", "PC205", "PC206"]),
        ("P5", "Did a physical condition of the mishap individual influence the
            unsafe act? If yes, please reply 1 and the most applicable error
            code;if no, please only reply 0. Then, please give the necessary
            analysis or reasoning for your choices.", PC300_info, ["PC301",
            "PC302", "PC304", "PC305", "PC306", "PC307", "PC312", "PC317",
            "PC319", "PC320", "PC321"]),
        ("P7", "Did the physical environment negatively affect the performance
            of the mishap individual? If yes, please reply 1 and the most
            applicable error code; if no, please only reply 0. Then, please give
            the necessary analysis or reasoning for your choices.", PE100_info,
            ["PE101", "PE106", "PE108", "PE109", "PE110", "PE112", "PE113",
            "PE114"]),
        ("P8", "Did the technological environment negatively affect the
            performance of the mishap individual? If yes, please reply 1 and the
            most applicable error code; if no, please only reply 0. Then, please
            give the necessary analysis or reasoning for your choices.",
            PE200_info, ["PE201", "PE202", "PE203", "PE204", "PE205", "PE206",
            "PE207", "PE208", "PE209"]),
        ("P9", "Did the Coordination/Communication Conditions of team members
            contribute to the unsafe act? If yes, please reply 1 and the most
            applicable error code; if no, please only reply 0. Then, please give
            the necessary analysis or reasoning for your choices.", PP100_info,
            ["PP101", "PP109", "PP111"]),
        ("P10", "Did the certification/experience conditions of pilot contribute
            to the unsafe act? If yes, please reply 1 and the most applicable
            error code; if no, please only reply 0. Then, please give the
            necessary analysis or reasoning for your choices.", PT100_info,
            ["PT101", "PT102", "PT103", "PT104", "PT105"]),
    ]

    # Iterate through all questions and invoke corresponding methods
    for p_id, question, info, codes in p_questions:
        p_bool, p_code = ask_precondition_question(text, info, question, codes,
            combined_conditions=combined_conditions)
        print(f"{p_id} Answer is: {p_bool}, {p_id} Code is: {p_code}")
```

# Appendix V.Data merging justice by LRT

Table 6.3: Likelihood ratio test results for within-period and overall model comparisons

| | **Departure** | **Enroute** | **Maneuvering** | **Arrival** |
|---|---|---|---|---|
| **Within 2008-2011** | 72.15(80)[27.79%] | 102.00(91)[79.77%] | 47.87(50)[44.08%] | 95.88(85)[80.28%] |
| **Within 2012-2015** | 122.70(85)[99.53%] | 48.06(61)[11.42%] | 77.27(60)[93.40%] | 72.93(80)[30.05%] |
| **Within 2016-2019** | 54.01(60)[30.69%] | 65.12(57)[78.50%] | 49.39(39)[87.70%] | 111.56(99)[81.70%] |
| **Within 2008-2019** | 130.67(26)[>99.99%] | 94.21(20)[>99.99%] | 98.65(14)[>99.99%] | 185.17(30)[>99.99%] |

Note: Rows 1–3 compare each period-specific model to its yearly models; Row 4 compares the overall model to the three period-specific models (2008-2011, 2012-2015, 2016-2019).

# Appendix VI. Marginal effects

Table 6.4: Marginal effects of explanatory variables

| **Significant variables** | **Departure** | | **Enroute** | | **Maneuvering** | | **Arrival** | |
|---|---|---|---|---|---|---|---|---|
| | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** |
| ***Pilot characteristics*** | | | | | | | | |
| Young pilot [2008-2011] | -0.2089 | | -0.1256 | | | | -0.7932 | |
| Young pilot [2012-2015] | | | | | -0.0581 | | | |
| Elder pilot [2008-2011] | | | | | | | 0.1297 | 0.0350 |
| Elder pilot [2012-2015] | 0.0229 | -0.0881 | | | | | | |
| Elder pilot [2016-2019] | | | | | | | 0.1182 | |
| Female pilot [2016-2019] | | | | | | | -0.3248 | |
| Class 1 [2012-2015] | | 0.1496 | | | | | -0.2150 | |
| Class 2 [2016-2019] | | | | 0.2807 | | 0.1976 | | |
| Class 3 [2016-2019] | | | | 0.1828 | | | | |
| Sport [2008-2011] | | | -0.1012 | | | | | |
| Sport [2012-2015] | | | | | | | -0.2115 | |
| Sport [2016-2019] | | | | | | | | 0.1472 |
| Left seat [2008-2011] | | | | -0.1100 | | | | |
| Left seat [2012-2015] | -0.0266 | | | -0.1939 | | | | |
| Left seat [2016-2019] | -0.0825 | | | | | | | |
| Front seat [2008-2011] | | | | | 0.0473 | | | |

| **Significant variables** | **Departure** | | **Enroute** | | **Maneuvering** | | **Arrival** | |
|---|---|---|---|---|---|---|---|---|
| | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** |
| Front seat [2016-2019] | | | | | 0.0670 | | | -0.1283 |
| Right seat [2008-2011] | | | | | 0.0608 | | | |
| Right seat [2016-2019] | | | | | 0.0921 | -0.1968 | | |
| Skill errors [2008-2011] | 0.1400 | | 0.0559 | | | | 0.1244 | |
| Skill errors [2012-2015] | 0.0250 | | 0.0451 | | | | | |
| Skill errors [2016-2019] | | | 0.1170 | 0.1476 | 0.0833 | | | |
| Decision making errors [2008-2011] | 0.1490 | | 0.1099 | | | | | |
| Decision making errors [2012-2015] | 0.0681 | 0.1399 | 0.0812 | | | | | |
| Decision making errors [2016-2019] | | | | | | | 0.2435 | |
| Perception errors [2008-2011] | 0.6335 | 0.2074 | 0.1492 | 0.1513 | | | 0.5027 | |
| Perception errors [2012-2015] | | | 0.1395 | 0.3866 | | | | |
| Perception errors [2016-2019] | | | 0.1923 | 0.5029 | | | | |
| Maintenance skill errors [2008-2011] | | | | | -0.1253 | | | |
| Maintenance skill errors [2012-2015] | 0.0736 | | | | | | | |
| Attention condition [2008-2011] | -0.2241 | | | | | | | |
| Attention condition [2012-2015] | | | 0.1368 | | | | 0.2752 | 0.1601 |
| Adverse physiologic [2008-2011] | 0.7560 | | 0.1574 | | 0.2015 | 0.0356 | 0.7023 | 0.1351 |
| Adverse physiologic [2012-2015] | 0.2713 | 0.3180 | 0.2736 | | 0.0898 | | | 0.2909 |
| Adverse physiologic [2016-2019] | | | 0.3690 | 0.3610 | 0.1958 | | 0.6174 | 0.3121 |
| Inadequate training [2008-2011] | 0.1851 | | | | | | 0.3487 | 0.0657 |
| Inadequate training [2012-2015] | 0.0745 | 0.2506 | | | 0.0702 | | 0.2006 | 0.1212 |
| Inadequate training [2016-2019] | 0.1851 | | 0.2413 | | | | 0.3497 | |
| Inadequate planning [2008-2011] | 0.2636 | | 0.0787 | | | | 0.2715 | |
| Inadequate planning [2016-2019] | | | | | | | 0.2883 | |
| ***Aircraft characteristics*** | | | | | | | | |
| Airplane [2008-2011] | -0.2437 | | -0.0981 | | | 0.0291 | -0.2047 | -0.0520 |
| Airplane [2012-2015] | -0.0531 | | | | | 0.2263 | -0.4967 | |
| Airplane [2016-2019] | -0.1080 | | | | | | -0.3126 | |
| Helicopter [2008-2011] | -0.2122 | | | | -0.1447 | | | |
| Helicopter [2012-2015] | -0.1071 | | | | | | -0.3764 | 0.1982 |
| Helicopter [2016-2019] | -0.1927 | | | | -0.1239 | | | |
| Glider [2008-2011] | | | -0.1585 | | | | -0.7556 | |
| Glider [2012-2015] | | | | | | | -0.1787 | |
| Homebuilt aircraft [2008-2011] | 0.2211 | 0.0577 | | | | | 0.2025 | |
| Homebuilt aircraft [2012-2015] | | 0.1994 | | | | | 0.1974 | 0.1436 |

| **Significant variables** | **Departure** | | **Enroute** | | **Maneuvering** | | **Arrival** | |
|---|---|---|---|---|---|---|---|---|
| | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** |
| Certificated aircraft [2008-2011] | | | | | 0.1380 | | | |
| Certificated aircraft [2016-2019] | | | | | | 0.2092 | -0.4476 | |
| Single engine [2008-2011] | | | | | | | -0.1494 | |
| Single engine [2012-2015] | | | 0.0519 | | | | | -0.0754 |
| Single engine [2016-2019] | | -0.1895 | | -0.1704 | | | | -0.1409 |
| Zero engine [2008-2011] | | | | | | | 0.7458 | |
| Zero engine [2016-2019] | | | | | | | 0.3743 | |
| Multi engine[2008-2011] | | | 0.0609 | 0.0895 | | | | |
| Fixed landing gear [2008-2011] | -0.0928 | | -0.0676 | | | | | |
| Fixed landing gear [2012-2015] | -0.0328 | -0.1557 | -0.0628 | | -0.0363 | -0.2074 | -0.1359 | -0.1523 |
| Fixed landing gear [2016-2019] | | | -0.0634 | | | -0.3250 | | |
| Aircraft systems issue [2008-2011] | | | | | -0.0656 | | -0.2040 | -0.1255 |
| Aircraft systems issue [2012-2015] | | | | | | | -0.1774 | |
| Aircraft power plant issue [2008-2011] | 0.1750 | | | | -0.0488 | | 0.1834 | |
| Aircraft power plant issue [2012-2015] | 0.0526 | | | | | | | |
| Aircraft power plant issue [2016-2019] | 0.2037 | | | | | | 0.4226 | 0.3029 |
| Aircraft operation issue [2008-2011] | 0.1165 | | 0.1419 | 0.0669 | 0.0694 | 0.0217 | | |
| Aircraft operation issue [2012-2015] | 0.0328 | | 0.0825 | | 0.0422 | | 0.0462 | |
| Aircraft operation issue [2016-2019] | 0.1135 | | | 0.2239 | 0.0918 | | 0.0922 | |
| Fluids/misc hardware issue [2008-2011] | 0.2156 | | | | | | | |
| Fluids/misc hardware issue [2012-2015] | | | | -0.2030 | | | 0.2989 | |
| Fluids/misc hardware issue [2016-2019] | | | | | | | 0.2203 | |
| ***Flight characteristics*** | | | | | | | | |
| Instrument flight rules [2008-2011] | | | | 0.1145 | | | 0.1825 | 0.1112 |
| Instrument flight rules [2012-2015] | | 0.2322 | | | | | | |
| Instrument flight rules [2016-2019] | | 0.4211 | | | | | | |
| Visual flight rules [2016-2019] | -0.4534 | | | | | | | |
| None plan [2016-2019] | | 0.3379 | | | | | | |
| Personal flying [2012-2015] | | | | 0.1246 | 0.0365 | | | |
| Instructional flying [2008-2011] | -0.3273 | | | | | | | |
| Instructional flying [2012-2015] | | | | | | | -0.1669 | |
| Instructional flying [2016-2019] | -0.2183 | | | | | | -0.2501 | -0.2405 |
| Business [2012-2015] | | | 0.1103 | | 0.0619 | | | |
| ***Environmental characteristics*** | | | | | | | | |
| Environmental characteristics | | | | | | | | |

| **Significant variables** | **Departure** | | **Enroute** | | **Maneuvering** | | **Arrival** | |
|---|---|---|---|---|---|---|---|---|
| | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** | **Pilot** | **Aircraft** |
| Daylight [2008-2011] | | | | -0.0826 | | | -0.1495 | |
| Daylight [2012-2015] | -0.0768 | | | | | | | |
| Daylight [2016-2019] | | -0.3312 | | | | -0.3020 | | |
| Night [2008-2011] | | 0.0972 | | | | 0.0443 | | |
| Night [2016-2019] | | -0.3358 | | 0.2130 | | | | |
| Clear [2012-2015] | | | | -0.1784 | | | | |
| Clear [2016-2019] | | | | | -0.0557 | | | |
| Few clouds [2008-2011] | | | | | | | | -0.0732 |
| Few clouds [2012-2015] | | | | | | | -0.0702 | |
| Scattered clouds [2008-2011] | | -0.0842 | | | | | | |
| Scattered clouds [2012-2015] | | | | -0.1377 | | 0.1950 | | |
| Broken clouds [2012-2015] | 0.0246 | | | | | | | |
| Broken clouds [2016-2019] | | | | 0.1233 | | | | |
| Overcast condition [2008-2011] | | | 0.0653 | 0.1162 | | | | |
| Overcast condition [2012-2015] | | | | | | | 0.0904 | |
| Overcast condition [2016-2019] | | | | | | | 0.2037 | |
| None clouds [2008-2011] | 0.0890 | | | | | | 0.1582 | |
| None clouds [2016-2019] | | | | | | | 0.2770 | |
| Visual meteorological condition [2008-2011] | -0.2264 | -0.1423 | | | -0.1217 | -0.0217 | | |
| Visual meteorological condition [2012-2015] | -0.1220 | | -0.1305 | -0.1704 | -0.0533 | -0.3953 | -0.3717 | -0.3745 |
| Visual meteorological condition [2016-2019] | | -0.3132 | -0.0885 | | | | | |
| Ceiling/visibility [2008-2011] | | | 0.1651 | 0.1655 | 0.1402 | | 0.2321 | 0.1119 |
| Ceiling/visibility [2012-2015] | | 0.1911 | | | | | | |
| Ceiling/visibility [2016-2019] | 0.7907 | | 0.3535 | 0.3805 | | | 0.5360 | 0.4740 |
| Wind [2008-2011] | -0.1564 | | | | | | | |
| Wind [2012-2015] | | | | | -0.0386 | | | |
| Wind [2016-2019] | -0.4136 | -0.2757 | | | | | -0.2426 | -0.1639 |
| Terrain [2008-2011] | | | | 0.1945 | | | | |
| Terrain [2016-2019] | -0.3110 | | | | | | | |
| Object/animal/substance [2016-2019] | -0.2907 | | | | | | | -0.2514 |
| Temperature [2008-2011] | 0.1500 | | | | -0.0831 | | 0.3081 | |