Advances in Human Factors of Transportation, Vol. 148, 2024, 77–87
https://doi.org/10.54941/ahfe1005197

Show the Way: Accelerating General Aviation Accident Investigations Through LLMs and HFACS

Liu Qingli[1], Yan Yuqi[1], Li Fan[1], and Feng Shanshan[2]

[1]Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, 999077, Hong Kong Special Administrative Region, Hong Kong
[2]Wecar Technology Co., Ltd., Shenzhen, China

ABSTRACT

General Aviation (GA), with the highest accident and fatality rates in civil aviation, undergoes lengthy accident investigations that include site analysis, witness interviews, cause identification, and detailed reporting. These expert-driven processes, often extending for months or years, not only require extensive manpower but also delay vital accident prevention initiatives in GA. The advent of large language models (LLMs), with groundbreaking capabilities in understanding and generating complex text, offers a potential solution to these challenges. This study aims to conduct a General Aviation Accident Cause Automatic Prediction System (GA-ACAPS), which leverages witness narratives (established early in the investigation) through LLMs. The research utilizes 2250 GA accident reports from the National Transportation Safety Board (NTSB), employing the Human Factors Analysis and Classification System (HFACS) for structured accident causation predictions. Three preliminary experiments were conducted to compare the prediction performance of three different prompting methods before the formal experiment. The results from the preliminary experiments underscore that integrating witness narratives with basic accident information significantly boosts the performance of GA-ACAPS. This optimized prompt was thus implemented in the formal study. The formal experiment's findings demonstrate that GA-ACAPS is proficient in predicting unsafe acts and specific preconditions of unsafe acts like the physical environment and personal readiness. This study endorses the potential of GA-ACAPS to serve as a dependable tool for investigators, aiming to narrow down probable causes of accidents and thereby increase the efficiency of investigations. Moreover, the application of LLMs in GA accident analysis heralds a new era of innovative approaches and essential insights, contributing to the advancement of aviation safety.

Keywords: Large language models, HFACS, GA accidents, Automated accident investigation system, Prompt

INTRODUCTION

General aviation (GA, 14CFR Part 91), refers to civil aviation operations excluding those conducted under Part 121 air carrier operations or Part 135 commuter and on-demand operations, faces high accident rates. Data from the National Transportation Safety Board (NTSB) shows a slight decline

in GA accidents in the United States from 2008 to 2022, while with a notable increase post-COVID-19 lockdowns (NTSB, 2023), as represented in Figure 1. In 2022, GA accidents alarmingly made up 94% of all civil aviation accidents.

Accident investigation is key for improving aviation safety, as it not only identifies the causes of accidents but also addresses safety vulnerabilities, thereby preventing similar future accidents (Zhong et al., 2020). Consequently, the accuracy and promptness in pinpointing these causes are of utmost importance. In the context of GA safety, however, a delicate balance often exists between the precision of identified causes and the speed of their disclosure. Using NTSB's investigation as an illustration, a team quickly forms post-accident for an on-site investigation, scrutinizing the scene, gathering evidence, and interviewing witnesses promptly. This is followed by collating additional data like pilot records, aircraft maintenance logs, and weather reports. After an initial report draft, it undergoes multiple reviews and revisions before the final report's public release (NTSB, 2024b). Thus, pinpointing a GA accident's cause can take months to years of expert analysis. Such protracted period not only casts doubt on necessary aircraft or operational modifications but also substantially delays preventive actions against similar future accidents.

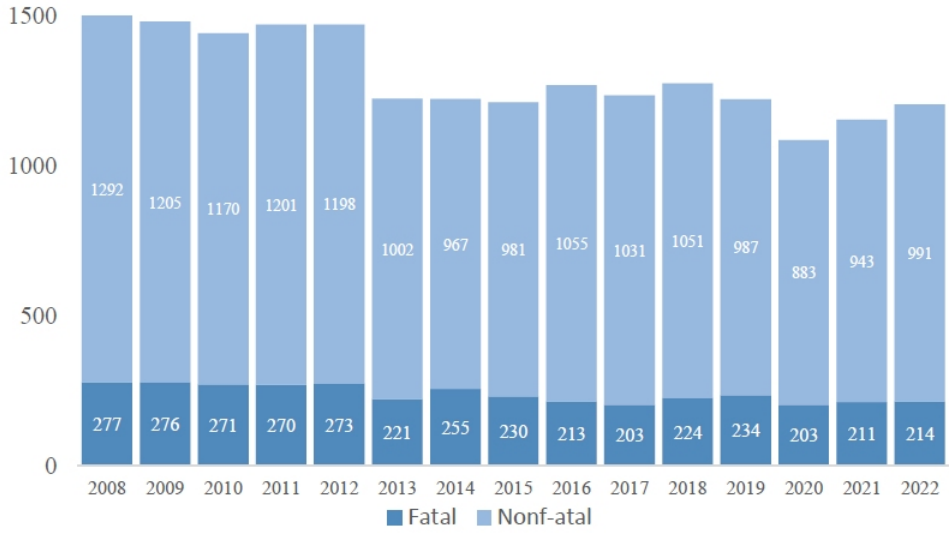

Figure 1: GA accidents by calendar year (adapted from NTSB, 2023).

The recent advent of large language models (LLMs) such as ChatGPT developed by OpenAI, renowned for their exceptional text comprehension and generation abilities without being trained in specific tasks (Thirunavukarasu et al., 2023; Touvron et al., 2023), offers a promising intelligent approach to expedite GA accident analysis. In this research, we attempt to leverage LLMs to create a General Aviation Accident Cause Automatic

Prediction System (GA-ACAPS) based on witnesses' narratives. Witness narratives, typically from pilots and passengers of the involved aircraft or pilots from other aircraft, are often obtained early in accident investigations and contain a wealth of accurate and fresh information about the events leading up to the accident. Determining accident causes from these narratives requires contextual comprehension and inferential reasoning that go beyond mere basic natural language processing (NLP), as witness statements mix subjective perceptions with objective facts in unstructured formats. In such cases, LLMs with advanced text analysis capabilities, are adept at extracting key information from these unstructured contents to predict potential accident causes.

Additionally, to structure and trace predicted accident causes, the Human Factors Analysis and Classification System (HFACS) is utilized in this research. Renowned for its advantageous taxonomy (Zheng et al., 2024), HFACS is extensively used in aviation accident investigations (Dönmez & Uslu 2020; Li et al., 2008). It classifies human factors into four levels: organizational influences, unsafe supervisions, preconditions for unsafe acts, and unsafe acts (Shappell & Wiegmann, 2000). By integrating the HFACS framework with the capabilities of LLMs (cantered around ChatGPT), the proposed GA-ACAPS not only can identify potential causes but also classify them within systematic layers.

However, developing GA-ACAPS using LLMs presents a primary challenge: optimizing prompt design to enhance its predictive accuracy. To tackle this, the study first carried out three preliminary experiments evaluating three different prompting strategies for predicting four layers of causes. Subsequently, the most effective prompt was chosen for the formal experiment.

It's important to clarify that this paper's contribution lies not in identifying final GA accident causes, but rather in the automated generation of probable causes mainly using witness narratives via GA-ACAPS. By leveraging LLMs, GA-ACAPS aims to offer investigators potential leads in the early stages of investigation, thereby reducing overall investigation time and accelerating the determination of final causes.

The present study begins with experimental framework including data preparation, preliminary experiment, and formal experiment design. Then, it is followed by the prediction results and discussion. Finally, the paper concludes with a summary of major findings and point out some lines in future work.

METHOD

Fig 2 represented the research workflow. It is noted that OpenAI provides access to GPT-4 via an interactive chatbot interface or an API, both offering identical functionalities, but the chatbot provides a simpler interface to run experiments. Therefore, we choose interactive chatbot with GPT-4 in this research. More details can be seen as follows.

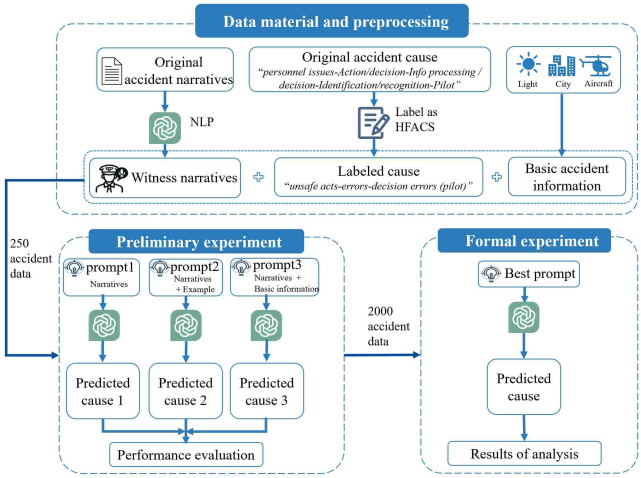

Figure 2: Research workflow.

Data Material and Preprocessing

The dataset, consisting of GA accident data from the United States, was randomly sourced from the NTSB database (NTSB, 2024a). For this study, it was refined by excluding information such as flight crew details, accident severity, and sequences of events, which are time-consuming to verify in investigations. As a result, the dataset was narrowed down to four elements: accident ID, narrative descriptions, accident causes, and a selection of basic accident information including light conditions, aircraft category, and occurrence city.

Prior to initiating the preliminary experiments, the dataset underwent the following processing steps:

1. The original narratives, containing pilot or witness statements and investigators' descriptions, were refined using ChatGPT to extract only the pilot or witness accounts. Records without witness statements were then identified and excluded. After that, 2,250 accident records with witness statements were randomly selected for the preliminary and formal experiments.
2. The causes of the selected accidents were manually labelled by experts following the HFACS framework. For instance, an original cause like 'personnel issues-Action/decision-Info processing / decision-Identification/recognition-Pilot' was labelled as 'unsafe acts-errors-decision errors (pilot)'. It's important to note that during the labelling process, certain identified causes did not conform to traditional HFACS subcategories, necessitating the expansion of the HFACS taxonomy to include 'task environment' and 'operational environment' as subcategories under 'preconditions for unsafe acts-environment factors' (see Figure 3). Furthermore, it's noteworthy that most accident causes recorded by the NTSB typically omit 'organizational influences' and 'unsafe supervision' categories.
3. The manually labelled data was divided into two sets: 250 cases for preliminary experiments, with the witness narratives from 200 cases used in three pre-experiments, and 50 narratives along with their corresponding labelled causes utilized as extra knowledge in one of the prompt strategies. The rest of the cases formed the test set for the formal experiments.

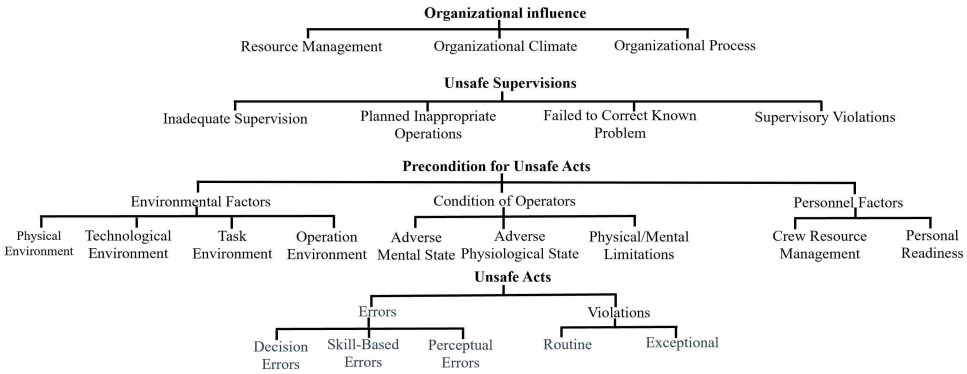

Figure 3: Improved HFACS (adapted from (Shappell & Wiegmann, 2000)).

Preliminary Experiment

Three preliminary experiments were designed to determine the optimal prompting strategy before the formal study. Prompt 1 used solely witness narratives; Prompt 2 added 50 corresponding accident causes for additional context; and Prompt 3 engaged ChatGPT with narratives and basic accident details for prediction. The results and outputs for these prompts were illustrated in Figures 4-6. The strategies were evaluated on their accuracy in classifying the four HFACS levels of causes, with the best-performing prompt chosen for the formal experiment.

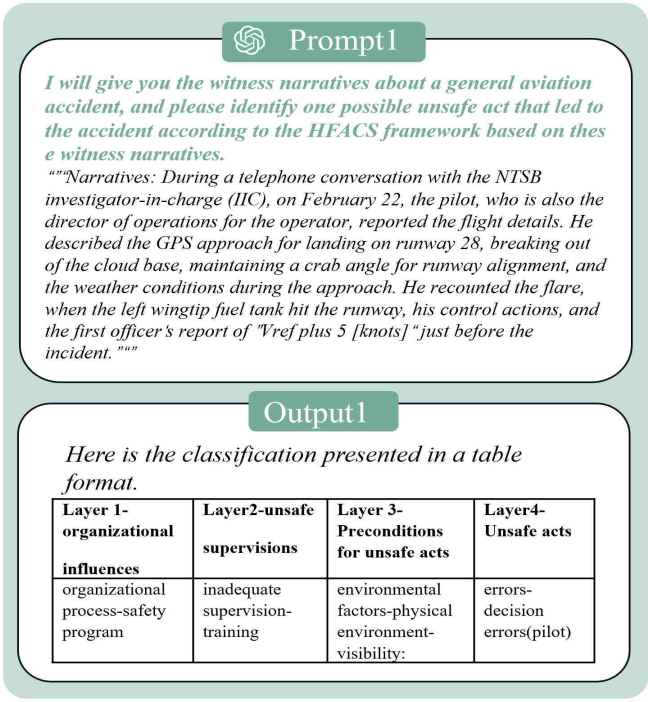

Figure 4: Prompt 1.

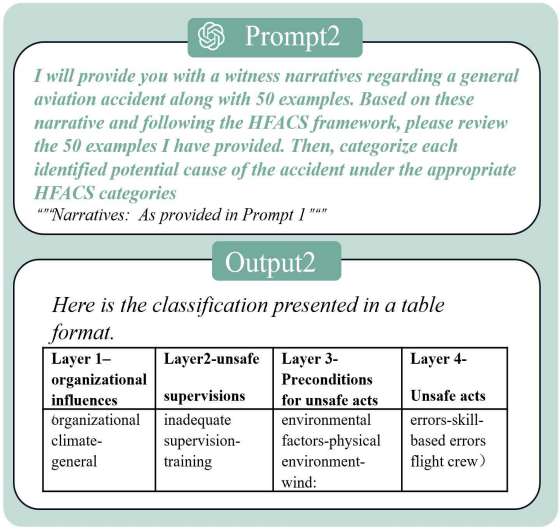

Figure 5: Prompt 2.

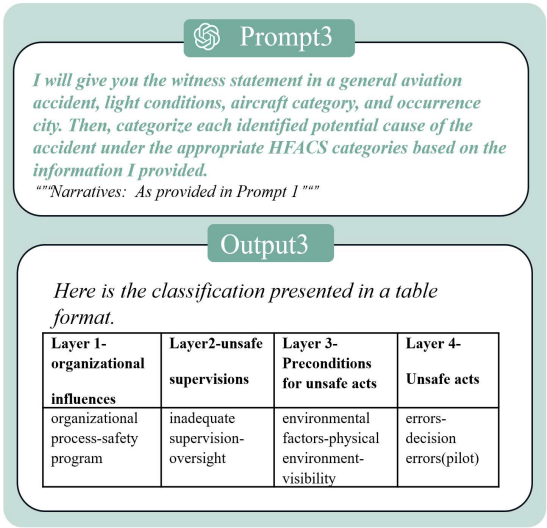

Figure 6: Prompt 3.

Table 1 illustrated the prediction performance of the three prompts across the four levels of accident causation. As shown in Table 1, due to only one instance each of “organizational influence” and “unsafe supervision” identified within the 200 tested reports, all three prompts exhibited extreme performance in these categories (100% or 0). However, for preconditions of “unsafe acts”, Prompt 3 was the most effective, while Prompt 1 yielded the highest accuracy for predicting “unsafe acts”, followed by Prompt 3. Therefore, after comprehensive consideration, Prompt 3 was selected for the formal experiment. Concurrently, the technical specifics of GA-ACAPS were finalized to automatically generate probable accident causes using witness narratives and basic accident information.

Table 1. Performance in terms of accuracy (%) achieved by three prompts.

	Prompt1	Prompt2	Prompt3
Organizational influences	0.00%	0.00%	0.00%
Unsafe supervisory	100.00%	100.00%	100.00%
Preconditions of unsafe acts	58.50%	61.50%	67.00%
Unsafe acts	81.00%	73.00%	78.00%

Formal Experiment

The formal experiment employed Prompt 3 to predict the causes of 2,000 GA accidents. The performance of the proposed GA-ACAPS was evaluated based on values of precision, recall, F1 score and confusion matrixes across four layers of cause categories. Results and discussions of the formal experiment were presented in the following section.

RESULTS AND DISCUSSIONS

The Results of Precision, Recall, and F1 Score

The performance of GA-ACPS in predicting different accident categories was presented in Table 2, the key indicators were precision and recall rates and F1 score metrics.

1. Precision: a measure of how many of the positive predictions made are correct (true positives).
2. Recall: a measure of how many of the positive cases the classifier correctly predicted, over all the positive cases in the data.
3. F1 Score: a measure combining both precision and recall. It is generally described as the harmonic mean of the two.

The three indicators can be calculated sequentially according to Eqs.(1)–(3) (Price & Bouvier, 2002):

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

where, TP is the number of true positive, FN is the number of false negative, FP is the number of false positive. The results in Table 2 indicated that 'inadequate supervision' achieved outstanding performance with perfect scores in all metrics. 'Skill-based errors' also exhibited robust results, particularly in recall (0.8685), highlighting the model's predictive accuracy in this category. However, the system faced challenges in accurately predicting categories such as 'perceptual errors', 'routine' as evidenced by their considerably lower scores. It is intuitive because pilots (the most common witness) do not proactively report their own violations or perceptual errors due to self-preservation or fear of reprimand. A striking observation was the high

recall (0.9232) yet lower precision in the 'technological environment' category, suggesting a tendency of GA-ACPS to overestimate predictions in this specific area. Due to the impact of imbalanced samples (where the proportion of actual accident cause categories varies significantly) on the results of precision, recall, and F1 score (Korkmaz, 2020), the confusion matrix below has been presented therefore to show the detailed and comprehensive prediction performance across the four HFACS categories.

Table 2. Performance in terms of precision, recall, and F1 score in formal experiment.

	Precision	Recall	F1 Score
Organizational Process	0.5556	1.0000	0.7143
Resource Management	0.5000	0.2000	0.2857
Organizational Climate	1.0000	0.2857	0.4444
Inadequate Supervision	1.0000	1.0000	1.0000
Physical environment	0.7299	0.5487	0.6265
Technological Environment	0.4739	0.9232	0.6263
Personal Readiness	0.6393	0.2058	0.3114
Other preconditions of unsafe acts	0.3929	0.0696	0.1183
Skill-Based Errors	0.7558	0.8685	0.8082
Decision Errors	0.4861	0.3070	0.3763
Perceptual Errors	0.2500	0.1250	0.1667
Routine	0.0800	0.1429	0.1026
Exceptional	0.5000	0.0263	0.0500

The Results of Confusion Matrices

In Figures 7–9, the confusion matrices articulated the GA-ACPS's prediction performance on HFACS categories. Diagonal entries indicate correct predictions; all others denote misclassifications. Figure 7 underscored the GA-ACPS's high accuracy in "inadequate supervision", with a perfect true positive rate, suggesting exceptional identification and classification capabilities for this category. In contrast, "organizational process" displayed a mix of true and false positives. Moreover, the infrequent predictions for "resource management" and "organizational climate" hint at their possible underrepresentation in the data. The less satisfactory results in these organizational factors could be attributed to the National Transportation Safety Board (NTSB) reports' limited emphasis on such elements. Figure 8 shows the strong ability of GA-ACPS to identify "technological environment" within the preconditions for unsafe acts. However, it often confused "physical environment" with "technological environment". As indicated in Figure 8, the "technological environment" was confused 300 times with "physical environment". Furthermore, "personal readiness" was moderately recognized but not distinctly. In the manual labelling of accident causes within the HFACS framework, an overlap has been noted. Certain root causes could be classified both as 'Adverse mental/ physiological state' or as 'personal readiness'. For instance, an incident attributed to alcohol consumption by a pilot - 'personnel issues-physical-impairment/incapacitation-alcohol-pilot' - might indicate

an 'adverse physiological state' due to the impairment from alcohol, or it could suggest a lack of 'personal readiness'. This ambiguity suggests the need for a more granular and expanded approach within the HFACS framework, allowing for clearer differentiation and potentially leading to more accurate categorization of accident causes. Additionally, in Figure 9, "skill-based errors" were mostly correctly identified, yet the model confused them with "decision errors". The minimal true positives for "perceptual errors" and "routine" suggest these categories might be blind spots for the system.

Figure 7: The confusion matrix in organizational influence and unsafe supervisions.

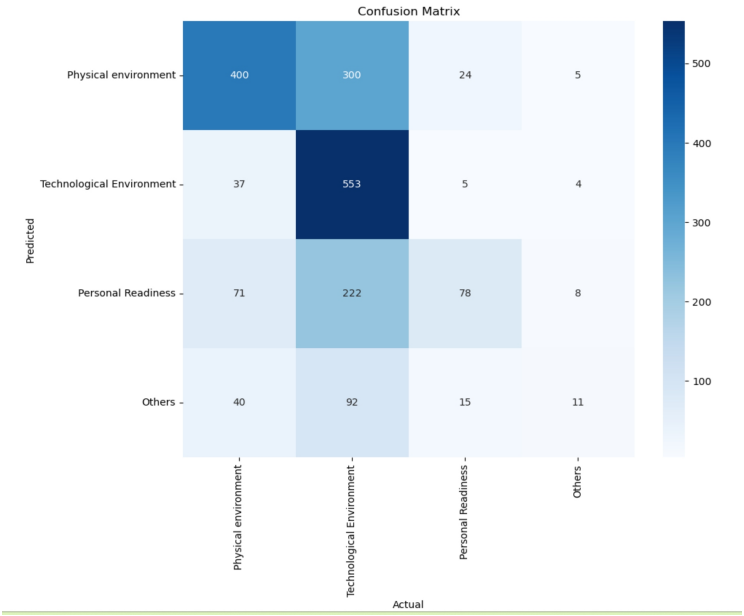

Figure 8: The confusion matrix in preconditions of unsafe acts.

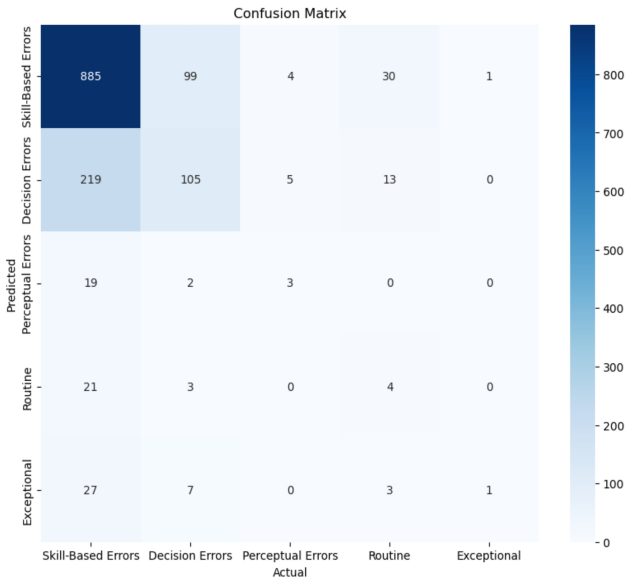

Figure 9: The confusion matrix in unsafe acts.

The above results indicate that while GA-ACPS does need refinement to improve its accuracy in less common yet significant categories, it still shows promise, particularly in identifying certain error types only using witness narratives and basic accident information. Therefore, it can serve as a valuable tool for accident investigators, providing them with initial insights that could potentially streamline the investigative process and enhance efficiency by guiding the direction of the investigation and helping to prioritize areas of focus.

CONCLUSION

This study investigated the performance of LLMs in predicting causes of GA accidents. Specifically, it harnessed LLMs' natural language processing to extract information from witness statements and applied reasoning for cause prediction. Following a series of preliminary tests to ascertain the most effective prompts, the prompt using both witness narrative and basic accident information emerged as the most successful and was therefore chosen for the formal investigation. Analysing 2,000 GA accident cases, the study found that the proposed GA-ACPS showed limitations in predicting "organizational factors" and "technological environment". However, it demonstrated a robust capability in identifying certain "preconditions for unsafe acts" and "unsafe acts". The findings suggest that the potential of GA-ACPS to provide valuable initial insights for accident investigators by predicting GA accident causes, thereby offering guidance, and improving the efficiency of investigations. Moreover, the study highlights the potential of Language Model systems (LLMs) to discern accident causes from witness narratives, marking a significant step forward in aviation safety research. The study's key limitation lies in the simple design of the prompts and the accuracy of the manual

labelling system used to evaluate GA-ACPS's performance. Future efforts will concentrate on refining both the prompts of LLMs and the manual labelling process.

REFERENCES

Dönmez, K., & Uslu, S. (2020). The effect of management practices on aircraft incidents. *Journal of Air Transport Management*, *84*.

Korkmaz, S. (2020). Deep Learning-Based Imbalanced Data Classification for Drug Discovery. *J Chem Inf Model*, *60*(9), 4180–4190.

Li, W. C., Harris, D., & Yu, C. S. (2008). Routes to failure: Analysis of 41 civil aviation accidents from the Republic of China using the human factors analysis and classification system. *Accident Analysis and Prevention*, *40*(2), 426–434.

NTSB. (2023). *US Civil Aviation Accident Dashboard: 2008-2022*. https://www.ntsb.gov/safety/StatisticalReviews/Pages/CivilAviationDashboard.aspx

NTSB. (2024a). *Aviation Investigation Search*. https://www.ntsb.gov/Pages/AviationQueryV2.aspx

NTSB. (2024b). *The Investigative Process*. https://www.ntsb.gov/investigations/process/Pages/default.aspx

Price, T. D., & Bouvier, M. M. (2002). The evolution of F1 postzygotic incompatibilities in birds. *Evolution*, *56*(10), 2083–2089.

Shappell, S. A., & Wiegmann, D. A. (2000). The human factors analysis and classification system--HFACS.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930–1940.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., & Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models *arXiv preprint ArXiv. /abs/2307.09288*.

Zheng, Q., Liu, X., Wang, W., & Han, S. (2024). A hybrid HFACS model using DEMATEL-ORESTE method with linguistic Z-number for risk analysis of human error factors in the healthcare system. *Expert Systems with Applications*, *235*.

Zhong, B., Pan, X., Love, P. E. D., Ding, L., & Fang, W. (2020). Deep learning and network analysis: Classifying and visualizing accident narratives in construction. *Automation in Construction*, *113*.