# The Process of Training ChatGPT Using HFACS to Analyse Aviation Accident Reports

**Declan Saunders, Kyle Hu & Wen-Chin Li**

Cranfield University, United Kingdom

## SUMMARY

This study investigates the feasibility of a generative-pre-trained transformer (GPT) to analyse aviation accident reports related to decision error, based on the Human Factors Analysis and Classification System (HFACS) framework. The application of artificial intelligence (AI) combined with machine learning (ML) is expected to expand significantly in aviation. It will have an impact on safety management and accident classification and prevention based on the development of the large language model (LLM) and prompt engineering. The results have demonstrated that there are challenges to using AI to classify accidents related to pilots' cognitive processes, which might have an impact on pilots' decision-making, violation, and operational behaviours. Currently, AI tends to misclassify causal factors implicated by human behaviours and cognitive processes of decision-making. This research reveals the potential of AI's utility in initial quick analysis with unexpected and unpredictable hallucinations, which may require a domain expert's validation.

## KEYWORDS

Artificial Intelligence, Aviation Safety, ChatGPT, Human Factors Analysis Classification System

## Introduction

The Human Factors Analysis and Classification System (HFACS) framework provides a systematic approach to the contributing factors of active failures and latent conditions leading to an accident or incident (Wiegmann & Shappell, 2003). Contributing factors are classified into a structured framework including 18 categories across four consequential levels (L1 to L4) of HFACS: starting from classifying causal factors into active failure categories at the lowest level, analysts work upwards through the framework to classify the associated precursor latent conditions using the presented taxonomies (Harris & Li, 2019). The aviation industry is increasingly interested in adopting artificial intelligence (AI) to improve efficiency, safety, and competitiveness (Kabashkin et al., 2023). AI application could be an efficient tool for organization's continuous learning, growth, and improvement by creating new knowledge from previous.

The development of a large language model (LLM) is encoding a given text by semantically linking its words and then decoding a response using internal statistical models (Bender et al., 2021). Prompt engineering is part of LLM: it influences the outputs of AI applications, as the prompt is the input of the model, and its engineering can result in significant output differences (Kaddour et al., 2023). The emergence of AI tools like ChatGPT, with advanced natural language processing capabilities, offers promising approaches for enhancing preliminary causal root analysis. The use of prompt engineering and a well-established classification framework could reach at least a fair agreement between human subject-matter experts and AI in accident analysis (Ziakkas & Pechlivanis, 2023). Based on the data-driven nature of AI applications, the outputs of accident analysis and classification may be able to provide the meticulous data management process and precise data labelling according to the required analysis framework to ensure sufficient data is collected for training, validation and testing purposes. The combination of AI and HFACS in analysing accident reports represents a significant advantage in cost efficiency in developing effective aviation safety management systems failures. HFACS provides a structured framework for identifying and categorizing human factors in aviation accidents. This study explores the potential and challenges of using AI to analyse decision errors in aviation accident reports based on the HFACS framework. It is to simplify the training process of using GPT to conduct detailed and precise classification at the initial stage of research.

## Methods

Participants: Six subject-matter experts in the aviation safety domain participated in this research. All participants were familiar with the Human Factors Analysis and Classification System (HFACS) and using the HFACS framework to analyse aviation accident reports. The accident reports represent a broad spectrum of human factor-related incidents/accidents in aviation, from flight deck human errors to organizational culture and safety management.

Prompt Engineering: A prompt is natural language text describing the task that an AI should perform. It is the process of structuring text that can be interpreted and understood by a generative AI model. The parameters were set as follows: temperature (0.01), top P (0.05) frequency penalty (0) and Presence penalty (0) (Giray, 2023).

Instruction: You are an expert on aviation safety and accident investigation. You will be provided with a detailed accident report. Please use the Human Factors Analysis and Classification System (HFACS) framework to classify the " decision errors" based on the attached accident report.

Definition: The "decision errors" based on HFACS has a diverse range of definitions, including "improper procedures/strategies", "misdiagnosed emergency", "wrong response to an emergency", "exceeded ability", "inappropriate manoeuvre", "poor decision", "improper in-flight planning", "improper remedial actions in an emergency", "inadequate knowledge of system procedure".

Example: The examples of "decision errors" from previous accident reports include "not following ATCO's instruction", "delayed RTO decision", "decided to land in bad weather", "too late to decide go-around", "select inappropriate procedures", "not confirmed cross-check", "late initiation of the flare", "improper prioritizes tasks", "not correct undesired flight path".

Output Indicator: The output should include a specific textual description related to " decision errors" that appeared on the accident report and indicate the page numbers on the accident report. Screen the textual keywords related to the definitions of " decision errors" or examples of " decision errors" based on the accident reports. Summarise each " decision error" involved in the accident as one bullet point and provide the exact page number on the accident report.

Algorithm: The authors decided to utilize GPT 4.0-turbo, a stable and robust platform developed from its predecessor, GPT 3.5 by OpenAI. Pre-trained with over 1.76 trillion parameters, across eight models consisting of 220 billion parameters, GPT 4.0-turbo can support greater textual generation with larger token allowances per message (Bastian, 2023). This platform is promoted with greater clarity and consistency in textual generation than previous iterations, supporting the evidence-based assessments that will be conducted to support accident casual investigation. OpenAI reports that GPT 4.0-turbo exhibits human-level performance on various professional and academic benchmarks, achievable to its optimized behaviour model to predict the next token in the sequence more consistent with the previous token (Achiam et al., 2023). This makes GPT 4.0-turbo the optimum platform for developing an academic-based algorithm to conduct complex hierarchical task analysis, reducing the likely 'hallucinations' in the output to improve the Agreement in the statement.

Process of Training: The selected accident report of GE235 was summarised and converted into structured prompts which were fed into the AI model as part of the fine-tuning process (Chen et al., 2023). Following the training, six subject matter experts (SMEs) in aviation safety evaluated the AI's performance. Each SME evaluated the AI's analysis of GE235 and was asked to assess AI's performance based on four criteria: Accuracy, Comprehensiveness, Satisfaction, and Agreement. Alongside these quantitative assessments, qualitative feedback was solicited from the experts to gain deeper insights into their evaluations of the rationale behind their scores. Fine-tuning practices were focused on the author's vast experience with deploying HFACS analysis in a corpus of scenarios, with specific level definitions supported by Wiegmann and Shappell's conclusions in 2003. This approach allowed for a comprehensive assessment of the AI's capabilities in the context of accident categorizations (Wiegmann & Shappell, 2003). The flowchart of the research design is shown below (Figure 1).

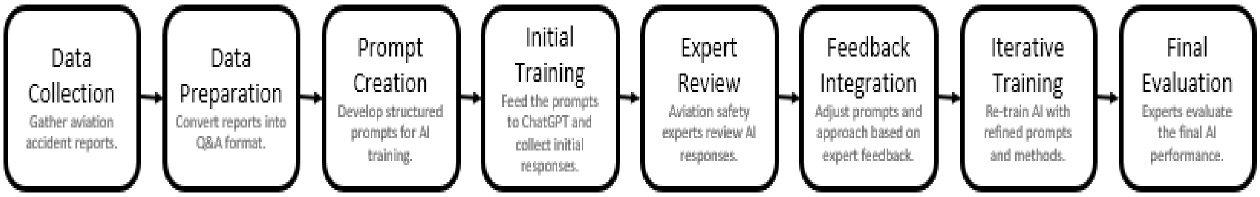


Figure 1. The Flowchart of Research Design on Using HFACS to Analyse Decision Errors Involved in the GE235 Accident Reports

## Result and Discussion

This study was concerned with the application of HFACS for "decision errors" by the pilots during the accident case of TransAsia Airways Flight 235 (GE235), with supporting feedback from SME's involved in the accident investigation. Throughout the procedure demonstrated in Figure 1, prompts were developed to capture instances of decision error as defined by Wiegmann and Shappell in 2003. Table 1 shows an example of prompts used to fine-tune Open AI. Initial textual outputs revealed a significant capacity for ChatGPT to perform complex data interpretation tasks for categorization and classification purposes, with notable diminution in applying HFACS to accident reports. The SME rated ChatGPT's performance on four independent scales: Accuracy, Comprehensiveness, Satisfaction, and Agreement, designed to identify limitations within the model to simulate human-like analysis processes required for effective HFACS assessment. The initial review by SME's revealed concerns about the limitation of dissemination between concepts of error, a consideration with human operators performing the same analysis. Feedback was focused on improving the GPT's ability to make these distinctions in human error, to consolidate the output into solely decision error related topics. Following this feedback, the prompts were revised with the considerations of the SME's to aid in the detection of decision error events, as well as independent research to aid in this distinction, based on the LLM's framework to predict the next token on the highest probability. Final evaluation ratings for decision error on GE235 were: Accuracy (76.19%), Comprehensiveness (71.43%), Satisfaction (63.10%), and Agreement (69.05%); indicating further fine-tuning is required to improve ChatGPT's functionality to analyse the causal factors related to aviation accidents.

The weightings represent a significant improvement towards the development of prompt engineering practises for decision error, although it's evident there is still lots of work required to meet the expectations of SME's in its conclusion drawing. Limitations associated with the innate principles of LLM's are evidenced by the discrepancy between the four scales and relatively low overall scores from expectations, likely due to the AI's inability to mimic the 'intellectual' properties required for correct dissemination between decision error and skill-based error. This often led to cases of misclassification of like-minded phrases between the two error forms, such as those associated with checklist procedures and disconnection of the autopilot. It has been concluded that these innate limitations to AI dissemination ability can be closed by user-allocated fine-tuning, to supplement these conversational areas and reduce the 'hallucinations' generated (Harris & Li, 2019).

Table 1: Example of prompt based on HFACS analysis for accident GE235 to train GPT.

| Content | Prompt |
|---|---|
| **Explain what decision error means** | {"role": "user", "content": "What is decision error according to the HFACS framework?"}, {"role": "assistant", "content": "Decision error represents intentional behaviour that proceeds as planned, yet the plan itself proves inadequate or inappropriate for the situation"} |
| **Introduce the crew decision errors for GE235** | {"role": "user", "content": "What decision errors did the crew commit, focusing on the approach procedures at Magong Airport?"}, {"role": "assistant", "content": "The crew shut down engine No.2, which was fully functional at the time, rather than the malfunctioning engine No.1"} |
| **Determine the precondition for decision error for GE235** | {"role": "user", "content": "Using the HFACS framework, what influences did TransAsia Airways have on the decision error of the crew?"}, {"role": "assistant", "content": "TransAsia Airways did not follow its procedures when selecting and training pilots flying for an upgrade. The airline also did not use widely avaliable crew resource management (CRM) guidelines to develop, implement, reinforce, and assess the effectiveness of their flight crew CRM training program"} |

This study underscores ChatGPT's potential in streamlining the preliminary stages of aviation accident analysis within the HFACS framework. However, it also presents significant limitations in the AI's ability to conduct detailed and nuanced analysis, particularly evident in its overemphasis on the pilot's operational behaviour over their actions, which are related to the complicated cognitive processes involved in decision error classification (Shanahan, 2023). This highlights the necessity of human oversight in interpreting AI findings, advocating for a hybrid approach that synergises AI efficiency with human expertise to address the human-centric approach to AI in aviation, set forward by the European Union Aviation Safety Agency (EASA; European Union Aviation Safety Agency, 2023).

**Example of Output on Decision Error based on Level-1 HFACS:**

After thoroughly reviewing the accident report, I have identified several instances of "Decision Error" following the HFACS framework. Here is a detailed classification:

1. Heading: "Engine No.1 Flameout Warning"
    - Page Number: 24
    - Extract from Report: "The crew shut down engine No. 2, which was fully functional at the time, rather than the malfunctioning engine No. 1."
    - Decision Error: Improper remedial actions in an emergency.
    - Category: Procedural Error
    - Subcategory: Wrong choice of action during an operation.
    - Level of Impact and Pre-Requisites:
        - Level 1: Decision Error
        - Level 2: Adverse mental state, Crew resource management
        - Level 3: Inadequate supervision

2. Heading: "Crew Handling of Aircraft after Engine No.1 Flameout Warning"
    - Page Number: 27
    - Extract from Report: "The crew's failure to perform the necessary flight control and power adjustments after the loss of engine No. 1."
    - Decision Error: Misdiagnosed emergency and improper in-flight planning.
    - Category: Knowledge-Based Mistake
    - Subcategory: Inadequate real-time risk assessment, Delayed a necessary action.
    - Level of Impact and Pre-Requisites:
        - Level 1: Decision Error
        - Level 2: Adverse mental state, Crew resource management
        - Level 3: Inadequate supervision
3. Heading: "Non-Standard Operation Procedure Compliance"
    - Page Number: 31
    - Extract from Report: "The crew did not follow the standard operating procedures for handling an engine flameout, leading to the shutting down of the wrong engine."
    - Decision Error: Selected inappropriate procedures.
    - Category: Procedural Error
    - Subcategory: Rushed a necessary action, Wrong choice of action during an operation.
    - Level of Impact and Pre-Requisites:
        - Level 1: Decision Error, Routine Violation
        - Level 2: Crew resource management
        - Level 3: Inadequate supervision
4. Heading: "Crew Resource Management"
    - Page Number: 34
    - Extract from Report: "Lack of effective communication and resource management among the crew during the emergency situation."
    - Decision Error: Ignored instruction or warning.
    - Category: Knowledge-Based Mistake
    - Subcategory: Failure to prioritize tasks adequately.
    - Level of Impact and Pre-Requisites:
        - Level 1: Decision Error
        - Level 2: Adverse mental state, Crew resource management
        - Level 3: Inadequate supervision

Notably, each of these instances reflects a significant deviation from expected procedural and knowledge-based responses to the emergent situations, influenced by various factors at different levels of HFACS categorization.

There is significant interest in the capabilities of AI models to perform a range of applications, given their analytical abilities with large volumes of data and categorization (Ray, 2023). Despite this, GPT and large language models are not without limitations. The quality of the output is heavily dependent on the quality and diversity of the training data (van Dis et al., 2023). Ethical considerations also come into play, especially concerning data privacy and the potential for the models to generate misleading or biased information (Zhuo et al., 2023). The field of large language models is still in a phase of rapid development. Ongoing research aims to address existing limitations and ethical concerns, to make these models more reliable and applicable across a broader range of tasks in aviation safety management and accident investigation (Ray, 2023).

## Conclusion and future research

While there is clear potential for GPT models to supplement accident analysis within the HFACS framework, there is still more work to synergise the two systems together for effective communication. Further, significant improvements in AI technology are required to close the gap between superficial displays of ‘intelligence’ and intelligence. This study focused on the highlights of GPT integration for causal accident investigation, focusing on the minimal analytical time for HFACS application and its superior categorization capacity, unrivalled by human operators. The nature of GPT makes it less susceptible to bias or discrimination when generating textual responses, which makes it suitable as an effective analysis method for human operation, as opposed to a standalone solution. This is evident in the tendency of GPT to output results which can be exaggerations of the prompt requirements, to meet the expected outcomes of the user. This was notable with SME’s comments on their concerns about the limitations of generative output to disseminate between decision error and skill-based error within this study's scope. As such, the human operator will be required to verify the quality and accuracy of data output to ensure the facts of the accident are captured.

With continued support from SME’s, the next steps will prioritize the training of GPT in all four levels of the HFACS framework, providing a complete analysis for review and further improvement. Development of new prompts will be centered around the work done in this study and the feedback from SME’s, who identified the limitation of GPT to disseminate between similar terms, for example decision error and skill-based error, which were generally classified incorrectly. To address these limitations, fine-tuning on a magnitude of accident reports will focus on cases of clear and distinguishable level classification to train the model to identify and report on the required information from accident data, before focusing on more challenging accident cases with less evident cases of human error. While the strengths of AI in quick context analysis position it as an effective complementary tool for causal analysis, it is restricted to being a human-operated system as opposed to a standalone solution. This is evidenced by the fundamentals of AI capability and capacity as of 2024, with limitations to its superficial display of ‘intelligence’ and resourcefulness to make these distinctions where a human operator could. The unrivaled computation time to conduct HFACS using GPT, along with its superior classification and categorization capacity support the development of these tools for future accident investigation practices, where small amounts of information can be used for analysis to support the investigation process. It is expected to see preliminary usage of these tools within the next couple of years, supported by our SME’s interest and continual support in developing and delivering an analytical tool into the hands of accident investigators around the world.

## References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., … Zoph, B. (2023). GPT-4 Technical Report. Computation and Language, vol 4. https://doi.org/10.48550/arXiv.2303.08774

Bastian, M (2023). GPT-4 has more than a trillion parameters – Report. The Decoder, AI in practice. https://the-decoder.com/gpt-4-has-a-trillion-parameters/

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. Computation and Language, 2. https://doi.org/10.48550/arXiv.2310.14735

European Union Aviation Safety Agency, (2023). Artificial Intelligence Roadmap 2.0: Human-centric approach to AI in Aviation. May 2023 Version 2.0, https://www.easa.europa.eu/en/document-library/general-publications/easa-artificial-intelligence-roadmap-20

Giray, L. (2023). Prompt Engineering with ChatGPT: A Guide for Academic Writers. Annals of Biomedical Engineering, 51(12), 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

Harris, D., & Li, W.-C. (2019). Using Neural Networks to predict HFACS unsafe acts from the pre-conditions of unsafe acts. Ergonomics, 62(2), 181–191. https://doi.org/10.1080/00140139.2017.1407441

Kabashkin, I., Misnevs, B., & Zervina, O. (2023). Artificial Intelligence in Aviation: New Professionals for New Technologies. Applied Sciences, 13(21), 11660. https://doi.org/10.3390/app132111660

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems, 3, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Shanahan, M (2023). Talking About Large Language Models. Computation and Language, 5. https://doi.org/10.48550/arXiv.2212.03551

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: five priorities for research. Nature, 614(7947), 224–226. https://doi.org/10.1038/d41586-023-00288-7

Wiegmann, D. A., & Shappell, S. A. (2003). A Human Error Approach to Aviation Accident Analysis. Routledge. https://doi.org/10.4324/9781315263878

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity. Computation and Language, 4. https://doi.org/10.48550/arxiv.2301.12867

Ziakkas, D., & Pechlivanis, K. (2023). Artificial intelligence applications in aviation accident classification: A preliminary exploratory study. Decision Analytics Journal, 9, 100358. https://doi.org/10.1016/j.dajour.2023.100358