Developing Prompts to Facilitate Generative Pre-Trained Transformer Classifying Decision-Errors in Flight Operations

Wen-Chin Li
Cranfield University
Wenchin.li@cranfield.ac.uk
+44 (0) 1793 785810

Declan Saunders
Cranfield University
Declan.saunders.173@cranfield.ac.uk
+44 (0) 1793 785810

Hamed Amanzadeh
easyJet
Hamed.Amanzadeh@easyjet.com
+44 (0) 7909 730731

Abstract. The emergence of artificial intelligence (AI) with advanced natural language processing offers promising approaches for enhancing the capacity of textual classification. The aviation industry is increasingly interested in adopting AI to improve efficiency, safety, and cost efficiency. This study explores the potential and challenges of using AI to analyse decision errors in flight operations based on the HFACS framework. In pre-training, the model is trained based on a large amount of data to predict the next word in a sequence which allows the model to learn relationships between the words and their meaning in the accident investigation reports. Initial discoveries demonstrated that the AI model could supply a consistent HFACS framework and populate these dimensions with moderate accuracy. Future research is focused on the development of this HFACS-GPT model through fine-tuning and deep learning, facilitating more reliable and consistent conversations.

Keywords: Artificial Intelligence, Aviation Safety, Generative Pre-trained Transformer, Human Factors Analysis and Classification System, Large Language Model

Introduction

A generative pre-trained transformer (GPT) is a neural network type of large language model (LLM) and a prominent framework for generative artificial intelligence. Prompt engineering is the process of structuring text that can be understood by a generative model across a wide range of applications toward the desired outcome (OpenAI, 2023). GPT models are based on the transformer architecture which is designed for natural language processing (NLP) tasks and has been widely used in applications such as translation, text classification, and text generation. Language modelling is a key task in the GPT architecture, as it is performed during the pre-training phase of the model. In pre-training, the model is trained based on a large amount of data to predict the next word in sequence based on the previous words. It allows the model to learn relationships between the words and their meaning in the training data. Transformers allow the model to focus on different areas of the input while processing (Rahali & Akhloufi, 2023). NLP has been instrumental in making AI accessible to the public, for instance, it is the technology that powers chatbots, enabling them to generate human-like responses. Users can interact with these chatbots in a conversational manner, making the technology more approachable and user-friendly (Qin et al., 2023).

The Human Factors Analysis and Classification System (HFACS) is currently the most widely adopted human-factors coding framework used in accident and incident analysis. HFACS is based on Generic Error Modelling System (GEMS) and the subsequent 'Swiss Cheese' model, which places unsafe acts of operators within an organizational context (Reason, 1997). HFACS provides investigators the ability to categorize influencing effects that contribute to human decision-making and human errors, which can be accounted between 40% to 70% of accidents (Li et al., 2014). Neural networks have been used to model such diverse applications as predicting landing speeds, aircraft maintenance risk, and naturalistic pilot decision-making (Duggan & Harris, 2001). The model-building approach in NN analysis also reflects the trend in human factors away from statistical tests of difference and toward developing models of human behaviour. The essential feature of NNs is that they learn the relationship(s) between inputs and outputs and self-correct. They are trained by exposure to historical data in supervised learning with known inputs and outputs. In the supervised learning set, the model commences with a 'best guess' and applies a set of weights (Harris & Li, 2019).

The process of HFACS analysis is inherently time-consuming, requiring investigators to engage in thorough reviews of the mass of materials succeeding at classification and causal factor analysis. This frequently required domain experts of accident investigations who must sift through extensive accident-related information and data. Furthermore, the multi-layered structure of HFACS can introduce complexities that demand additional time for investigators to fully understand the relationships, from technical issues to operational environments (Hsieh et al., 2018). Decision-making in the aviation domain is a joint function of the features of the tasks and the pilots' knowledge and experience relevant to flight operations. Aeronautical decision-making (ADM) is defined as a systematic approach to the mental process used by aircraft pilots to consistently determine the best course of action in response to a given set of circumstances and has been involved in the majority of accidents (Li, Harris, & Hsu, 2014). This initial research aims to develop prompt engineering to assist investigators in expediting the classification of decision errors in flight operations using HFACS and GPT.

Method

Materials: For AI pre-training to conduct an HFACS analysis, two specific accident reports with a considerable scope of decision-error-related findings, which provide clear and trainable content for ChatGPT. These were the accidents of TransAsia Airways Flight 235 and Air France Flight 447, two accidents with a complex series of decision errors, related to cognitive workload and Crew Resource Management (CRM). With the available AI modelling, there are limitations to the ability of models like ChatGPT to process, predict, and output in response to text-based conversations from the user. Instead, these chatbots require an NLP network to convert natural language into tokens to make its predictions and formulate there meaning, much in the same manner as hearing a foreign language with no previous exposure (Lauriola et.al., 2022). Conversion of natural language into tokens allows the AI modelling to understand concepts of language such as grammar, and syntax to make logical responses (Perry, T, 2024).

Research framework: The scope of the research is focused on modelling, training, and conversation output of GPT models to conduct HFACS analysis based on a series of instructions and intentions that are pre-defined (Chen, B, 2023; Harris & Li, 2019). Pre-training remains as a supervised generation tool, requiring the human operator to continuously update and indicate the area of interest to the model, to generate an output more precisely to HFACS framework. The objective is to investigate the ability of unsupervised models to conduct HFACS analysis based on previous instructions and expectations of accident data, and then to create fine-tuning data based on the authors human evaluation of these responses.

Prompt engineering: AI chatbots such as ChatGPT require specific and informative instructions to perform the desired intentions required by the end user. We convert these expectations into prompts, goal-orientated snippets of instructions with clear and precise instructions on what steps the model must take to achieve the goal. Prompt engineering involves training the model using existing data to help categorize the text-based information when performing probability distribution for generating the highest probability sequence to follow the chat. We can re-engineer these prompts, based on outputs from the model to further improve this process and increase the accuracy of the data output. Below in Figure 1 is an extract from Python for one of the prompts used in initial training.

```
>>> response = client.chat.completions.create(
...   model="gpt-3.5-turbo",
...   messages=[
...     {
...       "role": "system",
...       "content": "Step 1: "The goal is to provide a complete and comprehensive HFACS a
nalysis of the attached content provided. Your answers should be long and detailed.
```

Figure 1: Source code for prompt integration into python.

Approach: Considerations to the parameter calibration are: Top_P of 0.2, Frequency Penalty of 0.2 and a Presence Penalty of 0.8. This research creates a framework for ChatGPT to perform its initial generation and training on two accident reports selected for this study. These were dedicated to 5 logical steps as follows:

Step 1: Create a legacy dataset for prompt generation and fine-tuning of the AI model system.

Step 2: Generate HFACS-specific prompts which will be used to analyse the report and extract key findings based on this framework.

Step 3: Generate a set of instructions which will help guide the AI model to conduct the analysis and generate the conversation with the users.

Step 4: Conduct irritative analysis on the AI model for the legacy dataset and store the responses to re-engineer prompt-specific guidelines for the AI model to use in the generation of text chat.

Step 5: Conduct fine-tuning on the AI model to improve its understanding of context and specialized classification for accident findings based on the HFACS framework and repeat Steps 1 – 4.

Results

The results of using HFACS-GPT to analyze flight GE 235 and AF 447 demonstrated key extracts and findings from the accident data with clear reference to the area of sourcing from the report. The presentation of information requested by the accident data also followed expected guidelines, indicating the level, sub-categories and a number of the findings within the HFACS framework. For the initial tests, the relevant decision errors were screened from the reports.

As we explored the content more thoroughly and analytically, it was seen that the repeated pre-training of classification on decision errors demonstrated significantly varied content output using the same prompt and accident report. In Figure 2, the top left image is represented by a temperature selection of 0.1, which returned an ambiguous finding of adverse weather conditions which was not reported as decision error within the original accident report. However, the conversation was consistent with its allocation of extract references from the report, demonstrating a consistent prediction logic of Top P. Findings from TransAsia Airways Flight 235 included: the inappropriate flight control inputs which resulted in a series of stalls, the failure to follow the documented engine failure procedures to identify and act to the failure, the decision to not reject take-off and the pilot's decision to disconnect the autopilot instead of diagnosing the problem. The accuracy of this conversation is

less than optimum based on report findings, which is likely a result of the model's attention to maintaining the highly probable sequence of structure for output. Supporting this theory was the results in Figure 2. While using a temperature selection of 0.5, the conversation demonstrated more reliable focus onto the accident report, addressing the active failure of the crew to stabilize the aircraft and correctly identify the failure engine before acting, which likely exacerbated the scenario. The accuracy of the 2nd batch of iteration is sufficient and provides a favorable result for prompt re-design and training. The 3rd batch of iteration, shown as the bottom left conversation in Figure 2, failed to demonstrate consistency to the subject matter. While initially the report finding was almost identical to the 2nd iteration, the extract reference was not sufficient to make the statements reported, despite the accuracy of this finding to the report. This iteration was considered at a temperature selection of 1, giving more creativity to its interpretation and presentation of output. This is evident at the lack of strict adhesion to HFACS.

Figure 2: Iterations of HFACS analysis using ChatGPT, based on two accident reports for TNA 235 and AF 44.

Initial findings for Air France 447 HFACS-GPT analysis demonstrated a strong adhesion to the structure requirements for all three batches of trials, with some clear indication of the significant decision-errors analyzed from the accident report. However, further analysis into the accuracy of the content was less than satisfactory. In the 1st batch of iteration, seen in Figure 2 on the top right, the model focused on the licensing of the crew and the condition of the aircraft. While its addressment of the blockage of pitot tubes was correct, it was not considered as a decision-error (technical failure induced pilot's confusion and resulted in decision errors) and should not have been classified as a decision error. The finding is supported by the extract from the report with clear reference to its location, indicating strong consistency between expectations and actual output. In the 2nd batch of iteration, seen in Figure 2 in as the middle right conversation, there was much more reliable data associated to a decision-error of the crew, focusing on the failure to initiate the 'Unreliable IAS'

procedure when first reporting that the airspeed indicator was not operating correctly. However, this iteration failed to provide an extract from the report to support this finding, rather giving the location of the finding for the user to source manually. The lack of consistency to the initial prompt are evident of the lower restrictive behaviour to formulate conversations, which improve the AI's ability when sourcing findings for the HFACS analysis. The 3rd batch of iteration, seen in Figure 2 on the bottom right, highlighted a less obvious and critical decision-error of the captain prior to taking his scheduled rest, which could have created the confusion over PF and PM, noted as having a considerable effect to crew workload during the accident sequence. The structure and presentation of content were well defined, demonstrating the level-1 of HFACS framework which was provided with plausible content. The quantity of content needs to be expanded to allow for deeper insight into the accident sequence and the effect of each finding.

Discussion and Conclusion

Early analysis of HFACS-GPT modelling for text-generated conversions seems to demonstrate an intellectual and analytical context of the accident reports supplied, although this is still dependent on a 'supervised management' of the model to instruct and refine to continue the conversation. The reiteration between the two accidents demonstrated good accuracy and consistency with the middle option for temperature settings, generally giving a good example corresponding to the classification of the textual report. The selection of these parameters was based on suggestions from the AI model developers, OpenAI, to categorize these parameters to generate specific responses. The research findings suggest that a temperature selection of 0.5 is optimum for achieving the desired balance between creativity and precision in text generation for HFACS classification. This is underscoring the importance of nuanced temperature control in leveraging GPT for specific applications. This finding will serve as a guideline for future pre-training on GPT model to balance the creativity and precision of the AI for HFACS framework. Future investigation would consider the selection of temperature values in 0.1 increments, to the maximum value of 2 to find an optimum selection to base on training materials. Additionally, manipulation of Top_P could allow finer control over the content generation by reducing the probability of possible sequences to the top 10% or lower. Expectations are that the higher threshold of probability will reduce the hallucination in the model and keep a stricter focus to the factual findings and conclusions made.

Additionally, the research will explore the application of HFACS analysis into segments for GPT analysis. Users will be able to specialize the HFACS analysis to focus onto a particular sub-dimension, such as Decision-Error or Supervisory Violation, or can have the ability to conduct hierarchical analysis such as choosing Level 1 or Level 2 for a conversation output. This specialty will also incorporate the visualization of impact effects to each sub-dimension or category, to further support the human evaluation of hierarchical influence in accident causation and paths of impacts. Future study would focus on human evaluation of GPT's conversations from accident investigators close to the accident cases, and industry experts specialised in HFACS analysis. The goal will be to generate fine-tuning/training material for GPT, through careful selection and generation of applicable conversations based on their adherence to HFACS. Further testing will investigate the reliability and overall improvement in output conversations provided by GPT, when covering multiple levels of HFACS. Further suggestions would focus on a more generic understanding of GPT's functionality to generate responses based on probability distributions of the tokens generated, and how to manipulate this probability to improve accuracy within the model. Finally, to have functionality of GPT to run in a continuous loop of processing responses to accident data and constantly improving its own learning capacity in an unsupervised application, to incorporate a wider range of end users to apply in different domains.

The limitation of current research lies in its narrow scope, as it exclusively focuses on the TransAsia Airways Flight 235 and Air France Flight 447 cases, potentially restricting the generalizability of findings to a broader context. In the assessment of generated text quality, without parallel human

supervision, there is an element of interpretative variability of the model's predicted outputs. Additionally, the trial predominantly manipulates the temperature parameter, neglecting a thorough exploration of the other relevant parameter configurations such as top p, frequency and presence penalty. A more extensive study design incorporating diverse topics, comparative analyses, human evaluations, and an exploration of parameter variability would enhance the robustness and applicability of the findings in the broader landscape of natural language processing research.

References

Aviation Safety Council (2016). Aviation Occurrence Report 4 February, 2015 TransAsia Airways Flight GE235 ATR72-212A. Report number: ASC-AOR-16-06-001, Aviation Safety Council, Taipei, Taiwan. https://reports.aviation-safety.net/2015/20150204-0_AT76_B-22816.pdf

BEA (2012). Final Report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro – Paris. BEA official report. https://bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf

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

Duggan, S., & Harris, D. (2001). Modelling Naturalistic Decision Making using an Artificial Neural Network: Pilot's responses to a disruptive passenger incident. Human Factors and Aerospace Safety, 1, 145-166.

Harris, D., & Li, W-C. (2019). Using Neural Networks to Predict HFACS Unsafe Acts from the Preconditions of Unsafe Acts. Ergonomics, 62(2), 181-191. https://doi.org/10.1080/00140139.2017.1407441

Hsieh, M., Wang, E. M., Lee, W., Li, L., Hsieh, C., Tsai, W., ... Liu, T. (2018). Application of HFACS, fuzzy TOPSIS, and AHP for identifying important human error factors in emergency departments in Taiwan. International Journal of Industrial Ergonomics, 67, 171–179. https://doi.org/10.1016/j.ergon.2018.05.004

Lauriola, I., Lavelli, A. & Aiolli, F (2022). An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools. Neurocomputing, Volume 470, pp 443-456. Elsevier. https://doi.org/10.1016/j.neucom.2021.05.103

Li, W. C., Li, L. W., Harris, D., & Hsu, Y. L. (2014). The Application of Aeronautical Decision-making Support Systems for Improving Pilots' Performance in Flight Operations. Journal of Aeronautics, Astronautics and Aviation, 46(2), 114-123. https://doi.org/10.6125/14-0324-789

OpenAI. (2023). GPT-4 Technical Report. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2303.08774

Perry, T. (2024). What is Tokenization in Natural Language Processing (NLP)?. Machine Learning. https://www.machinelearningplus.com/nlp/what-is-tokenization-in-natural-language-processing/?utm_content=cmp-true

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). Is ChatGPT a General-Purpose natural language processing task solver? arXiv (Cornell University). https://doi.org/10.48550/arxiv.2302.06476

Rahali, A., & Akhloufi, M. A. (2023). *End-to-end transformer-based models in textual-based NLP*. *AI, 4*(1), 54-110, pp. 54–110. https://doi.org/10.3390/ai4010004

Reason, J. (1997). Managing the Risks of Organizational Accidents. Aldershot, UK: Ashgate.

Biography

Dr. Wen-Chin Li: Dr. Wen-Chin Li is a Reader in Human Factor and Digital Aviation at the Safety and Accident Investigation Centre, Faculty of Engineering and Applied Sciences at Cranfield University, since 2013. He holds a PhD in Aviation Safety and Human Factors from Cranfield University. He is a Fellow of the Royal Aeronautical Society (FRAeS), a Fellow of the Chartered Institute of Ergonomics and Human Factors (FCIEHF), and an Aviation Human Factors Specialist at the European Association of Aviation Psychology (EAAP).

Declan Saunders: Declan is a research assistant at Cranfield University Faculty of Engineering and Applied Sciences since 2023, after successfully completing his master's in Safety and Human Factors in Aviation. Declan holds a bachelor's degree in Aeronautical Engineering where he won two awards from the Institution of Mechanical Engineers (IMechE) for his thesis research. Declan specializes in human factors research related to: digital tower operations, single-pilot operations, hydrogen integration to aviation, HMI design principles and AI for accident classification.

Hamed Amanzadeh: Hamed holds a bachelor's degree in mathematics and a master's in aerospace engineering. After graduating, he began his career in the airline industry, where he earned his B1 license in turbine aircraft maintenance. With extensive experience in aviation safety, data analysis, and flight data monitoring, he has developed expertise in the application of safety principles. Hamed later completed a second master's in Safety and Human Factors in Aviation from Cranfield University, and currently serves as a Safety Data Analyst at easyJet, focusing on enhancing safety outcomes through data-driven insights in aviation.