aerospace

MDPI

Article

Examining the Potential of Generative Language Models for Aviation Safety Analysis: Case Study and Insights Using the Aviation Safety Reporting System (ASRS)

Archana Tikayat Ray [1,*], Anirudh Prabhakara Bhat [2], Ryan T. White [3], Van Minh Nguyen [3], Olivia J. Pinon Fischer [1,*] and Dimitri N. Mavris [1]

[1] Aerospace Systems Design Laboratory, School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[2] AI Fusion Technologies, Toronto, ON M5V 3Z5, Canada
[3] NEural TransmissionS Lab, Department of Mathematics and Systems Engineering, Florida Institute of Technology, Melbourne, FL 32901, USA
* Correspondence: archanatikayatray@gmail.com (A.T.R.); olivia.pinon@asdl.gatech.edu (O.J.P.F.)

Abstract: This research investigates the potential application of generative language models, especially ChatGPT, in aviation safety analysis as a means to enhance the efficiency of safety analyses and accelerate the time it takes to process incident reports. In particular, ChatGPT was leveraged to generate incident synopses from narratives, which were subsequently compared with ground-truth synopses from the Aviation Safety Reporting System (ASRS) dataset. The comparison was facilitated by using embeddings from Large Language Models (LLMs), with aeroBERT demonstrating the highest similarity due to its aerospace-specific fine-tuning. A positive correlation was observed between the synopsis length and its cosine similarity. In a subsequent phase, human factors issues involved in incidents, as identified by ChatGPT, were compared to human factors issues identified by safety analysts. The precision was found to be 0.61, with ChatGPT demonstrating a cautious approach toward attributing human factors issues. Finally, the model was utilized to execute an evaluation of accountability. As no dedicated ground-truth column existed for this task, a manual evaluation was conducted to compare the quality of outputs provided by ChatGPT to the ground truths provided by safety analysts. This study discusses the advantages and pitfalls of generative language models in the context of aviation safety analysis and proposes a human-in-the-loop system to ensure responsible and effective utilization of such models, leading to continuous improvement and fostering a collaborative approach in the aviation safety domain.

Keywords: Aviation Safety Reporting System; ASRS; aviation safety; human factors; large language models; LLM; ChatGPT; generative language models; GPT-3.5; aeroBERT; BERT; InstructGPT; prompt engineering; NLP

Citation: 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). *Aerospace* **2023**, *10*, 770. https://doi.org/10.3390/aerospace10090770

Academic Editor: Peng Wei

Received: 4 July 2023
Revised: 2 August 2023
Accepted: 29 August 2023
Published: 31 August 2023

1. Introduction

The annual number of reported incidents within the Aviation Safety Reporting System (ASRS) has been consistently increasing [1], a trend that is anticipated to persist in the foreseeable future. This projected growth is largely attributable to the ease of submitting incident reports, the integration of novel systems such as Unmanned Aerial Systems (UAS) into the National Airspace System (NAS), and the increase in air travel overall. Because the current initial processing time of one incident report by two ASRS safety analysts can take up to five business days [2], new approaches are sought to help facilitate and accelerate such tasks. The development of hybrid human–AI approaches, particularly those involving the use of Large Language Models (LLMs), is expected to enhance the efficiency of safety analyses and reduce the time required to process incident reports [3].

Prior studies, including those mentioned in [4], have utilized the ASRS dataset in conjunction with Large Language Models (LLMs) such as BERT. However, the conclusions drawn from these studies are limited for a variety of reasons. The first is the imposed maximum narrative length of 256 or 512 WordPiece tokens when using BERT, which may potentially lead to information loss. For comparison, this specified length is below the 25th percentile of incident narrative lengths used for this study, which stands at 747 WordPiece tokens. The second reason is the fact that BERT requires specific training or fine-tuning on domain-specific data. Generative language models, on the other hand, can learn from a wider range of information due to their expansive training on larger datasets. This not only makes them more adaptable to evolving linguistic trends and domain shifts but also enhances their performance in zero-shot tasks, without requiring specialized fine-tuning.

The use of generative language models in the field of aviation safety remains largely unexplored. These models can serve as effective "copilots" or assistants to aviation safety analysts in numerous ways. They can automate the analysis of safety reports, identify patterns or anomalies that highlight potential safety issues, and identify potential risks based on historical data, hence aiding in the development of proactive safety strategies. Their proficiency in Natural Language Processing (NLP) can be harnessed for summarizing incident reports and extracting crucial information of interest. Furthermore, these models can be employed as training tools to simulate various scenarios or create synthetic data as a means to test safety measures or fill data gaps. However, their utility relies heavily on their training and implementation, and they should complement rather than replace human expertise.

In light of the considerable potential of generative language models, the primary objective of this work is to conduct a comprehensive assessment of the applicability and significance of generative language models such as GPT-3.5 (ChatGPT) [5] in the context of aviation safety analysis, specifically the ASRS dataset. In the context of the ASRS dataset, these language models hold the potential to serve as instrumental tools to aid human safety analysts by accelerating the examination of incident reports while simultaneously preserving the consistency and reproducibility of their analyses. In particular, this paper focuses on the following tasks:

1. Generation of succinct synopses of the incidents from incident narratives using ChatGPT.
2. Comparison of the faithfulness of the generated synopses to human-written synopses.
3. Identification of the human factors contributing to an incident.
4. Identification of the entity involved in the incident.
5. Providing explanatory logic/rationale for the generative language model's decisions.

The assembled dataset, which includes the ground truths, generated outputs, and accompanying rationale, can be found on the HuggingFace platform [6]. This accessibility allows for additional examination and validation, thereby fostering further advancements in the field.

This paper is organized as follows. Section 2 provides detailed information regarding the ASRS, introduces LLMs, and discusses the use of LLMs in the context of the ASRS dataset. Section 3 elaborates on the methodology implemented in this study, with a particular focus on the dataset used, prompt engineering, and the methodology used for comparing the generated outputs to the ground truths. Section 4 discusses the findings of this work; presents examples of incident narratives, synopses, and human factors errors; and discusses the evaluation of accountability. Lastly, Section 5 summarizes this research effort, discusses its limitations, and suggests potential avenues for future work.

2. Background

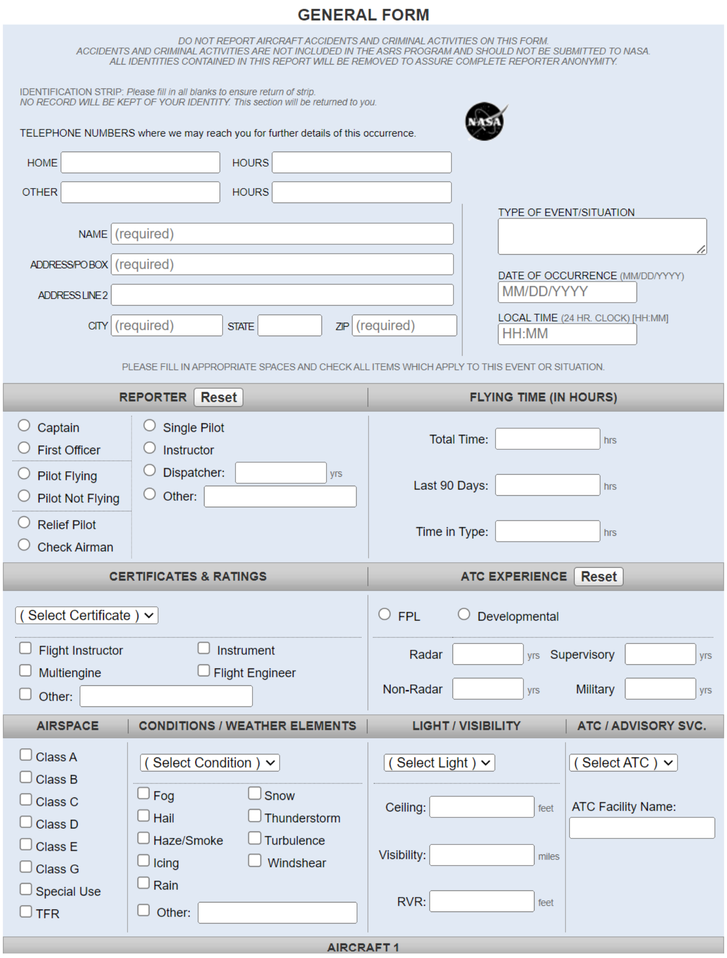
This section provides more information about the ASRS dataset and the way in which incident reports are gathered and analyzed by safety analysts to draw useful insights. This section also offers a comprehensive overview of LLMs as foundation models, specifically focusing on generative language models, as well as a discussion on the application of NLP in aviation safety analysis.

2.1. Aviation Safety Reporting System (ASRS)

The ASRS offers a selection of five distinct forms for the submission of incident reports by various personnel, as presented in Table 1. It is possible for multiple reports pertaining to the same event or incident to exist, which are subsequently merged by safety analysts at a later stage. A segment of the General Form for reporting incidents involving aircraft is depicted in Figure 1.

Table 1. The ASRS provides a range of five distinct forms for the submission of incident reports by different personnel. This can be accomplished through either an online form or an offline form, which is subsequently dispatched to the ASRS via postal mail [7].

Form Name	Submitted by
General Report Form	Pilot, Dispatcher, Ground Ops, and Other
ATC Report Form	Air Traffic Controller
Maintenance Report Form	Repairman, Mechanic, and Inspector
Cabin Report Form	Cabin Crew
UAS Report Form	UAS Pilot, Visual Observer, and Crew



GENERAL FORM

DO NOT REPORT AIRCRAFT ACCIDENTS AND CRIMINAL ACTIVITIES ON THIS FORM.
ACCIDENTS AND CRIMINAL ACTIVITIES ARE NOT INCLUDED IN THE ASRS PROGRAM AND SHOULD NOT BE SUBMITTED TO NASA.
ALL IDENTITIES CONTAINED IN THIS REPORT WILL BE REMOVED TO ASSURE COMPLETE REPORTER ANONYMITY.

IDENTIFICATION STRIP: *Please fill in all blanks to ensure return of strip.*
NO RECORD WILL BE KEPT OF YOUR IDENTITY. This section will be returned to you.

TELEPHONE NUMBERS where we may reach you for further details of this occurrence.

HOME HOURS
OTHER HOURS

NAME (required)
ADDRESS/PO BOX (required)
ADDRESS LINE 2
CITY (required) STATE ZIP (required)

TYPE OF EVENT/SITUATION
DATE OF OCCURRENCE (MM/DD/YYYY) MM/DD/YYYY
LOCAL TIME (24 HR. CLOCK) [HH:MM] HH:MM

PLEASE FILL IN APPROPRIATE SPACES AND CHECK ALL ITEMS WHICH APPLY TO THIS EVENT OR SITUATION.

REPORTER Reset
Captain
First Officer
Pilot Flying
Pilot Not Flying
Relief Pilot
Check Airman
Single Pilot
Instructor
Dispatcher: yrs
Other:

FLYING TIME (IN HOURS)
Total Time: hrs
Last 90 Days: hrs
Time in Type: hrs

CERTIFICATES & RATINGS
(Select Certificate)
Flight Instructor
Instrument
Multiengine
Flight Engineer
Other:

ATC EXPERIENCE Reset
FPL
Developmental
Radar yrs
Supervisory yrs
Non-Radar yrs
Military yrs

AIRSPACE
Class A
Class B
Class C
Class D
Class E
Class G
Special Use
TFR

CONDITIONS / WEATHER ELEMENTS
(Select Condition)
Fog
Snow
Hail
Thunderstorm
Haze/Smoke
Turbulence
Icing
Windshear
Rain
Other:

LIGHT / VISIBILITY
(Select Light)
Ceiling: feet
Visibility: miles
RVR: feet

ATC / ADVISORY SVC.
(Select ATC)
ATC Facility Name:

AIRCRAFT 1

Figure 1. This is part of the General Form used by pilots, dispatchers, etc., to report any incidents involving aircraft. The form contains fields asking about the *Reporter*, *Conditions/Weather elements*, *Light/Visibility*, *Airspace*, *Location*, *Conflicts*, *Description of event/situation*, etc., [8].

Figure 2 illustrates the pipeline for processing incident reports in the ASRS. The process begins with the ASRS receiving the reports in electronic or paper format. Subsequently, each report undergoes a date- and time-stamping procedure based on the receipt date. Two ASRS safety analysts then screen the report to determine its initial categorization and triage it for processing [2]. This screening process typically takes approximately five working days. Based on the initial analysis, the analysts have the authority to issue an *Alert Message* and share de-identified information with relevant organizations in positions of authority. These organizations are responsible for further evaluation and any necessary corrective actions [2].

Figure 2. The procedural flow for report processing commences with the submission of reports through either physical or electronic means. These reports are subsequently subject to scrutiny by safety analysts, and following the necessary de-identification procedures, they are integrated into the Aviation Safety Reporting System (ASRS) database [2].

Afterward, multiple reports related to the same incident are consolidated to form a single *record* in the ASRS database. Reports that require additional analysis are identified and entered into the database after being coded using the ASRS taxonomy. If further clarification is needed, the analyst may contact the reporter of the incident and any newly obtained information is documented in the *Callback* column. Following the analysis phase, any identifying information is removed, and a final check is conducted to ensure coding accuracy. The de-identified reports are then added to the ASRS database, which can be accessed through the organization's website. To maintain confidentiality, all original incident reports, both physical and electronic, are securely destroyed [2]. Table 2 shows some of the columns included in the ASRS dataset.

Table 2. Below is a list of columns from the ASRS dataset, along with additional accompanying information. This list is not exhaustive.

Column Name	Description
ASRS Record Number (ACN)	Unique identifier for each record in the ASRS database; **Example:** 881998, 881724, etc.
Date	The date on which the incident occurred is provided in a *yyyymm* format. This is done to de-identify incidents by removing "Day" information; **Example:** 201004, 201610, etc.
Local Time of Day	The incident time is categorized into specific time buckets to maintain anonymity and prevent the inclusion of exact incident times. These time buckets divide the 24-h period into four intervals; **Example:** 0001-0600, 0601-1200, 1201-1800, and 1801-2400
Human Factors	Human Factors in aviation refers to the discipline that examines the impact of human performance, cognition, and behavior on aviation incidents, with the aim of understanding and mitigating factors, such as human error, fatigue, communication breakdowns, and inadequate training, that contribute to accidents or near misses in the aviation industry; **Example:** Communication Breakdown, Confusion, Distraction, Fatigue, Human–Machine Interface, Situational Awareness, Time Pressure, Workload, etc.

Table 2. *Cont.*

Column Name	Description
Contributing Factors/Situations	The factors or circumstances that played a role in the incident's occurrence, as identified by the reporter (in the narrative) and/or safety analyst; **Example:** Human Factors, Environment Non-Weather-Related, Procedure, and Airspace Structure. Each incident can have multiple contributing factors.
Primary Problem	The main cause that led to the incident as identified by the safety analyst; **Example:** Human Factors, Environment Non-Weather-Related, Procedure, and Airspace Structure. However, each incident can have only one primary problem that led to the incident.
Narrative	The description of the incident provided by the reporter includes information about the chain of events, " *how the problem arose*", and various human performance considerations, such as perceptions, judgments, decisions, and factors affecting the quality of human performance, actions, or inactions; **Example:** A C680, checked on to frequency (very thick accent). I verified his Mode C and verified his assigned altitude of 11,000. I issued a 070 heading out of PVD VOR to intercept the Runway 4R localizer. He said 'roger, zero seven zero'. Moments later I noticed his altitude out of 10,000. I asked for an altitude verification and issued a climb. Then I pointed the aircraft out to the adjacent facilities who responded that there was no problem and point out approved. Continued with routine handling. Just a language barrier. Just a foreign pilot and language, although we use English as a common language in ATC, can be a barrier.
Synopsis	The summary of the incident written by safety analysts; **Example:** A90 Controller described a pilot error event when the flight crew of a foreign-registered aircraft descended below the assigned altitude during vectors to final.

In the ASRS database, information in different columns is populated either based on reporter-provided data or by safety analysts who analyze incident reports. For instance, the *Narrative* column is examined to populate related columns like *Human Factors*, *Contributing Factors/Situations*, and *Synopsis*.

With the increase in the number of incident reports over time, there is a need for a human-in-the-loop system to assist safety analysts in processing and analyzing these reports, which will help reduce the processing time, improve labeling accuracy [3], and ultimately enhance the safety of the NAS. LLMs, which are introduced in the section below, have the potential to help address this need.

2.2. Large Language Models (LLMs) as Foundation Models

This section provides an overview of LLMs and their pre-training and fine-tuning processes and highlights their significance as foundational models. Furthermore, it explores recent advancements in the field, with a particular focus on generative language models and their relevance to the present work on aviation safety.

LLMs, such as Bidirectional Encoder Representations from Transformers (BERT) [9] and the Generative Pre-trained Transformer (GPT) family [10–13], LLaMA [14], Llama 2 [15], LaMDA [16], and PaLM [17], are advanced NLP systems that have shown remarkable capabilities in understanding and generating human-like text. These models are built upon Transformer neural networks with attention mechanisms [18]. Neural networks, inspired by the functioning of the human brain, consist of interconnected nodes organized in layers that process and transform input data. The attention mechanism enables the model to focus on relevant parts of the input during processing, effectively capturing dependencies between different words and improving contextual understanding. Transformers' neural architectures have been particularly successful in NLP tasks, providing an efficient and effective way to process text.

The training process of LLMs involves two main stages: *pre-training* and *fine-tuning*. During pre-training, the model is exposed to vast amounts of text data from the Internet or other sources, which helps it learn patterns, grammar, and semantic relationships. This unsupervised learning phase utilizes large corpora of text to predict masked words, allowing the model to capture the linguistic nuances of the language. The pre-training stage often involves a variant of unsupervised learning called *self-supervised learning* [19], where the model generates its own training labels using methods such as Masked Language Modeling (MLM), Next-Sentence Prediction (NSP), generative pre-training, etc. This enables the model to learn without relying on human-annotated data, making it highly scalable. Unsupervised pre-training typically uses general-purpose texts, including data scraped from the Internet, novels, and other sources. This helps overcome the non-trivial cost and limits on scaling associated with data annotation.

After pre-training, LLMs are often fine-tuned on specific downstream tasks. This stage involves training the model on annotated data for NLP tasks like text classification, question answering, or translation. Fine-tuning allows the model to adapt its pre-learned knowledge to the specific requirements of the target task and domain, enhancing its performance and accuracy. The fine-tuning process typically involves using gradient-based optimization methods to update the model's parameters and minimize the discrepancy between its predictions and the ground-truth labels.

The general schematics of pre-training and fine-tuning LLMs are shown in Figure 3.

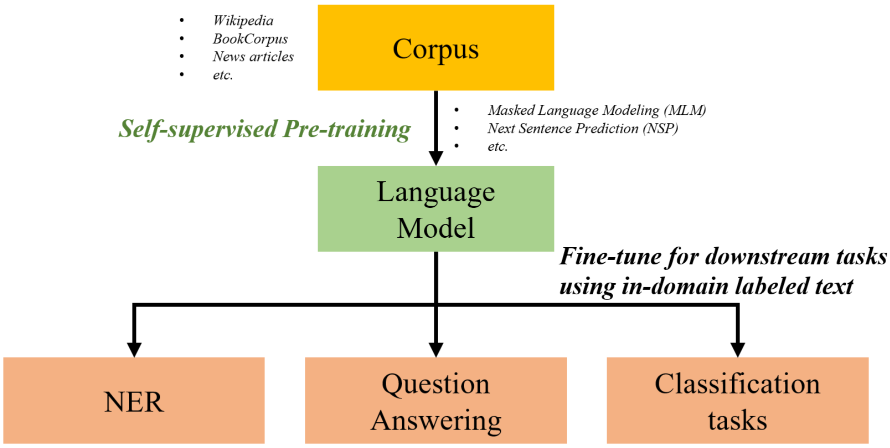


Figure 3. This figure demonstrates the training process of large language models (LLMs) in two stages: pre-training and fine-tuning. In the pre-training stage, the LLM learns from a large unlabeled corpus to capture language patterns and semantics. In the fine-tuning stage, the LLM is further trained on labeled corpora specific to downstream tasks, adapting its knowledge to improve performance in task-specific domains [20].

As mentioned, LLMs, including BERT and GPT, are often termed as *foundation models*. They provide the basis for an extensive range of specialized models and applications [21].

Fine-tuned models have been developed specifically for the aerospace field. These include, for example, aeroBERT-NER [20] and aeroBERT-Classifier [22], developed by fine-tuning variants of BERT on annotated aerospace corpora [20,22–24]. These models were designed to recognize aerospace-specific named entities and categorize aerospace requirements into different types, respectively [24].

The next subsection introduces a specific type of foundation model, namely generative language models.

2.2.1. Generative Language Models

An alternative to BERT's MLM and NSP pre-training is *generative pre-training*. This approach draws inspiration from statistical language models [25,26], which aim to generate text sequences by choosing word (token) sequences that maximize next-token probabilities conditioned on prior text. Neural language models [27] use neural networks to estimate the conditional next-token probabilities. During generative pre-training, a model is fed a partial sequence of text, with the remainder hidden from itself, and it is trained to complete the text. Hence, the corpora for generative pre-training can be unlabeled, as in the self-supervised training of BERT.

GPT [10] is a neural language model that employs a Transformer-based *decoder* [18] as its neural architecture. This is in contrast to the Transformer *encoder* of BERT, which is pre-trained with MLM and NSP. The decoder-only structure of GPT allows the model to perform diverse NLP tasks, such as classifying text, answering questions, and summarizing text, with minimal architectural changes.

In GPT-1 [10], generative pre-training is followed by task-specific supervised fine-tuning by simply adding a dense layer and softmax and fine-tuning for only a few training epochs. This approach is similar to BERT in that it requires sufficiently large annotated datasets for supervised fine-tuning. GPT-2 and GPT-3 place greater focus on zero-shot and few-shot learning, where the model must learn how to perform its tasks with zero or only a few examples of correct answers. GPT-2 [11] proposed conditioning its output probabilities on both the input and the desired task. Training a model for this multi-task learning through supervised means is infeasible, as it would require thousands of (dataset, objective) pairs for training. Therefore, GPT-2 shifts focus and demonstrates that strong zero-shot capabilities on some tasks without supervised fine-tuning can be achieved with a larger model (1.5B parameters). GPT-3 [12] scales GPT-2 up to 175B parameters, which greatly improves task-agnostic performance in zero-shot, one-shot, and few-shot settings without any supervised fine-tuning or parameter updates. It even outperforms some fine-tuned models, achieving state-of-the-art performance on a few NLP benchmarks.

However, GPT-3 has numerous limitations. It struggles to synthesize text by repeating itself, losing coherence in long-generated text, and including non-sequitur sentences. It lags far behind fine-tuned models in some NLP tasks, such as question answering. In addition, its responses to user prompts are not always aligned with the user's intent and sometimes show unintended behaviors, such as making up facts ("hallucinating"), generating biased or toxic text [28], and not following user instructions. These limitations stem from a fundamental incompatibility between the pre-training objective of generating the next token and the real objective of following user instructions safely and helpfully. InstructGPT (GPT-3.5) [5], the underlying LLM for ChatGPT by OpenAI, aims to correct this misalignment.

Since the mistakes GPT-3 makes are not easy to evaluate through simple NLP metrics, InstructGPT employs reinforcement learning with human feedback (RLHF) [29,30] after pre-training to dramatically improve performance. This is a three-step process:

1. **Supervised policy fine-tuning**: Collect a set of instruction prompts and data labelers to demonstrate the desired output. This is used for supervised fine-tuning (SFT) of GPT-3.
2. **Training a reward model**: Collect a set of instruction prompts, each with multiple different model outputs, and have data labelers rank the responses. This is used to train a reward model (RM) starting from the SFT model with the final layer removed.
3. **Optimizing a policy against the RM via RL**: Collect a set of prompts, outputs, and corresponding rewards. This is used to fine-tune the SFT model on their environment using proximal policy optimization (PPO).

Once InstructGPT fine-tunes GPT-3 through these steps, it becomes a standalone, off-the-shelf LLM that can effectively perform a diverse set of tasks based on text instructions, without the need for any additional training. Indeed, it has numerous benefits over GPT-3: labelers prefer InstructGPT outputs, it is more truthful and less toxic, and it generalizes better to tasks and instructions not seen during training.

In the context of this study, the terms InstructGPT, GPT-3.5, and ChatGPT are used synonymously, as they fundamentally represent the same technology utilized via an API.

The following section discusses the application of NLP more broadly in support of aviation safety analysis

2.2.2. NLP in Aviation Safety Analysis

There has been a decrease in the occurrence of incidents resulting from technical failures; however, incidents arising from human factors issues have emerged as the predominant underlying cause of the majority of incidents [31,32]. Several studies, such as [33–40], have looked into human factors issues in aviation. One of the most complex and difficult tasks when classifying aviation safety incidents is sub-classifying incidents stemming from human factors complications, which is a primary area of interest in this research.

Presently, incident/accident narratives are predominantly analyzed by safety analysts for the identification of factors that led to the incident/accident and to identify the root cause. The investigation conducted in [3] gathered labels from six individual annotators with aviation/human factors training, each working on a subset of 400 incident reports, culminating in a collective 2400 individual annotations. The outcomes indicated that there was a high level of disagreement among the safety analysts. This highlights the potential for LLMs to assist in incident classification, with subsequent verification by safety analysts.

In light of contemporary advancements in the field of NLP, numerous LLMs have been employed for the evaluation of aviation safety reports. The majority of research conducted in this sphere has largely concentrated on the classification of safety documents or reports [4,41].

In their study, Andrade et al. [4] introduced SafeAeroBERT, an LLM generated by initially training BERT on incident and accident reports sourced from the ASRS and the National Transportation Safety Board (NTSB). The model is capable of classifying reports into four distinct categories, each based on the causative factor that led to the incident. Despite its capability, SafeAeroBERT outperformed BERT and SciBERT in only two out of the four categories in which it was explicitly trained, thereby indicating potential areas for enhancement. In a similar vein, Kierszbaum et al. [41] proposed a model named ASRS-CMFS, which is a more compact model drawing inspiration from RoBERTa and is trained using domain-specific corpora. The purpose of the training is to perform diverse classifications based on the types of anomalies that resulted in incidents. From the authors' findings, it became evident that in most instances, the base RoBERTa model maintained a comparative advantage.

Despite the abundance of research and literature in the domain of aviation safety analysis [42,43], the application of generative language models remains largely unexplored within this field.

The following section discusses the dataset and methodology developed to demonstrate the potential of generative language models in the realm of aviation safety analysis.

3. Materials and Methods

This section details the dataset utilized in this work, the specific prompt employed, and the methodology adopted for interacting with ChatGPT.

3.1. Dataset

The ASRS database contains 70,829 incident reports added between January 2009 and July 2022. A total of 10,000 incident reports whose Primary Problem was labeled as *human factors* were downloaded for use in this study. This choice was motivated by the large number of incidents resulting from human factors compared to other causes [1].

3.2. Prompt Engineering for ASRS Analysis

This work leverages GPT-3.5 (via the OpenAI's ChatGPT API) [44] to analyze incident narratives, identify the human factors issues that led to incidents (Table 3), identify responsible entities, and generate incident synopses. As mentioned, the primary objective is to investigate and validate the potential uses of generative language models in the context of aviation safety analysis.

Table 3. List of human factors issues and their definitions. One or more of these factors can result in an incident or accident.

Human Factors Issue	Definition
Communication Breakdown	Failure in the exchange of information or understanding between pilots, air traffic controllers, or other personnel, leading to potential errors or safety issues in flight operations
Confusion	State where pilots, air traffic controllers, or other personnel are uncertain or lack clarity about flight information or procedures, potentially compromising flight safety or efficiency
Distraction	Any event, process, or activity that diverts attention away from a pilot's primary task of safely controlling the aircraft or prevents air traffic controllers from effectively managing flight operations
Fatigue	State of mental or physical exhaustion that reduces a pilot's ability to safely operate an aircraft or perform flight-related duties
Human–Machine Interface	Problems or difficulties in the interaction between pilots (or other personnel) and aviation equipment or systems, which can hinder operations and potentially compromise flight safety
Physiological—Other	Can include conditions like fatigue, hypoxia, barotrauma, dehydration, deep vein thrombosis, jet lag, spatial disorientation, effects of G-force, chronic noise and vibration exposure, radiation exposure, and disruptions to circadian rhythms, each resulting from the unique environmental and physical challenges of flight
Situational Awareness	Refers to a scenario where a pilot or crew has an incomplete, inaccurate, or misinterpreted understanding of their flight environment, which can potentially lead to operational errors or accidents
Time Pressure	Urgency or stress that pilots or air traffic controllers may experience when they have limited time to make crucial decisions or complete necessary tasks, often impacting safety and operational efficiency
Training/Qualification	Problems or challenges arising due to insufficient, inadequate, or improper training and certification of aviation personnel, including pilots, air traffic controllers, and maintenance crews, potentially impacting the safety and efficiency of aviation operations
Troubleshooting	Process of identifying and solving mechanical, technical, operational, or human factors-related problems that occur in the functioning of aircraft or in aviation operations, to maintain safety and efficiency
Workload	Tasks or responsibilities assigned to aviation personnel, such as pilots, air traffic controllers, or maintenance crews, exceed their capacity, potentially resulting in fatigue, errors, and safety risks
Other/Unknown	Problems, errors, or challenges occurring within aviation operations that cannot be readily categorized or identified under established categories and might require further investigation

Interacting with ChatGPT involved testing a variety of prompts before selecting the most suitable one. A prompt forms the initial input to a language model, shaping its subsequent output and significantly impacting the generated text's quality and relevance. Prompt engineering is the act of optimizing these prompts. This process refines the input to effectively guide the model's responses, improving its performance and output applicability. The temperature parameter of ChatGPT was set to near zero for this task. When the value is set to near zero, the output becomes predominantly predetermined and is well-suited for tasks that necessitate stability and yield the most probable outcome.

The initial step in the prompt engineering process involved assigning the *persona* of an aviation safety analyst to ChatGPT. Subsequently, ChatGPT was instructed to produce a brief synopsis based on the incident description. Initially, there were no restrictions on the length of the generated synopses, resulting in significant variations in length compared to the actual synopses. To address this, the lengths of the actual synopses were examined, and a maximum limit of two sentences was imposed on the generated synopses. Because the model appeared to omit the names of the systems involved at first, it was then specifically prompted to include system names and other relevant abbreviations.

Subsequently, the model was tasked with identifying various human factors issues responsible for the incident based on the provided incident narratives. While the model demonstrated the ability to identify these human factors issues, its responses exhibited significant variability due to its capability to generate highly detailed causal factors. This made it challenging to compare the generated responses with the ground truths, which encompassed twelve overarching human factors issues. Consequently, adjustments were made to the prompt to instruct the model to categorize the incidents into these twelve predefined classifications, where the model could choose one or more factors for each incident. Additionally, the model's reasoning behind the identification of human factors issues was generated via a prompt to provide an explanation for the decision made by the language model. Likewise, the model was directed to determine the entity accountable for the incident from a predetermined list of general options (such as ATC, Dispatch, Flight Crew, etc.) to prevent the generation of excessively specific answers, thereby facilitating the aggregation and subsequent evaluation process. The rationale behind these classifications was generated as well.

Lastly, the model was prompted to generate the output in a JSON format for which the keys were provided, namely *Synopsis*, *Human Factors issue*, *Rationale-Human Factors issue*, *Incident attribution*, and *Rationale-Incident attribution*. This was done to make the parsing of the model outputs easier. The structured format was then converted into a *.csv* file for further analysis using Python.

The prompt employed in this work can be found in Appendix A.

3.3. Analyzing ChatGPT's Performance

The *.csv* file generated was further analyzed to benchmark ChatGPT's performance against that of the safety analysts.

The quality of ChatGPT-generated incident synopses was analyzed first. Two approaches were taken to assess quality: (1) the similarity of ChatGPT-generated synopses to human-written synopses using BERT-based LM embeddings, and (2) the manual examination of a small subset of the synopses. These two approaches are discussed in more detail below.

3.3.1. Similarity Analysis Using BERT-Based LM Embeddings

When a sequence of text is fed to a BERT-based LM, it is first encoded by its pre-trained Transformer encoder into a numerical representation (i.e., an *embedding*). This same embedding may be fed to heads (or decoders) that perform various NLP tasks, such as sentiment analysis, text summarization, or named-entity recognition. Hence, this embedding contains deep syntactical and contextual information characterizing the text. For standard-size BERT models, the hidden representation of a text sequence made up of T WordPiece tokens is of the dimension $784 \times T$, and the row sum is commonly used as the embedding. Two text sequences are thought to be similar if their embedding vectors from the same LM have a similar angle, i.e., have high cosine similarity (close to 1), and dissimilar when they have low cosine similarity (close to 0). Hence, the cosine similarities of all pairs of human-written and ChatGPT-generated incident synopses were measured using several BERT-based LMs, including BERT [9], aeroBERT [20,22], and Sentence-BERT (SBERT) [45]. It is important to note that the latter uses a Siamese network architecture and fine-tunes BERT to generate sentence embeddings that the authors suggest are more suited to comparison by cosine similarity. This was followed by a manual examination of some of the actual and generated synopses to better understand the reasoning employed by ChatGPT for the task.

3.3.2. Manual Examination of Synopses

Subsequently, a comparison was made between the human factors issues identified by ChatGPT and those identified by the safety analysts. The frequency at which ChatGPT associated each human factors issue with incident narratives was taken into account. To visualize these comparisons, a histogram was constructed. Furthermore, a normalized multilabel confusion matrix was created to illustrate the level of concordance between ChatGPT and the safety analysts in attributing human factors issues to incident narratives. Each row in the matrix represents a human factors issue assigned to cases by the safety analysts. The values within each row indicate the percentage of instances where ChatGPT assigned the same issue to the corresponding narrative. In an ideal scenario where ChatGPT perfectly aligned with the safety analysts, all values except for those on the diagonal should be zero. Performance metrics such as precision, recall, and F1 score were used to assess the agreement between ChatGPT and the safety analysts.

Finally, an analysis was conducted regarding the attribution of fault by ChatGPT. The focus was placed on the top five entities involved in incidents, and their attributions were qualitatively discussed, supported by specific examples.

The subsequent section discusses the results obtained by implementing the methodology described in the current section.

4. Results and Discussion

The dataset of 10,000 records from the ASRS database was first screened for duplicates, as represented by the ASRS Case Number (ACN), and all duplicates were deleted, resulting in 9984 unique records. The prompt shown in Appendix A was run on each record's narrative via the OpenAI ChatGPT API, resulting in five new features generated by GPT-3.5, as outlined in Table 4.

Table 4. Features (columns) generated by GPT-3.5 (ChatGPT) based on ASRS incident narratives.

Generated Feature	Description
Synopsis	A synopsis of the narrative in 1–2 sentences that includes important details, such as the name of the system, and other relevant abbreviations, as necessary.
Human Factors Issue	A list of human factors issues predicted from the narrative, from the categories Communication Breakdown, Confusion, Distraction, Fatigue, Human–Machine Interface, Other/Unknown, Physiological—Other, Situational Awareness, Time Pressure, Training/Qualification, Troubleshooting, and Workload (mirroring the issues used in the ASRS), along with additional issues that ChatGPT was free to suggest.
Human Factors Issue (Rationale)	A 1–2 sentence description of the rationale ChatGPT used to decide which human factors issues were relevant.
Incident Attribution	An entity/entities to whom the incident can be attributed based on the narrative.
Incident Attribution (Rationale)	A description of the rationale ChatGPT used to attribute the incident to the specified party.

4.1. Generation of Incident Synopses

Two approaches were employed to assess the quality of ChatGPT-generated incident synopses: (1) the similarity of ChatGPT-generated synopses to human-written synopses using BERT-based LM embeddings, and (2) the manual examination of a small subset of the synopses.

As depicted in Figure 4, all the models (BERT, SBERT, and aeroBERT) found the synopses to be mostly quite similar. aeroBERT, with its fine-tuning on aerospace-specific language, evaluated the sequences as most similar, whereas the dedicated but general-purpose sequence-comparison model SBERT found the ChatGPT synopses to be similar but less so. SBERT is trained to excel in sentence similarity comparison, which inherently makes it more perceptive at distinguishing between sentences. This characteristic accounts for the wider distribution observed in the histogram. However, its limitation lies in the absence of domain-specific comprehension, leading to a lower median in the histogram.

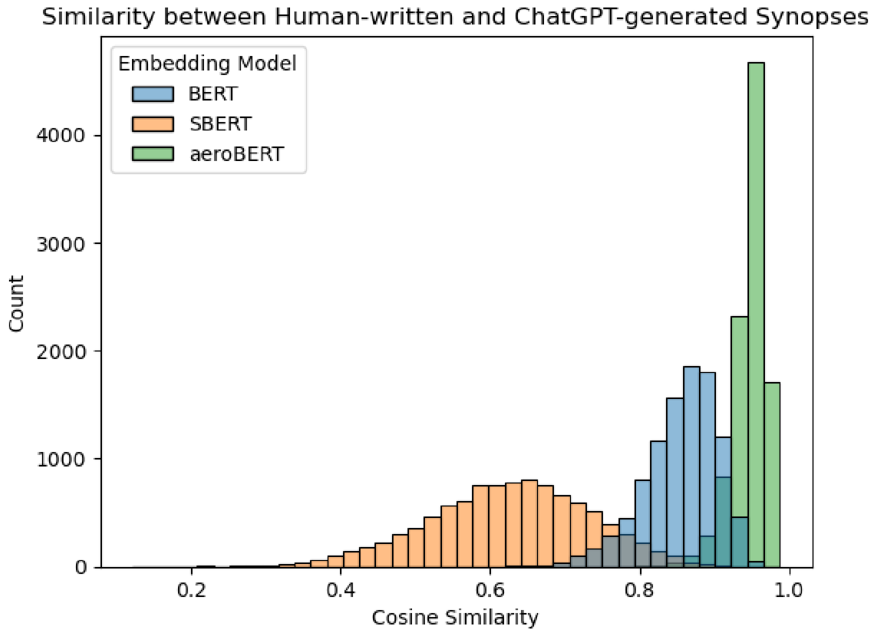


Figure 4. This histogram presents the cosine similarities computed by BERT, SBERT, and aeroBERT. Remarkably, aeroBERT, having undergone targeted fine-tuning on aerospace text, demonstrates a notably heightened level of similarity between the generated synopses and those composed by safety analysts in comparison to the other two models. Furthermore, aeroBERT exhibits superior proficiency in capturing semantic meaning beyond the mere lengths of the synopses being compared, contributing to its enhanced performance in capturing similarities within the aerospace domain.

Because the practicality of these similarities was not evident, the next step involved a manual evaluation of a sample of high- and low-similarity pairs of synopses from each of the three models.

Table 5 displays the top three synopses with the highest cosine similarity between the embeddings generated by each LM for the generated synopses and the synopses written by the safety analysts. Notably, the lengths of the generated and ground-truth synopses were very similar in this instance. As a general observation, the synopses generated by ChatGPT were found to be more comprehensive, presenting incident details in a chronological manner compared to the approach adopted by the safety analysts, who tended to include only the most critical elements in their synopses. For instance, upon analyzing the narrative for ACN 1759478, ChatGPT's synopsis encompassed the nuanced observation that the passenger was permitted to fly but subsequently removed upon arrival at the gate. In contrast, this specific detail was omitted in the synopsis generated by the safety analyst.

Regarding ACN 940308 (Table 5), an inaccuracy was observed where the model incorrectly abbreviated *Routine Overnight* as RON, whereas the correct abbreviation is *Remain Overnight*. This error arises from the incident narrative's usage of the incorrect abbreviation. To address such discrepancies, utilizing vector databases to store the meanings of abbreviations and other pertinent information could be beneficial. This approach would enable querying and extracting the appropriate information based on the context provided in a prompt, leading to more accurate and contextually relevant responses.

Similarly, Table 6 illustrates the three synopses with the minimum cosine similarity between the embeddings of the generated synopses and the corresponding synopses written by the safety analysts. A noteworthy observation from this data is the significant difference in length between the generated synopses and the ground truths for all three LLMs. The lengths of the generated synopses fluctuate depending on the narrative's length and the prompt's restriction to produce 1-2 sentences. In contrast, the ground-truth synopses are notably concise. Furthermore, there appears to be a degree of overlap for some of the synopses that obtained the lowest ranking in terms of cosine similarity. ACN 837587 "*A light twin and a single engine have a NMAC at Isla Grande*", for example, was ranked one of the lowest in terms of similarity when compared to the synopsis embeddings generated by both BERT and SBERT. However, the cosine similarity between the generated synopsis and the ground-truth synopsis, as determined using the aeroBERT embeddings, was computed to be 0.936. It is important to acknowledge that, typically, the embeddings produced by aeroBERT perceived the generated synopses and the ground-truth synopses as being more similar in comparison to those derived from BERT and SBERT.

The cosine similarities between human-written and ChatGPT-written synopses for aeroBERT, SBERT, and BERT show correlations of 0.25, 0.32, and 0.50, respectively, to the lengths of the human-written synopses (in words). This suggests that aeroBERT is less likely to attribute similarities based on the confounding variable of synopsis length.

Table 5. Table displaying the three top-matched synopses generated by the safety analysts and ChatGPT, as determined by cosine similarity (CS) scores calculated using different LLM embeddings.

ACN	Synopsis (Safety Analyst)	Synopsis (ChatGPT)	CS
	BERT Embedding		
963743	As a J3 slowed after landing the pilot over corrected a left drift so the aircraft veered right off the runway where the left wheel hit a prairie dog hole and broke the axle.	During landing at the home airport, the Piper J-3 Cub veered left and the pilot over-corrected, causing the aircraft to turn hard right and run off the right side of the runway. The left tire hit a prairie dog hole and bent the left axle.	0.963
1268182	A B-747 flight crew attempted to do a 180-degree turn on a runway due to taxiway closures. The Captain began the turn with about seven knots but was unable to complete the turn with that initial ground speed. The nose tire partially departed the runway and cables were used to get the Aircraft back on to the Runway. The aircraft was found to have no damage and cleared for flight.	The crew of Aircraft X attempted to perform a 180-degree turn on Runway 08, which was narrower than the minimum width required for the turn. The crew's decision to continue with the turn was influenced by the urgency of transporting high-priority cargo, including blood supplies. The aircraft skidded twice during the turn, and the crew decided to stop the aircraft and request tug assistance to complete the turn. The aircraft was eventually pulled back onto the runway and cleared for flight after inspection.	0.962
1230761	A Line Aircraft Maintenance Technician (AMT) reports finding the inboard end of a left aileron contacting a wing rib directly inboard of the aileron on a Cessna CE-560EP aircraft. Rivets and screws retaining the balance weight cover were gouging into the rib. Aileron had been installed three years earlier, but was found to be incorrect for the aircraft.	During a preliminary inspection of a Cessna CE560EP aircraft, the inboard end of the L/H aileron was found to be contacting the wing rib directly inboard of the aileron. Maintenance Control initially suggested grinding the aileron to get clearance, but a later E/O was issued to grind the rivet down and fly the aircraft under Ferry permit to get repaired properly. Three days later, it was discovered that the aileron was the wrong part number installed three years prior.	0.959
	aeroBERT Embedding		
1759478	Air carrier Captain reported a passenger appeared intoxicated during boarding, but was allowed to fly. During the flight the passenger was non-compliant with face mask policy.	During the flight, a passenger appeared to be intoxicated and was non-compliant with mask requirements. The flight crew notified customer service and had the passenger removed upon arrival at the gate.	0.987
1112916	An experienced pilot was distracted by conversation with a passenger and neglected to remove the tow bar during pre-flight. This was discovered when attempting to reposition into the hangar after the flight. The tow bar was found off the runway with no damage to the aircraft.	The pilot was distracted during pre-flight and failed to remove the aircraft tow bar. The tow bar was found 1,800 FT from the departure end of the runway after the flight. No apparent damage was found on the aircraft.	0.987
1158337	After landing in MIA, a fatigued flight crew realized that they were on Approach Control frequency and had not received landing clearance.	The flight crew forgot to switch to the Tower frequency and landed without clearance due to a lack of situational awareness and fatigue.	0.985

Table 5. *Cont.*

ACN	Synopsis (Safety Analyst)	Synopsis (ChatGPT)	CS
		SBERT Embedding	
1011027	A pilot reported loosing control of a Maule Amphibian on takeoff after he lifted off to soon in a crosswind causing a wing to stall and forward parts of the floats to contact the runway.	The pilot lost control of a Maule Amphibian during takeoff in strong crosswind conditions, causing the aircraft to contact the ground with the forward parts of the floats.	0.924
917365	A Line Mechanic was informed that a lock-pin from the aft service door evacuation slide inflation valve had not been removed after he replaced a MD-88 door slide. The lock-pin was allegedly found six months later during a cabin maintenance visit.	During a cabin maintenance visit, it was found that the lock pin from the aft service door evacuation slide inflation valve had not been removed after the replacement of evacuation slides in a MD-88 aircraft.	0.921
940308	A Line Mechanic loosely installed mount bolts during replacement of a Center Tank Fuel Quantity Transient Suppression Device (TSD) on a B737-300 aircraft to allow another Mechanic to connect a wire harness to the TSD. Mechanic realized after the aircraft departed that the mount bolts may not have been tightened.	During maintenance work on a B737-300 aircraft, a lack of communication between two Aviation Maintenance Technicians (AMTs) led to the temporary installation of the Center Fuel Quantity Transient Suppression Device (TSD) box with mount bolts finger tight. The AMT who installed the TSD harness was not informed about this, leading to loose bolts. The issue was later corrected by another AMT during a Routine Overnight (RON) inspection.	0.920

Table 6. The three worst-matched synopses generated by the safety analysts and ChatGPT, alongside their cosine similarity scores calculated using different language model embeddings.

ACN	Synopsis (Safety Analyst)	Synopsis (ChatGPT)	CS
		BERT Embedding	
837587	A light twin and a single engine have a NMAC at Isla Grande.	The pilot was unable to communicate effectively with the tower due to a congested frequency, resulting in a near mid-air collision with another aircraft on converging flight paths.	0.559
1889454	Air carrier ground personnel reported Hazmat cargo was transported on wrong aircraft.	Hazmat intended for a later mainline trip was mistakenly loaded onto an express trip.	0.621
1231062	CE525 Captain expresses confusion over RNAV SID's STAR's after modification by ATC, in general and the UECKR1 specifically.	The flight crew received conflicting instructions from ATC during clearance, which led to confusion and misinterpretation of the clearance.	0.623
		aeroBERT Embedding	
1671165	C172 Pilot reported a NMAC while training.	During air work, the pilot was distracted by maintaining airspeed and bank angle while performing an emergency descent, leading to a lack of situational awareness. The pilot did not take sufficient steps to confirm the other aircraft's position, and the two aircraft passed each other with a relatively small clearance.	0.890
1878408	Small Aircraft Instructor Pilot reported a NMAC.	On DATE at XA:20, a flight crew and their student were flying in the Northeast Practice Area in Arizona when they encountered another aircraft, Aircraft Y, that was not following the right of way standard right turn to avoid traffic. The incident was caused by a communication breakdown between the flight crew and the other aircraft on the NE Practice area frequency.	0.882
1638197	C310 pilot reported flying VFR into IMC.	The pilot cancelled IFR based on the belief that VMC prevailed in the area, which was not the case. The pilot had an unreasonable belief that the weather would improve and did not consider requesting immediate IFR clearance or performing a 180 turn. The incident was caused by the pilot's decision-making and actions.	0.894
		SBERT Embedding	
837587	A light twin and a single engine have a NMAC at Isla Grande.	The pilot was unable to communicate effectively with the tower due to a congested frequency, resulting in a near mid-air collision with another aircraft on converging flight paths.	0.123

Table 6. *Cont.*

ACN	Synopsis (Safety Analyst)	Synopsis (ChatGPT)	CS
1409509	C172 pilot reported an NMAC in the vicinity of IGX airport.	During a VFR flight utilizing flight following, the pilot failed to set the ascent rate on the autopilot, causing the aircraft to hold the current altitude instead of climbing to the requested altitude of 4500 feet. This led to a potential conflict with another aircraft, and the pilot had to descend further to maintain separation.	0.162
1028402	EMB145 Captain describes the factors that resulted in missing a crossing restriction during the GIBBZ1 RNAV arrival to IAD.	The flight crew deviated from a new arrival procedure due to the First Officer's inexperience in the aircraft and uncertainty about an acceptable rate of descent. The Captain's focus on the next waypoint ahead of the current one led to a delay in realizing the aircraft was too high.	0.218

4.2. Performance with Human Factors-Related Issues

In this section, we evaluate how the safety analysts and ChatGPT compare in attributing human factors issues to ASRS incidents. ChatGPT was prompted (see Appendix A) to attribute the human factors issues listed in Table 3 to each incident narrative and to explain its rationale in doing so.

Initially, an assessment was made regarding the rate at which ChatGPT associated each human factors issue with incident narratives. When compared with the assessments made by the safety analysts, it was observed that ChatGPT attributed human factors issues to narratives less frequently, with the sole exception being the *Training/Qualification* category, which is a systemic issue that ChatGPT assigned with a higher frequency (Figure 5). ChatGPT's propensity to frequently assign the *Training/Qualification* category might be associated with its algorithmic interpretation of the narratives. Since the model is designed to identify patterns and make decisions based on those patterns, the prevalence of *Training/Qualification* assignments could potentially reflect patterns within the training data related to language usage or context. Furthermore, *Training/Qualification* issues might be more straightforward and explicit in the narratives, leading to their higher representation. These elements could appear more clearly and concretely in the text, leading to their more frequent identification by the model. This does not necessarily imply a bias in the model but rather illustrates the intricacies involved in training a language model to understand and interpret human factors in complex real-world scenarios.

Notably, ChatGPT rarely assigned *Confusion*, *Human–Machine Interface*, *Other/Unknown*, and *Troubleshooting*. These human factors issues can be relatively nuanced and involve a complex interplay of human behavior and situational context. These elements could potentially be more challenging for the model to consistently identify due to their complexity and the subtlety with which they may be conveyed in the text. It should be emphasized that there were variations in the human factors issues identified by the safety analysts due to subtle differences in categories like *Confusion*, *Situational Awareness*, *Fatigue*, *Physiological—Other*, *Time Pressure*, etc., due to their broad definitions. For instance, in the case of the narrative pertaining to ACN 1758247 presented in Table 7, one safety analyst could potentially identify the related human factors issue as *Fatigue*. Conversely, another analyst might designate the issue as falling under the category of *Physiological—Other*, which was the case in this specific instance. The differences in these attributions underscore the challenge of categorizing complex human factors in aviation safety, as categories such as *Situational awareness*, *Confusion*, *Time Pressure*, and *Physiological—Other* can be broadly interpreted and may overlap. For instance, *Time Pressure* could result in *Confusion*, and both could affect *Situational awareness* and lead to various *Physiological—Other* conditions, including fatigue.

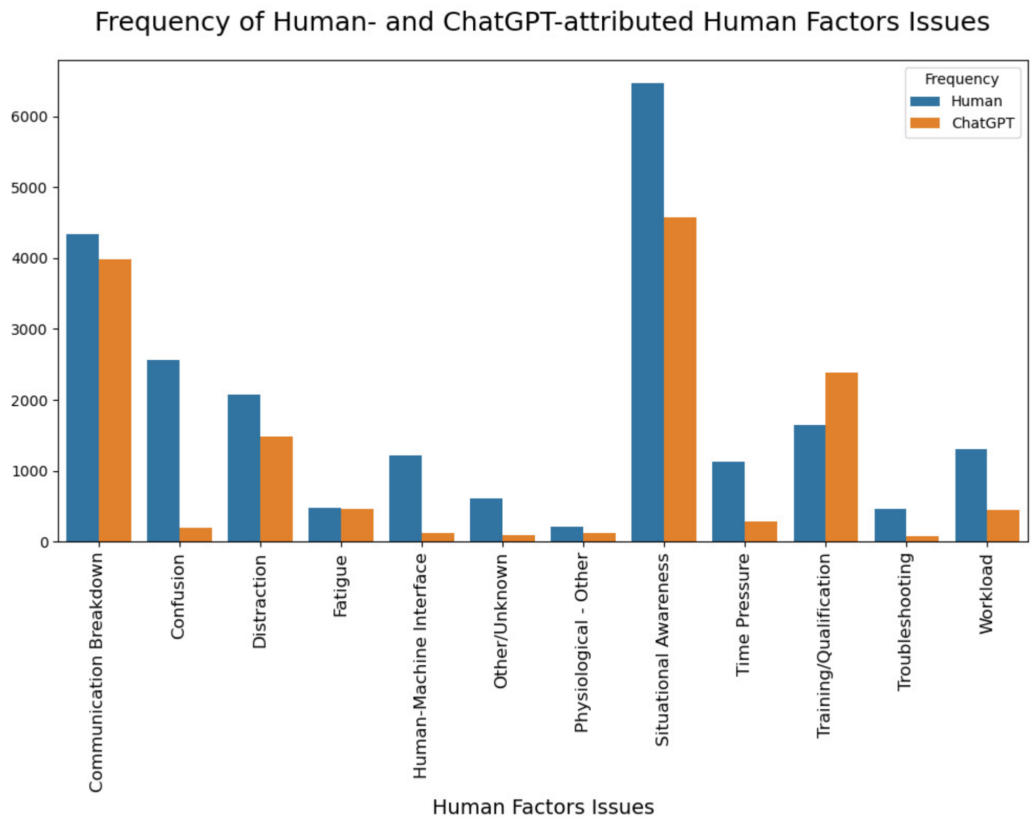


Figure 5. Frequency of human-attributed and ChatGPT-attributed human factors issues

Table 7 depicts a sample of five incident narratives where the safety analyst and ChatGPT did not agree on the identification of the human factors issue. For the narrative associated with ACN 1013382, a safety analyst attributed the installation of an incorrect Flight Management System (FMS) database to *Human–Machine Interface* and *Situational Awareness*, emphasizing the pilot's interaction with the system and their failure to fully grasp the situation. In contrast, ChatGPT classified the cause as *Training/Qualification*, pointing toward potential inadequacies in the pilot's training, specifically the thorough verification of the installed database. These different attributions underscore the complex nature of aviation safety issues, with the safety analyst's perspective focused on in-flight actions and interactions, whereas ChatGPT emphasized the need for comprehensive pre-flight training and checks.

Similarly, for ACN 834159, there was a discrepancy between the incident cause attribution assigned by the safety analyst and ChatGPT. The safety analyst identified the cause as *Troubleshooting*, likely focusing on the situation's process aspect and the decision to continue with the journey despite visible issues. In contrast, ChatGPT attributed the incident to *Maintenance*, possibly due to the mechanic's involvement and the re-occurrence of the problem, hinting at an underlying maintenance issue that was not resolved initially. The distinction in the interpretations might be due to the subjective nature of human analysis, which can lead to variations in classification among safety analysts. An analyst's interpretation may be influenced by their individual experiences, expertise, and perception of what constitutes the core issue in a given scenario. On the other hand, ChatGPT classified the issue under 'Maintenance' by associating keywords like 'mechanic', 'condensation', 'window', and 'blocked' with a maintenance problem, a classification approach that remained constant across various scenarios. This method can potentially lead to a more uniform and consistent categorization of incident causes, which is crucial for trend analysis and effective mitigation strategies.

Table 7. Five incident narratives where discrepancies exist between the human factors issues identified by the safety analyst and those recognized by ChatGPT.

ACN	Incident Narrative	Label (Analyst)	Label (ChatGPT)
1758247	After reporting at XA:15 for a XB:45 pushback to ZZZZ, we were delayed for a maintenance issue. Our CCO time was XD:28, we waived and agreed to extend for 2 h until XF:28. The time ran out and we were delayed because of legality issues. It took us 1 h to get to the hotel, because of late night and COVID-19 procedures. I entered my room at XG:30. I tried to find some food because the intention was to complete the 15 h flight to ZZZZ. I found no food so elected to try to rest. I rolled around trying to decompress and unwind from the on again off again rolling delay of the maintenance issue. I finally fell asleep around XJ:00. At XK:00, the hotel started to come alive. There were numerous door openings and closings. They were obviously crew members leaving for work. Not 45 min after numerous crew members departed for work housekeeping arrived to clean and disinfect the rooms. It seemed like due to COVID-19 the process took longer and was much louder than it should have been. By that time it was XN:00 and I had slept for maybe 60 min. I decided I was not fit to fly due to operational caused fatigue and informed the crew desk at XO:00.	Physiological—Other	Fatigue
1568978	(It was a) normal pattern to a base leg. After deploying flaps, the boost pump was accidentally switched on, which flooded the engine and caused a total loss of engine power. Due to the low altitude and the flaps being fully deployed, the aircraft could not glide to the runway, so I made the decision to take control and land off to the left side just short of the runway. The aircraft stopped rolling approximately 25 feet before the runway threshold. There appears to be no damage to the area where the landing occurred or to the aircraft itself. There is no sign of mechanical failure, only pilot error.	Other/Unknown	Distraction
1021950	July 2012. On Aircraft X, an ERJ-190 aircraft, I removed and replaced the Captain's Pitch-Trim switch in accordance with Aircraft Maintenance Manual (AMM). Did a Return to Service and Operational Test per AMM. Operational Check good, MEL was cleared. Two days later, Pilot in ZZZ1 reported Captain's yoke Elevator Trim (Pitch Trim) switch operates opposite to input. In ZZZ1, Captain's [Pitch] Trim switch was re-installed per AMM and Operational Check good, OK for service.	Communication Breakdown; Confusion; Situational Awareness; Troubleshooting	Maintenance
1013382	I was informed by operations that the aircraft that I flew had the wrong database installed in the FMS. It had a B777 database instead of an MD11 database. I did check the date of the database but did not check the numbers at the top to verify correct installation. After seeing the same thing at the top of the screen I rarely verify the correct database installed! We proceeded to destination without any issues or ATC questions so I am assuming the data base was close enough to operate a MD11. Have Maintenance verify the aircraft type and database type before installation then reverify after load.	Human–Machine Interface; Situational Awareness	Training/Qualification
834159	When I was doing my safety check I noticed that the window on door 2 L was covered in condensation. The mechanic came on removed the interior window and wiped off the condensation. He replaced the interior window stated it would be deferred and was 'good' for 100 h. Upon landing at our destination the window was again 100 percent blocked. 1R was 100 percent blocked and 2R was about 60 percent blocked. The plane continued to be released despite the serious safety concerns of the flight attendants. We went on to our next destination and again upon landing 2 L window was 100 percent blocked.	Troubleshooting	Maintenance

Next, a normalized multilabel confusion matrix [46] was generated, as shown in Figure 6. The confusion matrix helps visually represent the agreement (or disagreement) between the human factors issue attributions made by ChatGPT and those made by the safety analysts for each record in the ASRS database. Each row corresponds to a human factors issue as identified by the safety analysts. The values in each row signify the percentage of times ChatGPT attributed the same issue to the incident narrative as the safety analyst. The darker shades of blue indicate a higher agreement between the ChatGPT and safety analyst attributions. In an ideal scenario, if ChatGPT's assessments perfectly match those of the safety analysts, all the values outside the diagonal should be zero, depicted by the color white.

We note that ChatGPT agreed with the safety analysts more than 40% of the time for the following classes: *Communication Breakdown*, *Fatigue*, *Situational Awareness*, and *Training/Qualification* (note the dark shades of blue on the diagonal for these issues). It agreed substantially less for the other classes. The most common source of disagreement was when ChatGPT chose not to attribute any human factors issues (as shown in the right-most column of the multilabel confusion matrix in Figure 6, where moderate shades of blue are common). In addition, ChatGPT attributed many more incidents to *Communication Breakdown*, *Distraction*, *Situational Awareness*, and *Training/Qualification* compared to the safety analysts (these columns are shaded a bit more than the other columns).

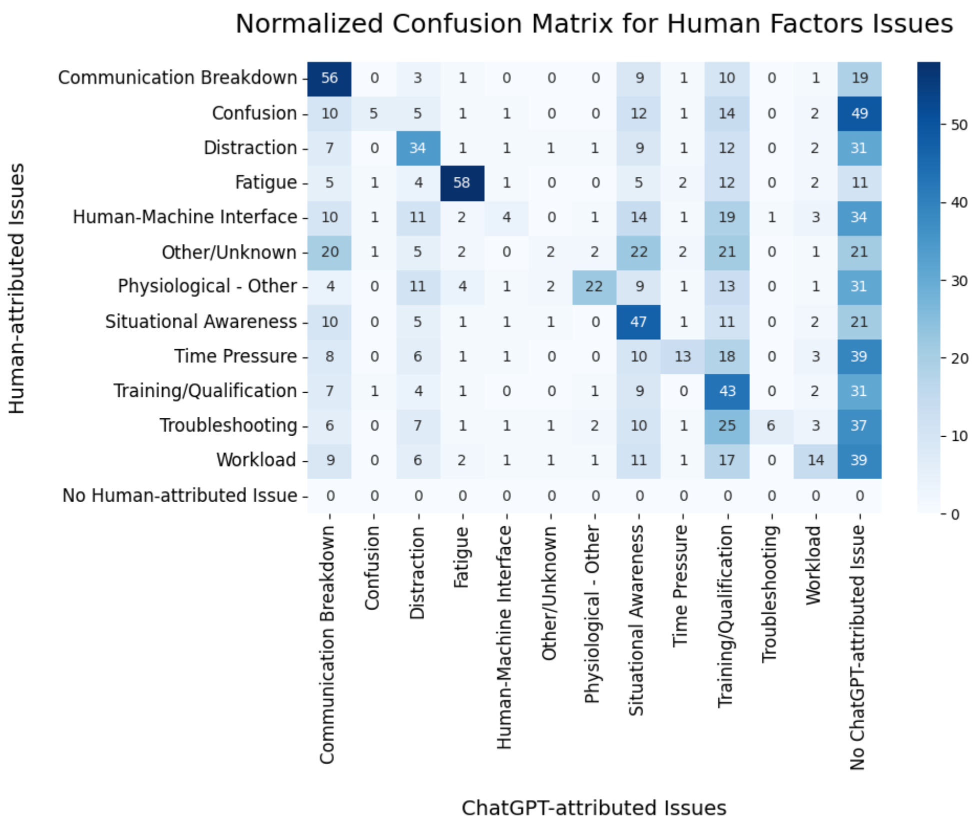


Figure 6. Normalized multilabel confusion matrix for human factors issues.

The lower precision and recall values achieved by ChatGPT (Table 8) in attributing human factors issues, compared to the safety analysts, can be understood through various perspectives that highlight both the strengths and limitations of humans and LLMs.

Firstly, it is important to acknowledge the subjectivity of human interpretation. Safety analysts, with their individual experiences and knowledge, may have varied interpretations of the same incident narrative, which can lead to inconsistencies in the identification of human factors issues. The process relies heavily on personal judgment, leading to potential inconsistencies in classifications (Table 7), even among different analysts.

Secondly, the identification of human factors issues in aviation incidents is inherently complex. A single event can trigger a cascade of other factors, and the precise categorization of these factors can be difficult due to their interconnectedness. On the other hand, ChatGPT, which is trained on vast amounts of data, offers a more uniform approach to categorization. It identifies patterns in the narratives and bases its attributions on these patterns, leading to better consistency. However, ChatGPT is generally more conservative in assigning human factors issues compared to safety analysts, often missing factors that a safety analyst may deem significant, thus leading to lower recall values.

The precision value of 0.61 implies that when ChatGPT identified a human factors issue, safety analysts agreed with this categorization around 61% of the time. The variations in recall and precision values between ChatGPT and the safety analysts indicate that there is room for growth and fine-tuning in the model.

Table 8. Classification report of ChatGPT predictions vs. human-attributed human factors issues.

Class	Precision	Recall	F1 Score	Support
Communication Breakdown	0.67	0.62	0.64	4332
Confusion	0.67	0.05	0.1	2570
Distraction	0.53	0.38	0.44	2072
Fatigue	0.71	0.69	0.7	481
Human–Machine Interface	0.44	0.04	0.08	1210
Other/Unknown	0.19	0.03	0.05	609
Physiological—Other	0.42	0.25	0.32	208
Situational Awareness	0.74	0.52	0.61	6475
Time Pressure	0.59	0.15	0.24	1132
Training/Qualification	0.32	0.47	0.38	1649
Troubleshooting	0.45	0.07	0.12	455
Workload	0.48	0.16	0.24	1305
Weighted Average	0.61	0.38	0.43	22,498

The integration of human expertise with LLM capabilities presents a promising solution to this challenge. While safety analysts bring experience-based insight and intuition to the table, ChatGPT offers large-scale, consistent data analysis. Together, they can enhance the accuracy and efficiency of incident-cause attribution in aviation safety, leveraging both the strengths of human judgment and the precision of LLMs. This collaborative approach can help ensure a more comprehensive, balanced, and consistent analysis, enhancing the overall effectiveness of safety measures in aviation.

4.3. Assessment of Responsibility

ChatGPT was also leveraged to ascertain the involvement of different entities in the event. It is essential to clarify that the goal here was not to impose penalties on the entities identified but to harness these specific occurrences as rich resources to bolster the base of aviation human factors safety research. This becomes especially significant considering the widely accepted notion that human performance errors underpin more than two-thirds of all aviation accidents and incidents.

The responses from ChatGPT were not strictly confined to the options offered in the prompt. Instead, it skillfully produced associated entities based on the provided narrative. Additionally, in instances where involvement could be distributed among multiple entities, the model exhibited its ability to effectively identify such cases. Table 9 highlights the top five entities, as identified by the model, involved in incidents tied to human factors.

In the scope of this effort, the 'flight crew' included all personnel on board the flight, except for passengers. A collective number of 5744 incidents were linked to the flight crew. The model was also capable of identifying incidents arising from Air Traffic Control (ATC), with the model demonstrating its ability to recognize distinct entities, such as ATC (Approach Control), ATC (Ground Control), ATC (Kona Tower), ATC (Military Controller), ATC (TRACON Controller), ATC (Indy Center), and so on. The model's ability to specify these particular towers can be explained by the presence of that information in the incident narratives, as those details were not anonymized. These diverse categories were consolidated into a singular ATC classification by the authors.

Table 9. List of top five entities identified by the model, as associated with incidents where the predominant contributing factor was determined to be human factors.

Entity Associated with Incident	Count
Flight Crew	5744
ATC	1838
Ground Personnel	805
Maintenance	738
Flight Crew, ATC	133

To fully grasp the mechanism behind ChatGPT linking a specific event to a particular entity, the model was prompted to articulate the underlying reasoning behind each association. Table 10 shows various examples of this to promote a comprehensive understanding of the principles that influence association reasoning. A meticulous examination of certain event reports revealed that the entity identified by ChatGPT, along with the rationale behind the identification, demonstrated a commendable degree of precision in the model's responses and justifications. For instance, as depicted in the initial case (ACN 1805938) shown in Table 10, the lead Flight Attendant, overly concerned with a passenger's non-compliance with mask policies, violated sterile cockpit procedures by alerting the flight crew via a chime during landing rollout to request a supervisor, a gesture typically associated with emergencies or critical aircraft issues. As a result, the linking of the incident to the flight crew by ChatGPT seems suitable. Furthermore, the reasoning behind this attribution is well-founded.

Table 10. Three incident narratives, the assessment of responsibility, and the rationale used by ChatGPT for the attribution.

ACN	Incident Narrative	Incident Attribution (ChatGPT)	Rationale (ChatGPT)
1805938	The lead Flight Attendant was so uptight with a passenger not complying with the Mask policies, (Flight Attendant) thought it was prudent to chime the cockpit during landing rollout and advise the flight crew to notify a Supervisor to meet the aircraft. This was a direct violation of sterile cockpit procedures. Because flight crews repeatedly experience chimes from the cabin while in simulator training, they are most associated with fire or critical aircraft issues. This is never a good time to chime the cockpit unless it is an emergency. Especially for a mask compliance issue. Relaxed emphasis from Leadership on the tattling on passengers for improper use of masks (never saw this type of behavior from FA when it came to early seat-belt releases. Reissued sterile cockpit etiquette with a bulletin on scenario that are and are not acceptable for chiming the cockpit during landing rollout.	Flight Crew	The flight attendant's actions caused a distraction to the flight crew during landing rollout.
874307	I was working Aircraft X inbound to ADS. Along with this aircraft I had several other VFR pop up aircraft that had to be low level to remain VFR, typically around 1500 feet. Aircraft X was at 2000 on a vector for an ILS approach approximately 15 miles southeast of ADS. I observed the aircraft in proximity to a 2200 foot MVA. I, for some reason, registered that the aircraft was VFR and instructed the aircraft to maintain VFR at or below 2500 so that he could maintain obstacle clearance. The pilot complied and descended to 1600 to VFR conditions. I pointed out the antenna and the pilot reported it in sight. It was then that I realized that the aircraft was in fact IFR and I climbed him immediately to 3000. By that time the aircraft was inside the MVA at 1600. Recommendation, this event could have been avoided had I had a better scan. It was complicated with bad weather and having to watch aircraft that were deviating and descending/climbing in areas where they would not normally be being worked by other positions.	ATC	The incident was caused by the controller's error in providing incorrect instructions to the pilot.
868384	Aircraft was presented to Crew in ZZZ with a deferred item open, stating the forward coffee maker and spigot were inop'ed at a previous Station and the deferred MEL included draining the A320 of potable water and deferring the entire water system. While reviewing the Maintenance Release, the First Officer reported the forward lavatory toilet flushed normally. Water (H20) quantity was checked and it was found to be full, in direct contrast to the MEL instructions for the deferral on the Maintenance Release. Maintenance Control was contacted and a Maintenance Report item sent. Contract Maintenance called to aircraft in ZZZ and he reinstated the water system as no faults could be located. A few other write-ups were handled by ZZZ Maintenance Technician and he left the aircraft, with the Deferral placard still located on the forward Cabin Intercommunication Data System (CIDS) panel. This item was not discovered until en-route to ZZZ1. Aircraft was not serviced with potable water in ZZZ, so it operated at least one leg in violation of the MEL. The ZZZ Maintenance Technician stated the aircraft appeared to be not configured correctly for the 'No' potable water operation as all the valves had been left open. If the aircraft had a frozen water system as originally expected in the first write-up, how can one drain a water system that is frozen? Rhetorical question but that was the procedure listed under the MEL.	Maintenance	The incident was caused by a maintenance error in deferring the water system and not properly configuring the aircraft for 'No' potable water operation.

Similarly, for ACN 874307, ChatGPT identified ATC as the responsible entity. This determination was based on the narrative where the ATC's misinterpretation of the aircraft's status led to instructions that put the aircraft at risk of violating the Minimum Vectoring

Altitude (MVA). Importantly, both the safety analyst and ChatGPT identified *Situational Awareness* as the underlying human factors issue. The situation was complicated by adverse weather and irregular aircraft movements, which could have disrupted ATC's situational awareness, leading to the mistaken instruction. This particular alignment between the safety analyst's and ChatGPT's attributions underlines the language model's capacity to accurately identify complex human factors issues, mirroring the insight of a safety analyst.

As mentioned, Table 10 contains a column detailing the rationale used by ChatGPT to arrive at the responsibility attribution. Soliciting a rationale from the model for its identification of a particular human factors issue not only enhances transparency in the decision-making process but also fosters confidence in the model's outcomes. It facilitates an understanding of the model's reasoning, which can prove instrumental in pinpointing inaccuracies, thereby enabling model refinement and boosting the precision of incident responsibility assignment. This is especially important in regulatory settings, where providing a discernible trail of how conclusions were reached supports auditability, a key component of accountability. In addition, a visible rationale enables a critical examination of potential bias in the responsibility assignment process, promoting the creation of fairer and more equitable conclusions.

5. Conclusions

The primary objective of this study was to assess the applicability and suitability of generative language models, particularly ChatGPT, as tools for aviation safety analysis. ChatGPT was deployed to generate incident synopses based on provided narratives. These generated synopses were then compared to the ground-truth synopses found in the *Synopsis* column of the ASRS dataset. This comparative analysis involved using embeddings generated by LLMs (BERT, aeroBERT, and SBERT) and manually comparing the synopses. Upon manual evaluation, it was observed that synopses with higher cosine similarities tended to exhibit consistent similarities in terms of length. Conversely, synopses with lower cosine similarities showed more pronounced differences in their respective lengths.

Subsequent to this, the human factors issues linked to an incident, as determined by safety analysts, were compared to those identified by ChatGPT based on the incident narrative. In general, when ChatGPT identified a human factors issue, safety analysts agreed with this categorization around 61% of the time. ChatGPT demonstrated a more cautious approach in assigning human factors issues compared to the safety analysts. This may be ascribed to its limitation in not being able to infer causes that extend beyond the explicit content described within the narrative, given that no other columns were provided as inputs to the model.

Lastly, ChatGPT was employed to determine the entity to which the incident could be attributed. As there was no dedicated column serving as the ground truth for this specific task, a manual inspection was undertaken on a limited dataset. ChatGPT attributed 5877, 1971, 805, and 738 incidents to Flight Crew, ATC, Ground Personnel, and Maintenance, respectively. The rationale and underlying logic provided by ChatGPT for its attributions were well-founded. Nonetheless, due to the sheer volume of incidents used in this study, a manual examination of approximately 1% of the incident reports was performed. Validation of the results of generative models remains a challenge since the manual examination of large amounts of data is humanly impossible. Using embeddings from various LLMs to compute the cosine similarities, as conducted as part of this effort, represents a first valid step toward validating the outputs of ChatGPT or other similar generative language models.

The aforementioned results lead to the inference that the application of generative language models for aviation safety purposes presents considerable potential. However, it is suggested that these models be utilized in the capacity of "co-pilots" or assistants to aviation safety analysts, as opposed to being solely used in an automated way for safety analysis purposes. Implementing ChatGPT-like models as assistants to aid safety analysts in analyzing incident reports could significantly transform the analysts' workflow. The AI assistant could automate routine tasks such as data extraction, report summarization, and trend identification, freeing up the analyst's time for more critical tasks. With its ability to swiftly process and analyze large volumes of incident data, the AI assistant could potentially uncover patterns and anomalies that might be challenging for a human analyst to detect. Moreover, it could offer faster information retrieval by accessing relevant databases and safety records, providing timely and accurate information to support decision making. Over time, the AI assistant could learn and improve from its interactions with the analyst, becoming increasingly effective. It could also provide decision support by suggesting insights and highlighting areas of concern, although the ultimate decision-making authority would remain with the human analyst. The system's design could encompass a user-friendly natural language interface, integration with incident databases, pre-training, fine-tuning on domain-specific data, and a human-in-the-loop approach to decision making. Ethical considerations would be paramount, ensuring the system avoids bias, respects privacy, and prevents over-reliance on automation. With regular updates and a feedback mechanism, the AI assistant's performance would be expected to continuously improve, supporting safety analysts and fostering a collaborative approach to incident analysis in the aviation domain. Such AI assistants have already demonstrated successful implementation in various industries, including customer support, healthcare, retail and e-commerce, education, and the legal industry, among others.

Future work in this area should primarily focus on broadening and validating the application of generative language models for aviation safety analysis. More specifically, fine-tuning models such as ChatGPT on domain-specific data could enhance their understanding of the field's nuances, improving the generation of incident synopses, identification of human factors, and assessment of responsibility. The assessment of responsibility, as detailed on the ASRS landing page, is conducted to aid in enhancing the training of a specific group (such as pilots, maintenance personnel, and so on), and is in no way intended to impose punitive measures. In addition, the significant positive correlations between the synopsis length and the cosine similarity suggest the need for future experiments to be conducted to isolate and account for this bias.

Broadening the scope of the assessment of responsibility to encompass additional entities, such as passengers, weather conditions, or technical failures, and incorporating more detailed sub-categories could enhance the precision and effectiveness of these models. To solidify the ground truths for the assessment of responsibility, future research should scale manual inspection or employ other methods on a larger dataset. Following the suggestion to utilize these models as "co-pilots", the development of human–AI teaming approaches is another promising avenue. By designing interactive systems where safety analysts can guide and refine model outputs, or systems providing explanatory insights for aiding human decision making, both efficiency and accuracy could be enhanced. Finally, assessing the generalizability of these models across other aviation datasets, such as ATC voice communications, and other safety-critical sectors, such as space travel, maritime, or nuclear industries, would further solidify their wider applicability and suitability.

Author Contributions: Conceptualization, A.T.R., A.P.B., O.J.P.F. and D.N.M.; methodology, A.T.R., A.P.B., R.T.W. and V.M.N.; Software, A.T.R., R.T.W. and V.M.N.; validation, A.T.R. and R.T.W.; formal analysis, A.T.R. and A.P.B.; investigation, A.T.R., A.P.B. and R.T.W.; data curation, A.T.R. and A.P.B.; writing—original draft preparation, A.T.R. and R.T.W.; writing—review and editing, A.T.R., A.P.B., O.J.P.F., R.T.W., V.M.N. and D.N.M.; visualization, A.T.R. and R.T.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset used for this work can be found on the Hugging Face platform, URL: https://huggingface.co/datasets/archanatikayatray/ASRS-ChatGPT (accessed on 4 July 2023).

Acknowledgments: The authors would like to thank Karl R. Vliet for his subject matter expertise in Air Traffic Control (ATC) procedures, aviation incident/accident investigation, and training and operation of the ATC facility and personnel, which were crucial for this work.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ATC	Air Traffic Control
ASRS	Aviation Safety Reporting System
BERT	Bidirectional Encoder Representations from Transformers
CSV	comma-separated values
FAA	Federal Aviation Administration
GPT	Generative Pre-trained Transformer
JSON	JavaScript Object Notation
LaMDA	Language Models for Dialog Applications
LLaMA	Large Language Model Meta AI
LLM	Large Language Model
MLM	Masked Language Modeling
MVA	Minimum Vectoring Altitude
NAS	National Airspace System
NASA	National Aeronautics and Space Administration
NLP	Natural Language Processing
NSP	Next-Sentence Prediction
PaLM	Pathways Language Model
PPO	Proximal Policy Optimization
RL	reinforcement learning
RLHF	reinforcement learning from human feedback
RM	Reward Model
SFT	Supervised Fine-Tuning
UAS	Unmanned Aerial Systems

Appendix A

The prompt used for this work is presented below.

Listing A1. Prompt used for this work.

```

import openai
openai.api_key  = "YOUR-API-KEY"

def get_completion(prompt, model="gpt-3.5-turbo"):
    messages = [{"role": "user", "content": prompt}]
    response = openai.ChatCompletion.create(
        model=model,
        messages=messages,
        temperature=0, # this is the degree of randomness of the model's output
    )
    return response.choices[0].message["content"]

prompt = f"""
You are an aviation safety analyst who analyzes aviation incident reports.

Can you write a synopsis of the narrative in 1-2 sentences? Make sure to include the important details such as the  name of the system, and other relevant abbreviations, as necessary.


What are the main human factor issues that led to the incident based on the narrative? Choose single or multiple causes (as necessary) from the following options:
Communication breakdown,
Confusion,
Distraction,
```

```
Fatigue,
Human-Machine Interface,
Physiological-Other,
Situational Awareness,
Time Pressure,
Training/Qualification,
Troubleshooting,
Workload,
Other / Unknown.
Also, provide the rationale about how did you decide on the human factor
    issues that led to the incident in 1-2 sentences.


Based on the narrative, the incident can be attributed to which of these
    entities:
ATC (air traffic control),
Dispatch,
Flight crew,
Ground Personnel,
Maintenance,
Aircraft Manufacturer,
Other.
Provide the rationale behind the attribution.


The output should be in a JSON format with the keys, "Synopsis", "Human
    Factor issue", "Rationale - Human Factor issue", "Incident attribution",
"Rationale - Incident attribution".


Narrative: ```{narrative}```
"""

response = get_completion(prompt)
print(response)
```

References

1. ASRS Program Briefing PDF. Available online: https://asrs.arc.nasa.gov/docs/ASRS_ProgramBriefing.pdf (accessed on 16 May 2023).
2. ASRS Program Briefing. Available online: https://asrs.arc.nasa.gov/overview/summary.html (accessed on 16 May 2023).
3. Boesser, C.T. Comparing Human and Machine Learning Classification of Human Factors in Incident Reports from Aviation. 2020. Available online: https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1330&context=etd2020 (accessed on 16 May 2023).
4. Andrade, S.R.; Walsh, H.S. SafeAeroBERT: Towards a Safety-Informed Aerospace-Specific Language Model. In *AIAA AVIATION 2023 Forum*; American Institute of Aeronautics and Astronautics (AIAA): San Diego, CA, USA, 2023. [CrossRef]
5. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
6. Tikayat Ray, A.; Bhat, A.P.; White, R.T.; Nguyen, V.M.; Pinon Fischer, O.J.; Mavris, D.N. ASRS-ChatGPT Dataset. Available online: https://huggingface.co/datasets/archanatikayatray/ASRS-ChatGPT (accessed on 16 May 2023).
7. Electronic Report Submission (ERS). Available online: https://asrs.arc.nasa.gov/report/electronic.html (accessed on 16 May 2023).
8. General Form. Available online: https://akama.arc.nasa.gov/asrs_ers/general.html (accessed on 16 May 2023).
9. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019 ; pp. 4171–4186. [CrossRef]
10. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf (accessed on 16 May 2023).
11. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
12. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
13. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
14. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.

15. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288.
16. Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. Lamda: Language models for dialog applications. *arXiv* **2022**, arXiv:2201.08239.
17. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. Palm: Scaling language modeling with pathways. *arXiv* **2022**, arXiv:2204.02311.
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
19. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef]
20. Tikayat Ray, A.; Pinon Fischer, O.J.; Mavris, D.N.; White, R.T.; Cole, B.F. aeroBERT-NER: Named-Entity Recognition for Aerospace Requirements Engineering using BERT. In *AIAA SCITECH 2023 Forum*; American Institute of Aeronautics and Astronautics (AIAA): National Harbor, MD, USA, 2023. [CrossRef]
21. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2021**, arXiv:2108.07258.
22. Tikayat Ray, A.; Cole, B.F.; Pinon Fischer, O.J.; White, R.T.; Mavris, D.N. aeroBERT-Classifier: Classification of Aerospace Requirements Using BERT. *Aerospace* **2023**, *10*, 279. [CrossRef]
23. Tikayat Ray, A.; Cole, B.F.; Pinon Fischer, O.J.; Bhat, A.P.; White, R.T.; Mavris, D.N. Agile Methodology for the Standardization of Engineering Requirements Using Large Language Models. *Systems* **2023**, *11*, 352. [CrossRef]
24. Tikayat Ray, A. Standardization of Engineering Requirements Using Large Language Models. Ph.D. Thesis, Georgia Institute of Technology, Atlanta, GA, USA, 2023. [CrossRef]
25. Weaver, W. Translation. In *Machine Translation of Languages*; Locke, W.N., Boothe, A.D., Eds.; MIT Press: Cambridge, MA, USA, 1952 ; pp. 15–23. Available online: https://aclanthology.org/1952.earlymt-1.1.pdf (accessed on 16 May 2023).
26. Brown, P.F.; Cocke, J.; Della Pietra, S.A.; Della Pietra, V.J.; Jelinek, F.; Lafferty, J.; Mercer, R.L.; Roossin, P.S. A statistical approach to machine translation. *Comput. Linguist.* **1990**, *16*, 79–85.
27. Bengio, Y.; Ducharme, R.; Vincent, P. A Neural Probabilistic Language Model. In *Advances in Neural Information Processing Systems*; Leen, T., Dietterich, T., Tresp, V., Eds.; MIT Press: Cambridge, MA, USA, 2000; Volume 13.
28. Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; Smith, N.A. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. *arXiv* **2020**, arXiv:2009.11462.
29. Ziegler, D.M.; Stiennon, N.; Wu, J.; Brown, T.B.; Radford, A.; Amodei, D.; Christiano, P.; Irving, G. Fine-tuning language models from human preferences. *arXiv* **2019**, arXiv:1909.08593.
30. Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; Christiano, P.F. Learning to summarize with human feedback. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3008–3021.
31. Graeber, C. The role of human factors in improving aviation safety. *Aero Boeing* **1999**, *8*.
32. Santos, L.; Melicio, R. Stress, Pressure and Fatigue on Aircraft Maintenance Personal. *Int. Rev. Aerosp. Eng.* **2019**, *12*, 35–45. [CrossRef]
33. Saleh, J.H.; Tikayat Ray, A.; Zhang, K.S.; Churchwell, J.S. Maintenance and inspection as risk factors in helicopter accidents: Analysis and recommendations. *PLoS ONE* **2019**, *14*, e0211424. Available online: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0211424 (accessed on 16 May 2023). [CrossRef] [PubMed]
34. Dumitru, I.M.; Boşcoianu, M. Human factors contribution to aviation safety. *Int. Sci. Comm.* **2015**, *49*. Available online: https://www.afahc.ro/ro/afases/2015/afases_2015/air_force/dumitru_%20boscoianu.pdf (accessed on 16 May 2023).
35. Hobbs, A. Human factors: The last frontier of aviation safety? *Int. J. Aviat. Psychol.* **2004**, *14*, 331–345. [CrossRef]
36. Salas, E.; Maurino, D.; Curtis, M. Human factors in aviation: An overview. *Hum. Factors Aviat.* **2010**, 3–19.
37. Cardosi, K.; Lennertz, T. Human Factors Considerations for the Integration of Unmanned Aerial Vehicles in the National Airspace System: An Analysis of Reports Submitted to the Aviation Safety Reporting System (ASRS). 2017. Available online: https://rosap.ntl.bts.gov/view/dot/12500 (accessed on 16 May 2023).
38. Madeira, T.; Melício, R.; Valério, D.; Santos, L. Machine learning and natural language processing for prediction of human factors in aviation incident reports. *Aerospace* **2021**, *8*, 47. [CrossRef]
39. Aurino, D.E.M. Human factors and aviation safety: What the industry has, what the industry needs. *Ergonomics* **2000**, *43*, 952–959. [CrossRef] [PubMed]
40. Hobbs, A. An overview of human factors in aviation maintenance. *ATSB Safty Rep. Aviat. Res. Anal. Rep. AR* **2008**, *55*, 2008.
41. Kierszbaum, S.; Klein, T.; Lapasset, L. ASRS-CMFS vs. RoBERTa: Comparing Two Pre-Trained Language Models to Predict Anomalies in Aviation Occurrence Reports with a Low Volume of In-Domain Data Available. *Aerospace* **2022**, *9*, 591. [CrossRef]
42. Yang, C.; Huang, C. Natural Language Processing (NLP) in Aviation Safety: Systematic Review of Research and Outlook into the Future. *Aerospace* **2023**, *10*, 600. [CrossRef]
43. Tanguy, L.; Tulechki, N.; Urieli, A.; Hermann, E.; Raynal, C. Natural language processing for aviation safety reports: From classification to interactive analysis. *Comput. Ind.* **2016**, *78*, 80–95. [CrossRef]
44. OpenAI. ChatGPT API; gpt-3.5-turbo. 2023. Available online: https://openai.com/blog/introducing-chatgpt-and-whisper-apis (accessed on 4 June 2023).

45. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 3982–3992. [CrossRef]
46. Heydarian, M.; Doyle, T.E.; Samavi, R. MLCM: Multi-Label Confusion Matrix. *IEEE Access* **2022**, *10*, 19083–19095. [CrossRef]