Article

Not peer-reviewed version

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 * , Anirudh Prabhakara Bhat , Ryan T White , Van Minh Nguyen , Olivia J Pinon Fischer * , Dimitri N Mavris

Posted Date: 4 July 2023

doi: 10.20944/preprints202307.0192.v1

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

 doi:10.20944/preprints202307.0192.v1

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, Georgia, 30332, USA
2 AI Fusion Technologies, Toronto, Ontario, M5V 3Z5, Canada
3 NEural TransmissionS Lab, Department of Mathematics and Systems Engineering, Florida Institute of Technology, Melbourne, Florida, 32901, USA
* Correspondence: archanatikayatray@gmail.com (A.T.R.); olivia.pinon@asdl.gatech.edu (OJ.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 Language Learning Models (LLMs), with aeroBERT demonstrating the highest similarity due to its aerospace-specific fine-tuning. A positive correlation was observed between synopsis length and their cosine similarity. In a subsequent phase, human factor issues involved in incidents as identified by ChatGPT were compared to human factor issues identified by safety analysts. A concurrence rate of 61% was found, with ChatGPT demonstrating a cautious approach towards attributing human factor issues. Finally, the model was used to attribute incidents to relevant parties. As no dedicated ground truth column existed for this task, a manual evaluation was conducted. ChatGPT attributed the majority of incidents to the Flight Crew, ATC, Ground Personnel, and Maintenance. This study opens new avenues for leveraging AI in aviation safety analysis.

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

---

1. Introduction

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

Prior studies, including those mentioned in [2], have utilized the ASRS dataset in conjunction with Large Language Models (LLMs), such as BERT. However, the conclusions drawn from these studies may be limited for a variety of reasons. The first is the imposed maximum narrative length

 doi:10.20944/preprints202307.0192.v1

of 256 or 512 WordPiece tokens when using BERT, which may potentially lead to information loss. For comparison, this specified length is below the 25^{th} percentile of incident narrative lengths, 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 to 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) [3] 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 consistency and reproducibility in their analyses. In particular, this paper evaluates the efficacy of generative language models in automatically performing the following tasks from free-form incident narratives:

1. Generate succinct synopses of the incidents.
2. Compare the faithfulness of the generated synopses to human-written synopses.
3. Discern the human factor(s) contributing to an incident.
4. Pinpoint the party responsible for the incident.
5. Provide explanatory logic/rationale for the generative language model's decisions.

The assembled dataset, which includes the ground truth, generated outputs, and accompanying rationale, can be found on the HuggingFace platform [4]. 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 truth. Section 4 discusses the findings of this work and presents examples of incident narratives, synopses, human factor errors, and responsible parties. 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2023 doi:10.20944/preprints202307.0192.v1

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. 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 ASRS via postal mail [5].

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

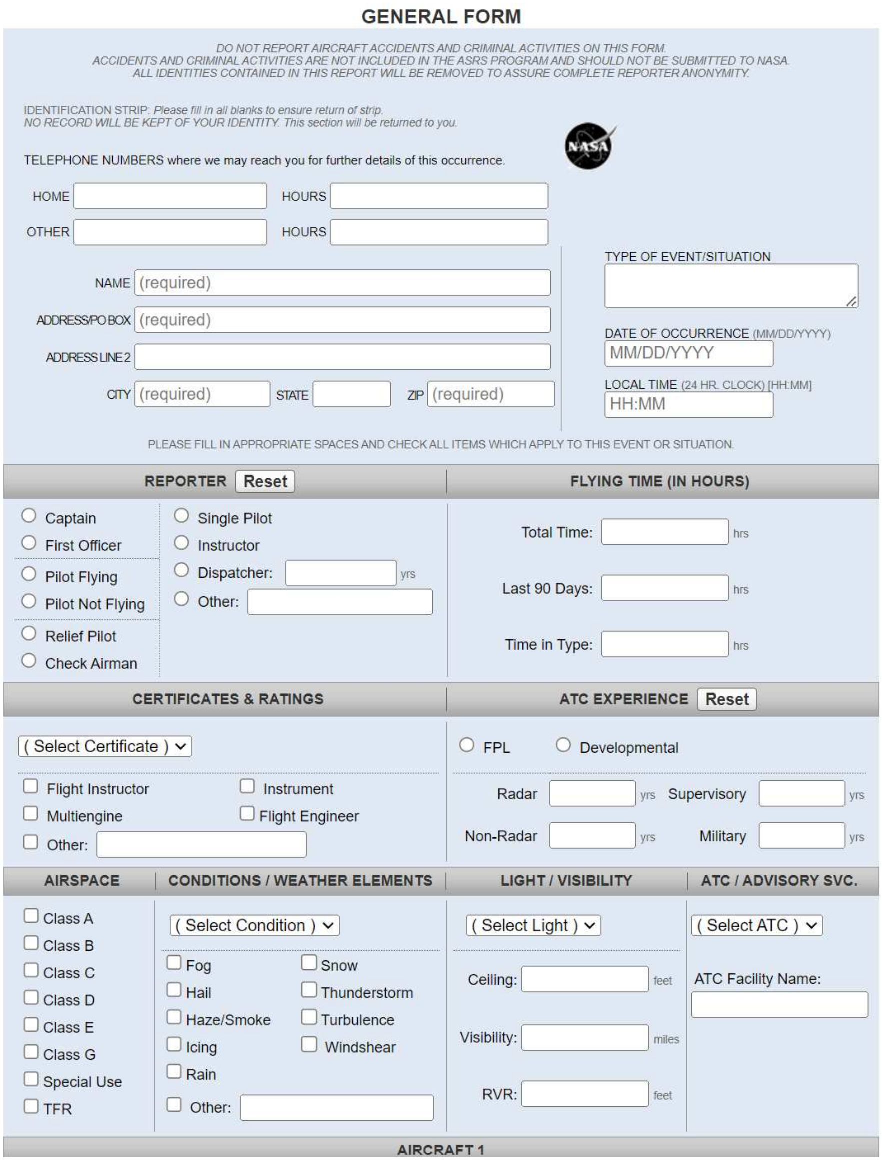
GENERAL FORM
DO NOT REPORT AIRCRAFT ACCIDENTS AND CRIMINAL ACTIVITIES ON THIS FORM.
ACCIDENTS AND CRIMINAL ACTIVITIES ARE NOT INCLUDED IN THE ASRS PROGRAM AND SHOULD NOT BE SUBMITTED TO NASA.
ALL IDENTITIES CONTAINED IN THIS REPORT WILL BE REMOVED TO ASSURE COMPLETE REPORTER ANONYMITY.
IDENTIFICATION 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 form is 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. [6] A part of the General form is shown in this figure.

Figure 2 illustrates the pipeline for processing incident reports in the ASRS. The process begins with 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 analysts then screen the report to determine its initial categorization and triage for processing [1]. 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 [1].

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 [1]. Table 2 shows some of the columns included in the ASRS dataset.

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 [1].

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 with reducing the processing time, improving labeling accuracy, and ultimately enhancing 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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2023 doi:10.20944/preprints202307.0192.v1

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-hour 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.
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 are some examples. 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, Airspace Structure are some examples. 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, 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 11000. 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 10000. 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.

Column Name Description

2.2.1. Large Language Models (LLMs) as Foundation Models

LLMs, such as Bidirectional Encoder Representations from Transformers (BERT) [7] and the Generative Pre-trained Transformer (GPT) family [8–11], LLaMA [12], LaMDA [13], PaLM [14] are advanced natural language processing systems that have shown remarkable capabilities in understanding and generating human-like text. These models are built upon Transformer neural networks with attention mechanisms [15]. 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 a large corpus 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* [16], 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.

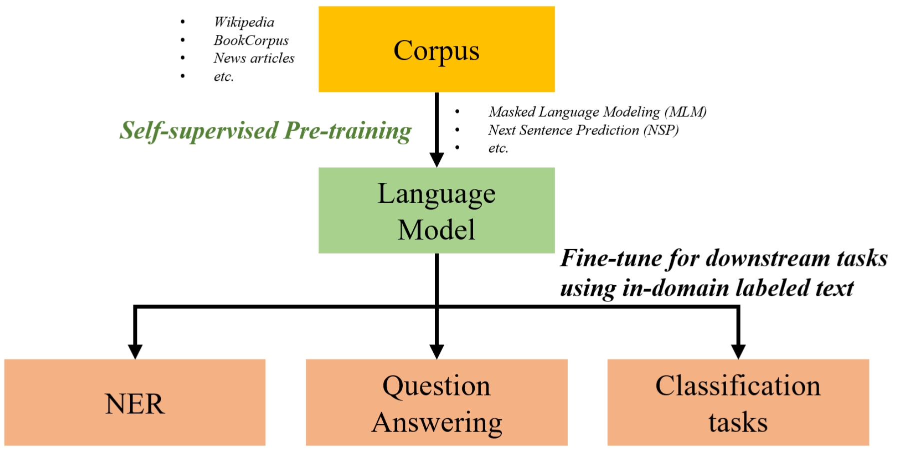

Figure 3. This figure demonstrates the training process of large language models (LLMs) in two stages: pre-training and fine-tuning. In the pre-training stage, the LLM learns from a large unlabeled corpus to capture language patterns and semantics. In the fine-tuning stage, the LLM is further trained on labeled corpora specific to downstream tasks, adapting its knowledge to improve performance in task-specific domains [17].

In summation, LLMs, including BERT and GPT are often termed as *foundation models*. They provide the basis for an extensive range of specialized models and applications [18].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2023 doi:10.20944/preprints202307.0192.v1

With specific regard to the aerospace field, there exist fine-tuned models, namely aeroBERT-NER [17] and aeroBERT-Classifier [19], developed by fine-tuning variants of BERT on annotated aerospace corpora [17,19–21]. These models were designed to recognize aerospace-specific named entities and to categorize aerospace requirements into different types, respectively [21].

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

2.2.2. 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 [22,23], which aim to generate text sequences by choosing word (token) sequences that maximize next-token probabilities conditioned on prior text. Neural language models [24] use neural networks for estimating 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 corpus for generative pre-training can be unlabeled as in the self-supervised training of BERT.

The GPT [8] is a neural language model that employs a Transformer-based *decoder* [15] as its neural architecture. This should be contrasted with 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. Rather, GPT performs task-specific input transformations indicating which task should be done, all of which can be fed to the same Transformer.

In GPT-1 [8], 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 [9] proposed conditioning its output probabilities on both the input and the desired task. Training a model for this multi-task learning by supervised means is infeasible, as it would require thousands of (dataset, objective) pairs for training. Therefore, GPT-2 shifts focus and demonstrates strong zero-shot capabilities on some tasks without supervised fine-tuning can be achieved with a larger model (1.5B parameters). GPT-3 [10] scales GPT-2 up to 175B parameters, which greatly improves performance in 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 to achieve 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 [25], 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) [3] aims to correct this misalignment. InstructGPT is the underlying LLM for ChatGPT by OpenAI [3].

Since the mistakes GPT-3 makes are not easy to evaluate through simple NLP metrics, InstructGPT employs reinforcement learning with human feedback (RLHF) [26,27] 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.

 doi:10.20944/preprints202307.0192.v1

2. **Train 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. **Optimize 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 perform a diverse set of tasks effectively 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.

Generative language models can be instrumental in advancing aviation safety analysis by streamlining various processes and tasks. These models can significantly expedite the review and assessment of incident reports, by automatically identifying and categorizing crucial information, thus aiding in faster decision-making and implementation of preventive measures. In terms of forecasting, generative language models can detect patterns and trends from historical data to anticipate potential safety issues, thereby contributing to proactive safety management. The ability of these models to generate concise summaries of complex reports further enhances their utility by ensuring that key insights are readily available for stakeholders. Additionally, these models can provide nuanced insights into the contextual factors influencing safety incidents by analyzing communication data and safety reports. Therefore, generative language models can significantly augment safety analysis by introducing efficiencies and providing multifaceted insights. Despite the promises, it is imperative to assert that generative language models should not supersede the role of human safety analysts. Rather, they ought to be incorporated within a system that maintains human involvement at its core. In addition, most of these applications of generative language models still remain unexplored.

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.

2.2.3. NLP in aviation safety analysis

There has been a decrease in the occurrence of incidents resulting from technical failures, however, the incidents arising from human factor issues have emerged as the predominant underlying cause of the majority of incidents [28,29]. Several studies, such as [30–35], have looked into human factor issues in aviation. One of the most complex and difficult tasks when classifying aviation safety incidents is to sub-classify incidents stemming from human factor complications, which is a primary area of interest in this research. The investigation conducted in [36] gathered labels from six individual annotators, each working on a subset of 400 incident reports, culminating in a collective 2,400 individual annotations. The outcomes indicated that there was a high level of disagreement among the human annotators. This highlights the potential for language models 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 [2,37].

In their study [2], Andrade et al. introduce SafeAeroBERT, a 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 superseded BERT and SciBERT only in 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. [37] propose 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, 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. 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 when compared to other causes. The high number of aviation incidents linked to human factors can be attributed to several factors such as the aviation environment being complex and stressful, which can lead to errors. People have physical and mental limits, and issues like fatigue, stress, and misunderstanding can contribute to mistakes. Differences in crew training and skills can also cause disparities. Sometimes, cockpit design can lead to confusion, and some airline training cultures may not focus enough on safety.

3.2. Prompt Engineering for ASRS analysis

This work leverages GPT-3.5 (via the OpenAI's ChatGPT API) [38] to analyze incident narratives, identify human factor issues leading to incidents (Table 3), pinpoint responsible parties, 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.

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 guide the model's responses effectively, improving its performance and output applicability. The temperature parameter of ChatGPT was set to zero for this work. When the value is set to 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2023 doi:10.20944/preprints202307.0192.v1

Table 3. List of *human factor* issues and their definitions are presented here. One or multiple of these factors can result in an incident or accident.

Human Factor 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 hinders 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 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, in order 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2023 doi:10.20944/preprints202307.0192.v1

Subsequently, the model was tasked with identifying various human factor issues responsible for the incident based on the provided incident narratives. While the model demonstrated the ability to identify these human factor 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 truth, which encompassed twelve overarching human factor 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 factor 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 party 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 Factor issue*, *Rationale - Human Factor issue*, *Incident attribution*, and *Rationale - Incident attribution*. This 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 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.

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 performs various NLP task, 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 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 (near 1), and dissimilar when they have low cosine similarity (near -1). Hence, the cosine similarity of all pairs of human-written and ChatGPT-generated incident synopses were measured using several BERT-based LMs, including BERT [7], aeroBERT [17,19], and Sentence-BERT (SBERT) [39]. 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.

Subsequently, a comparison was made between the human factor issues identified by ChatGPT and those identified by human safety analysts. The frequency at which ChatGPT associated each human factor issue with incident narratives was taken into account. To visualize these comparisons, a histogram was constructed. Furthermore, a normalized multi-label confusion matrix was created to illustrate the level of concordance between ChatGPT and safety analysts in attributing human factor issues to incident narratives. Each row in the matrix represents a human factor issue assigned to cases by the human annotators. 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 aligns with the human annotators, all values except for those on the diagonal would be zero.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2023 doi:10.20944/preprints202307.0192.v1

Performance metrics such as precision, recall, and F1 score were used to assess the agreement between ChatGPT and safety analysts.

Finally, an analysis was conducted regarding the attribution of fault by ChatGPT. The focus was placed on the top five parties involved in incidents, and their attribution was discussed qualitatively, 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 9,984 unique records. The prompt 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, outlined in Table 4.

Table 4. Features 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 Factor Issue	A list of human factor 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, Workload (mirroring issues used in ASRS) along with additional issues ChatGPT was free to suggest.
Human Factor Issue (Rationale)	A 1-2 sentence description of the rationale ChatGPT used to decide which human factor issues were relevant.
Incident Attribution	A party or parties 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 synopsis: (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) find the synopses mostly quite similar. aeroBERT with its fine-tuning to aerospace-specific language evaluates the sequences as most similar, while the dedicated but general-purpose sequence-comparison model SBERT finds the ChatGPT synopses as similar but less so.

 doi:10.20944/preprints202307.0192.v1

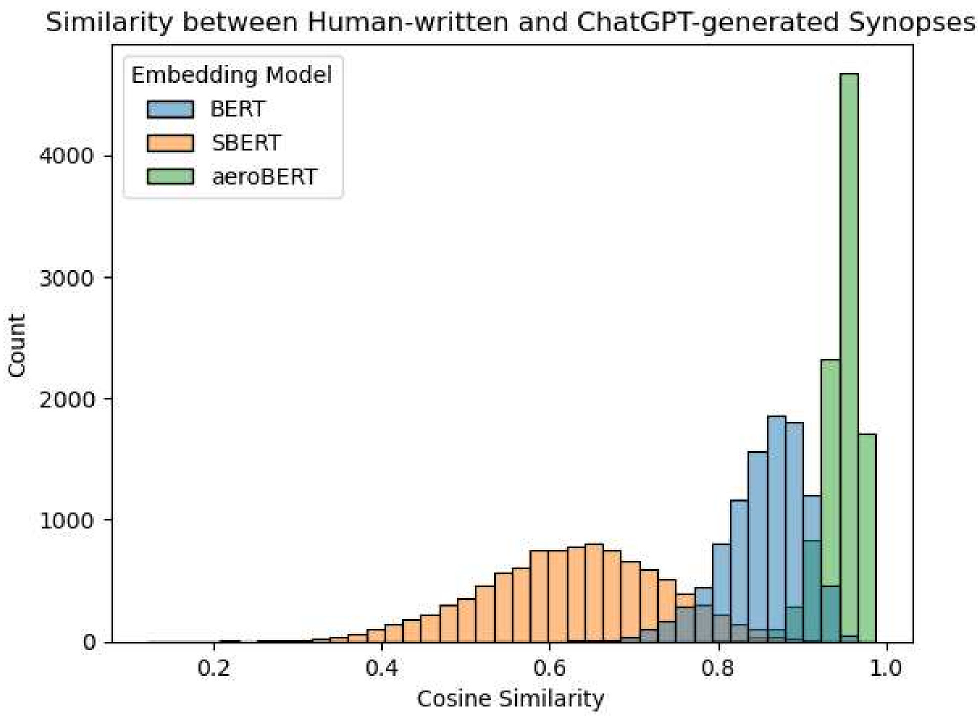

Figure 4. Histogram of cosine similarities as computed by BERT, SBERT, and aeroBERT

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

Table 5 displays the three synopses that have the highest cosine similarity between the embeddings generated by each LM of generated synopses and the synopses written by safety analysts. It is important to highlight that the lengths of the generated and ground truth synopses are very similar in this case. This is particularly relevant considering the prompt directed the generative model to produce a synopsis based on the incident narrative while maintaining a length constraint of 1-2 sentences, as depicted in Table 4.

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 safety analysts. A noteworthy observation from this data is the significant difference in length between the generated synopses and the ground truth for all three LMs. The length of the generated synopses fluctuates 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 concerning the synopses that obtained the lowest ranking in terms of cosine similarity. For example, the generated synopsis corresponding to "*A light twin and a single engine have a NMAC at Isla Grande*" (ground truth) was given the lowest ranking by both BERT and SBERT models. However, the cosine similarity when employing BERT embeddings was higher relative to that of 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 aeroBERT is less likely to attribute similarities based on the confounding variable of synopsis length.

 doi:10.20944/preprints202307.0192.v1

Table 5. Table displaying the three top matched synopses generated by the safety analyst and ChatGPT as determined by cosine similarity (CS) scores calculated using different LM 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. Table displaying the three worst matched synopses generated by the safety analyst 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

 doi:10.20944/preprints202307.0192.v1

Table 6. *Cont.*

ACN	Synopsis (Safety Analyst)	Synopsis (ChatGPT)	CS
		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
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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 July 2023 doi:10.20944/preprints202307.0192.v1

4.2. Performance with Human Factors Related Issues

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

First, the frequency with which ChatGPT attributes each human factors issue to incident narratives was considered. ChatGPT assigns almost all human factors issues to narratives less frequently than human annotators. The only exception is *Training/Qualification*, which ChatGPT assigns more frequently. Notably, ChatGPT almost never assigns *Confusion*, *Human-Machine Interface*, *Other/Unknown*, and *Troubleshooting*. It should be emphasized that there were variations in the human factor issues as identified by safety analysts due to subtle differences in categories like *Confusion* and *Fatigue*.

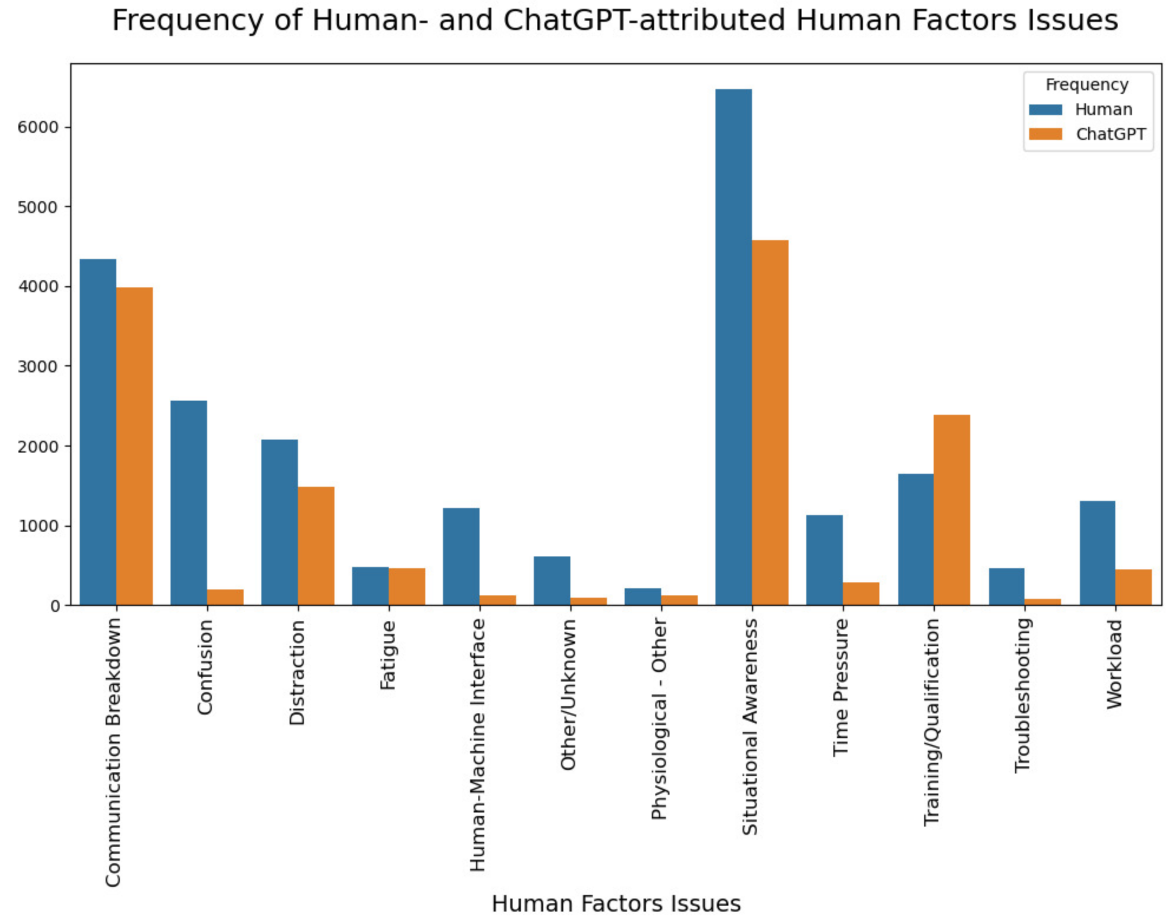

Figure 5. Frequency of Human-attributed and ChatGPT-attributed Human Factors Issues

Next, the normalized multi-label confusion matrix [40] in Figure 6 shows the degree of agreement between ChatGPT and the human annotators for human factors issues attributed to each record in the ASRS. Each row represents a human factor issue attributed to cases by the human annotators. The numbers in the row are the percentages of times ChatGPT attributed the same issue to the narrative. If ChatGPT agreed perfectly with the human annotators, all numbers except the numbers on the diagonal would be 0.

We note that ChatGPT agrees with human annotators more than 40% of the time on the following classes: *Communication Breakdown*, *Fatigue*, *Situational Awareness*, and *Training/Qualification*. It agrees substantially less for the other classes. The most common source of disagreement is when ChatGPT chooses not to attribute any human factors issues (see the rightmost column of the multilabel confusion matrix in Figure 6). In addition, ChatGPT attributes many more incidents than humans do to the factors textitCommunication Breakdown, *Distraction*, *Situational Awareness*, and *Training/Qualification*.

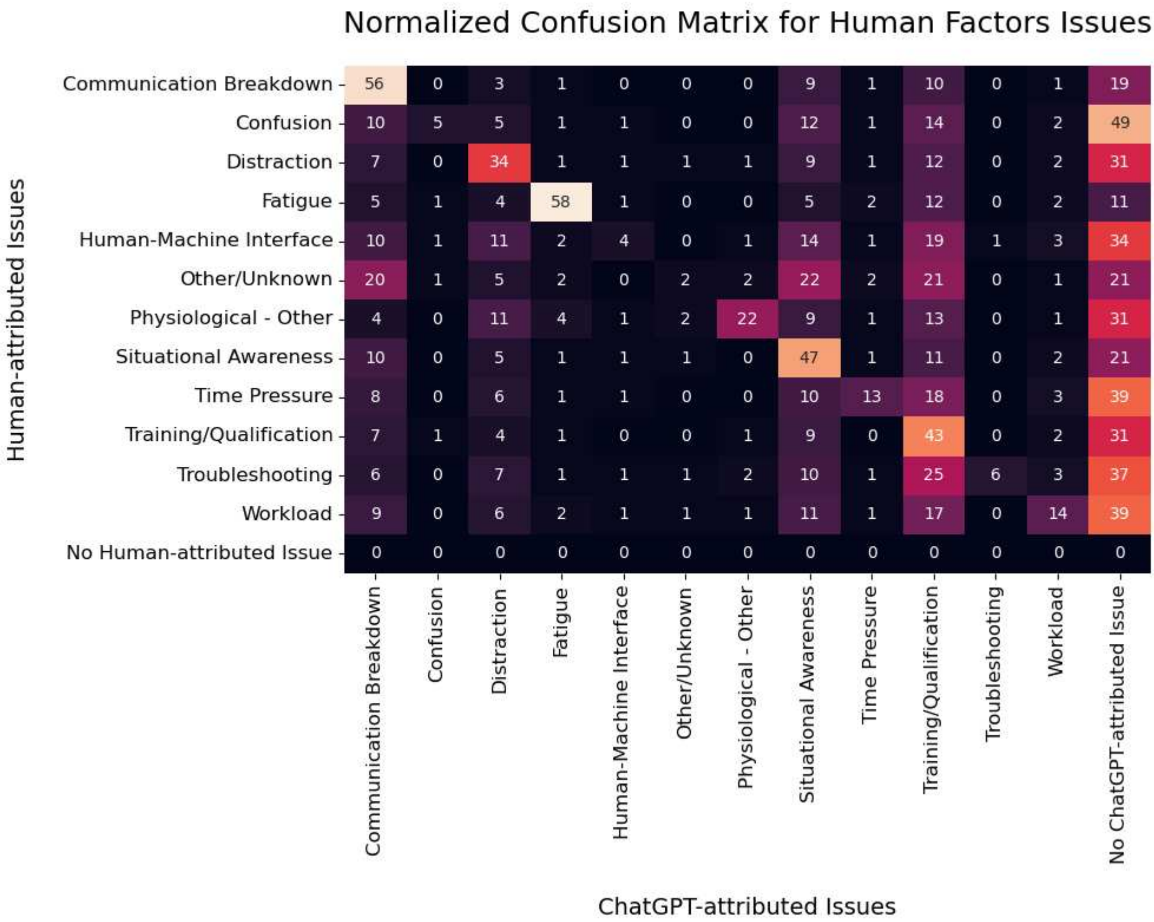

Figure 6. Normalized Multilabel Confusion Matrix for Human Factors Issues

To provide more of a comprehensive assessment, Table 7 provides precision, recall, and F1 score measuring ChatGPT agreement with humans for each issue. Generally speaking, when ChatGPT claims a human factor issue applies, human annotators agree 61% of the time (i.e., precision). The most common source of disagreement is that ChatGPT is generally more conservative in assigning human factors issues than humans and, hence, has lower recall values for most issues. However, it is crucial to note that there can be significant variations in annotations performed by safety analysts, as demonstrated in [36].

Table 7. 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	22498

4.3. Attribution of Fault

ChatGPT was further employed to identify the party the incident could be attributed to. It is important to emphasize that the intent is not to impose punitive measures on the implicated parties, but rather to leverage these specific incidents as instructive examples during training.

The outputs from ChatGPT did not rigidly follow the options provided in the prompt and proceeded in generating parties responsible based on the narrative. Moreover, in scenarios where responsibility could be apportioned to several parties, the model demonstrated the capability to do so. Table 8 provides the top five parties responsible for incidents that can be attributed to human factors.

In the context of this task, the term 'flight crew' encompassed all individuals aboard the flight, barring the passengers. A total of 5744 incidents were attributed to errors by the flight crew. Similarly, the model adeptly pinpointed incidents resulting from Air Traffic Control (ATC) errors with considerable specificity, including those from ATC (Approach Control), ATC (Ground Control), ATC (Kona Tower), ATC (Military Controller), ATC (TRACON Controller), ATC (Indy Center), and so forth. These were combined into a single ATC category by the authors.

Table 8. List of top five parties responsible for incidents where the primary contributing factor was identified as human factors.

Party Responsible for Incident	**Count**
Flight Crew	5744
ATC	1838
Ground Personnel	805
Maintenance	738
Flight Crew, ATC	133

To comprehend the reason behind ChatGPT associating a particular incident to a specific party, it was encouraged to produce the underlying rationale for each attribution. Table 9 provides some examples to facilitate a deeper understanding of the principles governing attribution rationale. Upon a detailed review of certain incident accounts, it was noted that the entity to which ChatGPT attributed the incident, and the logic underlying such attribution, seemed to indicate a high level of accuracy in the model's responses and their justifications. For example, in the first case presented in Table 9, the lead Flight Attendant, overly concerned with a passenger's non-compliance to 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. Consequently, the

 doi:10.20944/preprints202307.0192.v1

attribution of the incident to the flight crew, as made by ChatGPT, is appropriate. Furthermore, the reasoning behind such attribution is well-founded.

Asking the model for a rationale behind a certain decision provides a degree of transparency in the identification process, promoting trust in the conclusions drawn. Furthermore, the rationale can help in error identification and subsequent model optimization, thereby improving the precision of incident responsibility attribution. In regulatory contexts, the rationale supports auditability by providing a clear trail of how conclusions were reached, which is essential for accountability. Moreover, it allows for an assessment of any potential bias in assigning responsibility, leading to fairer and more balanced conclusions.

Table 9. Table displaying the three top matched synopses generated by the safety analyst and ChatGPT as determined by cosine similarity (CS) scores calculated using different LM embeddings.

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 1,500 feet. Aircraft X was at 2,000 on a vector for an ILS approach approximately 15 miles southeast of ADS. I observed the aircraft in proximity to a 2,200 foot MVA. I, for some reason, registered that the aircraft was VFR and instructed the aircraft to maintain VFR at or below 2,500 so that he could maintain obstacle clearance. The pilot complied and descended to 1,600 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 3,000. By that time the aircraft was inside the MVA at 1,600. 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.

5. Conclusions

The primary objective of this study was to evaluate the potential utility of generative language models based on the current state-of-the-art generative LM (ChatGPT) for aviation safety analysis. ChatGPT was deployed to generate incident synopses derived from the narrative. These generated synopses were then juxtaposed with the ground truth synopses found in the *Synopsis* column of the ASRS dataset. This comparison was performed utilizing embeddings generated by LLMs.During

 doi:10.20944/preprints202307.0192.v1

manual evaluation, it was observed that synopses exhibiting higher cosine similarity tended to exhibit consistent similarities in terms of length. In contrast, synopses with lower cosine similarity displayed more noticeable discrepancies in their respective lengths.

Subsequent to this, a comparison was conducted between the human factor issues linked to an incident by safety analyses to those identified by ChatGPT on the basis of the narrative. In general, there was a 61% concurrence between decisions made by ChatGPT and those made by safety analysts. ChatGPT demonstrated a more cautious approach in assigning human factor issues in comparison to human evaluators. 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 no other columns were provided as inputs to the model.

Lastly, ChatGPT was employed to determine the party the incident could be attributed to. As there was no dedicated column serving as 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 every individual incident record was not conducted.

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 models in such a manner would serve to alleviate the analysts' workload while simultaneously enhancing the efficiency of their work.

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 like 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 attributing fault. To reiterate, the identification of responsibility will aid in enhancing the training of a specific group (such as pilots, maintenance personnel, and so forth), as opposed to imposing punitive measures. The significant positive correlations between synopsis length and cosine similarity suggests future experiments to isolate and account for this bias.

Expanding the scope of incident attribution to include other parties such as passengers, weather conditions, or technical failures could further refine the application of these models. To solidify the ground truth for incident attribution, 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 human 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 safety-critical sectors such as space travel, maritime, or nuclear industries would further solidify their wider applicability.

Author Contributions: **Archana Tikayat Ray**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing—original draft preparation, Writing—review and editing **Anirudh Prabhakara Bhat**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing—review and editing **Ryan T. White**: Investigation, Methodology, Software, Validation, Visualization, Writing—original draft preparation, Writing—review and editing **Van Minh Nguyen**: Methodology, Software **Olivia J. Pinon Fischer**: Writing—review and editing **Dimitri N. Mavris**: Writing—review and editing

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

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

 doi:10.20944/preprints202307.0192.v1

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
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
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.

```

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)
```

Listing 1: Prompt used for this work

References

1. ASRS Program Briefing. https://asrs.arc.nasa.gov/overview/summary.html. (accessed: 05.16.2023).
2. Andrade, S.R.; Walsh, H.S., SafeAeroBERT: Towards a Safety-Informed Aerospace-Specific Language Model. In *AIAA AVIATION 2023 Forum*; [https://arc.aiaa.org/doi/pdf/10.2514/6.2023-3437]. doi:10.2514/6.2023-3437.
3. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; others. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* **2022**, *35*, 27730–27744.
4. Tikayat Ray, A.; Bhat, A.P.; White, R.T.; Nguyen, V.M.; Pinon Fischer, O.J.; Mavris, D.N. ASRS-ChatGPT Dataset. doi:10.57967/hf/0830.
5. Electronic Report Submission (ERS). https://asrs.arc.nasa.gov/report/electronic.html. (accessed: 05.16.2023).
6. General Form. https://akama.arc.nasa.gov/asrs_ers/general.html. (accessed: 05.16.2023).
7. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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); Association for Computational Linguistics: Minneapolis, Minnesota, 2019; pp. 4171–4186. doi:10.18653/v1/N19-1423.
8. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; others. Improving language understanding by generative pre-training **2018**.
9. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; others. Language models are unsupervised multitask learners. *OpenAI blog* **2019**, *1*, 9.
10. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; others. Language models are few-shot learners. *Advances in neural information processing systems* **2020**, *33*, 1877–1901.
11. OpenAI. GPT-4 Technical Report, 2023, [arXiv:cs.CL/2303.08774].

12. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; others. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* **2023**.
13. Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; others. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* **2022**.
14. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; others. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* **2022**.
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. Advances in Neural Information Processing Systems; Guyon, I.; Luxburg, U.V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., Eds. Curran Associates, Inc., 2017, Vol. 30.
16. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural computation* **2006**, *18*, 1527–1554.
17. 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*. doi:10.2514/6.2023-2583.
18. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N.S.; Chen, A.S.; Creel, K.A.; Davis, J.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L.E.; Goel, K.; Goodman, N.D.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D.E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.F.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P.W.; Krass, M.S.; Krishna, R.; Kuditipudi, R.; Kumar, A.; Ladhak, F.; Lee, M.; Lee, T.; Leskovec, J.; Levent, I.; Li, X.L.; Li, X.; Ma, T.; Malik, A.; Manning, C.D.; Mirchandani, S.P.; Mitchell, E.; Munyikwa, Z.; Nair, S.; Narayan, A.; Narayanan, D.; Newman, B.; Nie, A.; Niebles, J.C.; Nilforoshan, H.; Nyarko, J.F.; Ogut, G.; Orr, L.; Papadimitriou, I.; Park, J.S.; Piech, C.; Portelance, E.; Potts, C.; Raghunathan, A.; Reich, R.; Ren, H.; Rong, F.; Roohani, Y.H.; Ruiz, C.; Ryan, J.; R'e, C.; Sadigh, D.; Sagawa, S.; Santhanam, K.; Shih, A.; Srinivasan, K.P.; Tamkin, A.; Taori, R.; Thomas, A.W.; Tramèr, F.; Wang, R.E.; Wang, W.; Wu, B.; Wu, J.; Wu, Y.; Xie, S.M.; Yasunaga, M.; You, J.; Zaharia, M.A.; Zhang, M.; Zhang, T.; Zhang, X.; Zhang, Y.; Zheng, L.; Zhou, K.; Liang, P. On the Opportunities and Risks of Foundation Models. *ArXiv* **2021**.
19. 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*. doi:10.3390/aerospace10030279.
20. 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; 2023. doi:10.20944/preprints202305.1325.v1.
21. Tikayat Ray, A. Standardization of Engineering Requirements Using Large Language Models. PhD thesis, Georgia Institute of Technology, 2023. doi:10.13140/RG.2.2.17792.40961.
22. Weaver, W. Translation. In *Machine Translation of Languages*; Locke, W.N.; Boothe, A.D., Eds.; MIT Press: Cambridge, MA, 1949/1955; pp. 15–23. Reprinted from a memorandum written by Weaver in 1949.
23. 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. *Computational linguistics* **1990**, *16*, 79–85.
24. Bengio, Y.; Ducharme, R.; Vincent, P. A Neural Probabilistic Language Model. Advances in Neural Information Processing Systems; Leen, T.; Dietterich, T.; Tresp, V., Eds. MIT Press, 2000, Vol. 13.
25. Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; Smith, N.A. REALTOXICITYPROMPTS: Evaluating Neural Toxic Degeneration in Language Models.
26. 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 preprint arXiv:1909.08593* **2019**.
27. 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. *Advances in Neural Information Processing Systems* **2020**, *33*, 3008–3021.
28. Graeber, C. The role of human factors in improving aviation safety. *Aero Boeing* **1999**, *8*.

29. Santos, L.; Melicio, R. Stress, Pressure and Fatigue on Aircraft Maintenance Personal. *International Review of Aerospace Engineering* **2019**, *12*, 35–45. doi:10.15866/irease.v12i1.14860.
30. 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.
31. Dumitru, I.M.; Boşcoianu, M. Human factors contribution to aviation safety. *International Scientific Committee* **2015**, *49*.
32. Hobbs, A. Human factors: the last frontier of aviation safety? *The International Journal of Aviation Psychology* **2004**, *14*, 331–345.
33. Salas, E.; Maurino, D.; Curtis, M. Human factors in aviation: an overview. *Human factors in aviation* **2010**, pp. 3–19.
34. Cardosi, K.; Lennertz, T.; others. 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**.
35. 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.
36. Boesser, C.T. Comparing human and machine learning classification of human factors in incident reports from aviation **2020**.
37. 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*. doi:10.3390/aerospace9100591.
38. OpenAI. ChatGPT API. https://openai.com/blog/introducing-chatgpt-and-whisper-apis, 2023. gpt-3.5-turbo.
39. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Association for Computational Linguistics: Hong Kong, China, 2019; pp. 3982–3992. doi:10.18653/v1/D19-1410.
40. Heydarian, M.; Doyle, T.E.; Samavi, R. MLCM: Multi-Label Confusion Matrix. *IEEE Access* **2022**, *10*, 19083–19095. doi:10.1109/ACCESS.2022.3151048.