



Augmenting an Incident Dataset with ChatGPT

Jon Ricketts^{ab} 

^a Corresponding author email: j.ricketts@cranfield.ac.uk

^b Cranfield University, UK

Keywords:

natural language processing, incident reporting, semantic search, hazard identification

Peer-Reviewed

Gold Open Access

Zero APC Fees

[CC-BY-ND 4.0 License](#)

Received: 05-Mar-2022

Accepted: 05-Mar-2022

Online: 05-Mar-2022

Cite As:

Ricketts J, Augmenting an Incident Dataset with ChatGPT. Journal of System Safety. 2024;59(1):9-15. doi: 10.56094/jss.v59i1.273

ABSTRACT

The field of Natural Language Processing (NLP) is evolving at a rapid rate, impacting ways of working across multiple industries including that of System Safety. One area of NLP is the development of advanced language models, notably ChatGPT—which is essentially a powerful artificial intelligence chatbot powered by a large language model. This paper takes an incident report dataset and augments it with ChatGPT to improve searching capability and provide answers to safety related queries. It is shown that incident datasets can be further adapted for knowledge retrieval to support safety queries, however, a major limitation to deploying this method elsewhere are data protection policies. The underpinning vector database (used to retrieve relevant incident reports) demonstrated a useful semantic search ability for more accurate and meaningful searches of incident datasets. It is considered that if the outputs provide evidence or sources behind answers, and are used for advisory purposes then they can form useful tools for information and knowledge retrieval in System Safety.

ABBREVIATIONS

AI	Artificial Intelligence
ACN	ASRS Record Number
ASRS	Aviation Safety Reporting System
GPT	Generative Pre-trained Transformer
LLM	Large Language Model
NLP	Natural Language Processing

INTRODUCTION

ChatGPT (developed by OpenAI) is a powerful Artificial Intelligence Chatbot that is underpinned by Large Language Models (LLM) (OpenAI, 2023). At the time of writing, it had received extensive media coverage with its ability to provide answers to questions, conversational dialogue and even produce computer code in response to specific requests. However, ChatGPT is not perfect and can provide hallucinations (providing incorrect but plausible outputs), generate verbose responses, and only having been trained on data present up until 2021 (Chatterjee et al., 2023).

This paper explores the augmentation of an incident report dataset with ChatGPT for System Safety, notably utilizing it to support typical queries asked during safety assessment activities. The data selected for this paper is an extract of publicly accessible Aviation Safety Reporting System (ASRS) reports that describe accidents and incidents revealing causes, consequences and hazards. The ability to interface with such a large repository of information in a human-like way could be of real benefit to safety practitioners, unlocking knowledge that would otherwise only be revealed through targeted lexical searches.

Knowledge comprises of structured and organized information after cognitive processing and verification (Duan et al., 2017). It is a key component when undertaking a safety assessment (e.g., performing hazard identification) where lessons learnt from previous incidents can be incorporated into a system, further improving safety going forward.

This paper is structured as follows; the Background section provides a brief overview of ChatGPT, the ASRS incident reporting database and knowledge within safety assessment. The next section describes how ChatGPT was utilized for System Safety, before the Results section provides sample responses from the newly adapted ChatGPT. The Discussion examines the results, highlighting issues and limitations for future work.

BACKGROUND

CHATGPT

This section aims to provide a high-level overview of ChatGPT.

Human language is complex, comprising of a catalogue of words (lexicon) alongside structural rules (grammar) allowing meaning when combining the words into sentences (Manning et al., 1999). As such, human language presents a challenge for developing accurate NLP systems. The ongoing advancement of NLP has seen the deployment of more advanced machine learning and AI techniques to solve NLP tasks.

ChatGPT adapts a machine learning model; Generative Pre-trained Transformer (GPT) for dialogue, providing a conversational ability with humans.

GPTs are a type of LLM that consist of a neural network that uses self-attention to process sequential data. They can typically be fine-tuned for various natural language tasks such as translation or text classification. LLM is a term used to describe large pre-trained language models, where the larger scale model or increased data size demonstrates model improvement (Shanahan, 2022).

ChatGPT takes an existing GPT model and rather than training it on even more labelled data (a resource intensive task) it is further trained using human feedback, through reinforcement learning. This helps reduce some of the undesirable outputs such as (OpenAI, no date):

- Hallucination – Where the model makes up responses.
- Bias – Output of toxic responses.
- Poor interpretability – It is unclear how the model arrived at a given response.

Although this process does not entirely mitigate the issues above, it should reduce them.

ASRS DATABASE

ASRS is an incident reporting system that allows those involved within aviation to report aviation incidents. Its purpose is to collect, analyze, and respond to voluntarily submitted aviation safety incident reports to lessen the likelihood of aviation accidents (NASA, no date).

Incident reports, as a large body of structured data represent a significant source of knowledge, which help to;

- Understand why accidents do not occur (Johnson, 2003).
- Provide insight into system failures, human error and regulatory weakness (Van der Schaff, 1991).
- Promulgate learning from experience.

ASRS contains a publicly available database of de-personalised reports available to researchers—it is this database that this paper utilizes.

KNOWLEDGE WITHIN SAFETY ASSESSMENT

Conducting safety assessments draws together a number of components, some of which are shown in Figure 1.



Figure 1: Safety assessment components

One of the core attributes is the identification of hazards in the system and evaluating possible scenarios which lead to unwanted consequences. The identification of hazards is paramount, as no action can be taken to design out, avoid, or reduce the effects of unidentified hazards. A number of structured and systematic techniques exist for identifying and evaluating hazards. For

example, in the civil aerospace domain Functional Hazard Analysis (FHA) are recommended to identify the scope and criticality of functional failures; then analyses such as Fault-Tree Analysis (FTA), Failure Modes and Effects Analyses (FMEA), and Common Cause Analysis (CCA) methods are recommended as suitable analyses for safety certification (Society of Automotive Engineers, 1996). To reliably conduct these techniques, useage data and knowledge is required.

Incident datasets such as ASRS provide a vast source of information that can supplement individuals’ knowledge when undertaking a safety assessment. The definition of given risks is conditional upon knowledge (Aven et al., 2018). The introduction of ChatGPT provides an option to ‘unlock’ this knowledge when conducting safety assessments. This may be through improved searching of data or providing specific answers to queries. The ability to have a ‘fuzzy’ search engine (to some extent can be fulfilled by ChatGPT) within incident databases has previously been suggested (Kletz, 2008).

The usefulness of such a system goes beyond safety assessment support to assisting accident investigations and safety management system monitoring.

AUGMENTING ASRS WITH CHATGPT

The following section describes how ASRS data was augmented with ChatGPT, using the methodology and Python code described by Briggs (2023) as a guide.

Rather than ask ChatGPT specific questions that it is unlikely to answer, a method called Retrival Augmented Generation is applied. This takes the user’s query, to search (and retrieve) relevant reports from the ASRS data, then provides these with the query to ChatGPT. With this additional context of ASRS reports, ChatGPT should be able to answer the query. The general process in relation to the original ChatGPT querying process is shown in Figure 2.

A publicly available dataset of ASRS reports collated by Hoole (2022) was used for this paper, where each report was condensed to its unique identification, report date, narrative and synopsis. For this paper the dataset was filtered to only

include reports for the Boeing 737 (all models) which totalled 5136 reports.

Initially, this dataset was used to build a vector database. A vector is an array of numbers, representing more complex objects such as ASRS reports in a continuous high dimensional space—forming embeddings that map the semantic features of the reports. Embeddings can then be used for in machine learning applications such as searching relevant reports.

The vector database was hosted by Pinecone software (<https://www.pinecone.io/>) which indexes and stores the report embeddings for similarity searching and fast retrieval. Once each ASRS report is assigned a vector then nearest

neighbours (similar reports) can be easily found against a user query, therefore allowing a semantic search capability revealing knowledge that would otherwise have been difficult to find through traditional lexical (key word) searches.

With the vector database constructed, code can be produced to take a user query, search the vector database, retrieving relevant reports before feeding these reports (as context) and the original query to ChatGPT to provide an answer.

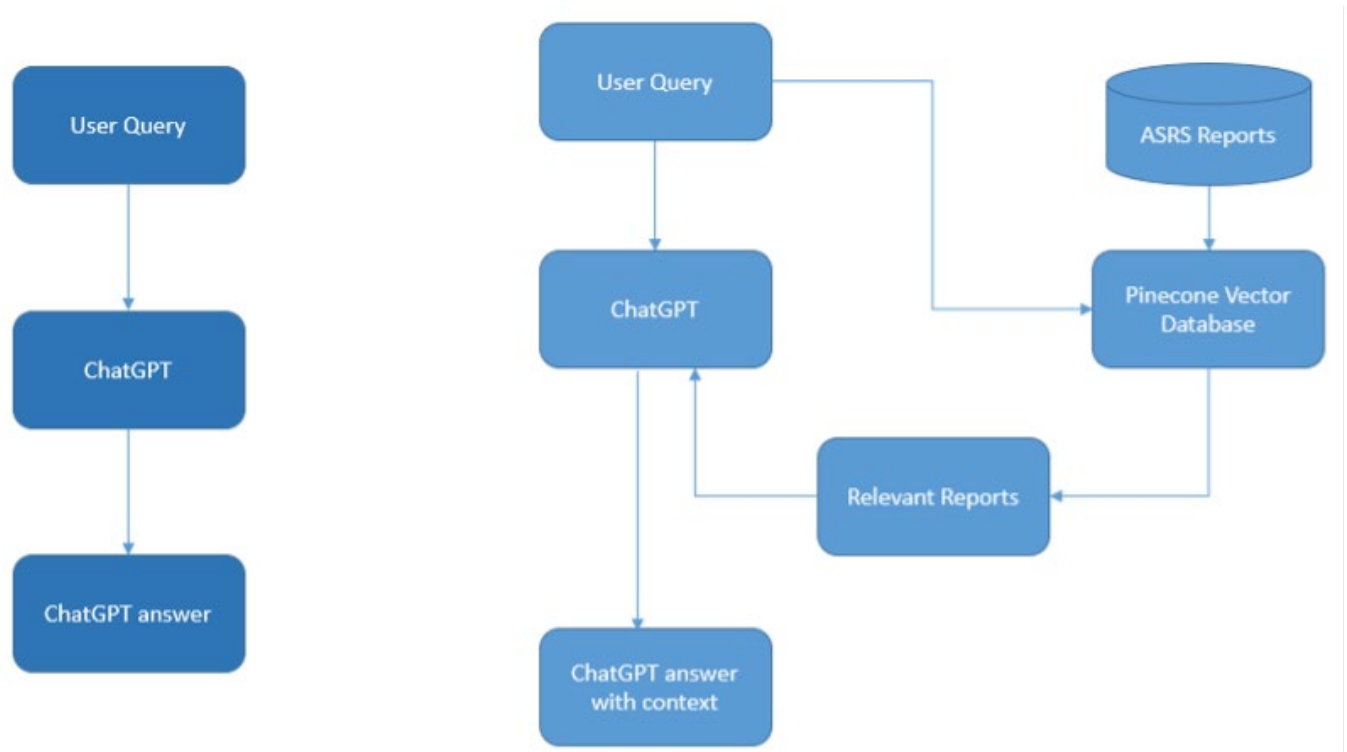


Figure 2: Original ChatGPT query process (shown on the left) alongside the Retrieval Augmented Generation process (depicted on the right)

RESULTS - SAFETY ASSESSMENT SUPPORTED BY CHATGPT

To assess the usefulness of the newly developed framework, a number of general queries can be run.

For example, a simple query was presented to the original ChatGPT; “has the aircraft ever had a fuel leak?” to which ChatGPT responded with; “No, the Boeing 737 has never had a fuel leak. The aircraft has an excellent safety record and has been in service since 1967.” Clearly, this is highly unlikely, plus ChatGPT goes further to state the aircraft has an excellent safety record, failing to account for 219 hull losses of the type (Aviation Safety Network, no date).

The same query was presented to the Retrieval Augmented Generation method described in the previous section. In this instance, two reports featuring fuel leaks (ACN 1782088 & 1070068) were retrieved and provided to ChatGPT as a context, allowing ChatGPT to respond with; “Yes, the aircraft has experienced fuel leaks on two occasions. The first instance was reported by the B737-800 pilot who refused the aircraft due to fuel leaking from the refuel manifold, which was later found to be caused by rag/cloth fragments preventing proper closure of the valve. The second instance occurred during a flight, where fuel was observed spraying from the right wing near the winglet area. The pilot suspected a fuel leak, declared an emergency, and returned for landing. However, upon landing, the airport rescue and firefighting (ARFF) team did not observe any fuel leaking”.

It can be seen, for the example query above, that the accuracy of ChatGPT’s responses has greatly improved with Retrieval Augmented Generation. The retrieval of similar reports matching the query is a useful tool in itself, ChatGPT provides the further step of summarising the retrieved reports and providing a human-like response.

A greater challenge is provided when asking questions to hypothesize what could happen (e.g., “What if?” questions asked during a safety assessment). A new query was created; “What are the consequences of a hydraulic failure?” to which ChatGPT responded with “Based on the context

provided, it appears that the narrator, who is the pilot-in-command (PIC), and their first officer were dealing with a hydraulic issue on their aircraft during a flight. They noticed a loss of A hydraulic fluid but could not find any leaks. They monitored the hydraulic level closely and observed a small loss during the flight. They requested maintenance assistance, performed checks, and shut off the A engine driven pump to prevent further loss of hydraulic fluid. They also switched the autopilot to the first officer’s side without informing them, which caused a breakdown in crew coordination. They communicated with dispatch and maintenance, and eventually decided to divert to their destination due to the need for significant maintenance. The narrator reflects on the need for better communication with their first officer, the importance of considering all possibilities in a situation, and the possibility of declaring an emergency. They also acknowledge the diligence of the mechanics and suggest checking aircraft for possible issues even if no leaks are found after hydraulic servicing”. Some useful information is provided, however, a list of possible consequences would be a better answer rather than a summary of reports mentioning hydraulic leaks.

In this case, the code was written to return the two closest matching reports. An instant improvement would be to return a greater number of matching reports, providing more context to ChatGPT and improving the response (i.e., removing the reference to fuel leaks on two occasions, as there are almost certainly more).

DISCUSSION

This paper has demonstrated the augmentation of an incident dataset with ChatGPT for safety-specific question and answering.

Where this paper used publicly accessible ASRS reports, future versions could equally use alternative data sources such as collections of safety assessment reports or other incident datasets (or a mixture). However, this is where a key limitation is found with data privacy. ChatGPT is accessed via an Application Programming Interface (API), therefore the data is submitted to a third party, although it is not a problem for data in the public domain. Sensitive data with personal details or restrictive data policies (e.g., government

data) cannot be utilized in this way.

A deeper risk remains around over-reliance on such tools within System Safety settings. In the context of this paper, it might lead to the user blindly believing the output of the model—however, if the model is set-up in such a way to show the source/evidence of its answer then that can alleviate this risk (e.g., extracts and references to ASRS reports).

Where ChatGPT could answer general safety-type questions (e.g., has there been a fuel leak?), it struggled with hypothetical questions, where knowledge is required to answer. This is most likely due to the data used to train ChatGPT not being inclusive of typical safety text such as incident reports, safety assessment reports, etc. A potential solution to this is to investigate bespoke LLMs which have been trained on large quantities of safety text, therefore accounting for unique terminology and context.

The vector database used to provide reports relevant to the user query shows promise without the additional ChatGPT feature. The database could easily be used to underpin a semantic search ability for more accurate and meaningful searches of incident datasets.

A further consideration are the costs associated with running a bespoke question and answering service. The main computational components are the creation of the word embeddings (OpenAI) and vector database (Pinecone). Both have costs; where the generation of embeddings had a one-off cost of \$3 and maintaining a vector database costs \$70 per month. For a year, this would cost ~\$840. For a large organization with a lot of data, this might be an easily absorbed cost, especially considering the cost of alternative manual methods. For the smaller organization, this represents an additional burden.

A wider public concern is how ChatGPT will impact jobs and how disruptive will it be? For System Safety, the author does not envisage ChatGPT (or similar technology) replacing safety practitioners, rather forming an efficient tool to use for common safety tasks, such as supporting safety assessments. If the system is set up in such a way to provide evidence for outputs then it should allow for the user to gain more trust. Where Microsoft has invested in OpenAI, there are plans to incorporate technology such as

ChatGPT into everyday Office products (Zhao et al., 2023)—methods like what is described in this paper may soon be commonplace. Therefore, it is likely that safety practitioners will require basic knowledge in AI, plus the benefits and limitations going forward.

CONCLUSION AND FUTURE WORK

This paper has taken an incident dataset and augmented it with ChatGPT to support safety assessment and general safety queries. It has been shown that ChatGPT can be further utilized for knowledge retrieval to provide general safety assessment support, however, a major limitation to deploying this method elsewhere are data protection policies. It is proposed if the model is setup so that outputs provide evidence or sources behind answers, and are used for advisory purposes then they can form useful tools for System Safety practitioners.

Future work may consider further trials of this concept, perhaps even in actual safety assessment tasks. The creation of bespoke datasets for safety tasks would also allow machine learning models to be fine-tuned on safety-specific data, therefore, improving accuracy when supporting System Safety tasks.

ACKNOWLEDGEMENTS

J Ricketts is pleased to acknowledge the contribution of the IMechE Whitworth Senior Scholarship award in supporting this research.

COMPETING INTERESTS

No Competing Interests.

ORCID IDS

Jon Ricketts  <https://orcid.org/0000-0001-9487-9092>

REFERENCES

- [1] OpenAI, "Introducing ChatGPT," 2023. [Online]. Available: <https://openai.com/blog/chatgpt>. [Accessed: 29-Mar-2023].
- [2] J. Chatterjee and N. Dethlefs, "This new conversational AI model can be your friend, philosopher, and guide. and even your worst enemy," *Patterns*, vol. 4, no. 1, pp. 1–3, 2023, doi: <https://doi.org/10.1016/j.patter.2022.100676>.
- [3] Y. Duan, L. Shao, G. Hu, Z. Zhou, Q. Zou, and Z. Lin, "Specifying Architecture of Knowledge Graph with Data Graph, Information Graph, Knowledge Graph and Wisdom Graph," in *IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, 2017, pp. 327–332, doi: <https://doi.org/10.1109/SERA.2017.7965747>.
- [4] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.
- [5] M. Shanahan, "Talking About Large Language Models," *arXiv Prepr.*, pp. 1–13, 2022.
- [6] OpenAI, "ChatGPT: Optimizing Language Models for Dialogue." [Online]. Available: <https://openai.com/blog/chatgpt/>. [Accessed: 10-Feb-2023].
- [7] NASA, "Aviation Safety Reporting System. Program Briefing." [Online]. Available: <https://asrs.arc.nasa.gov/overview/summary.html>. [Accessed: 04-Apr-2023].
- [8] C. W. Johnson, *A Handbook of Incident and Accident Reporting*. Glasgow: Glasgow University Press, 2003.
- [9] T. Van der Schaff, *Near Miss Reporting as a Safety Tool*. Butterworth Heinemann, 1991.
- [10] Society of Automotive Engineers, "ARP 4761 Guidelines and Methods for conducting the Safety Assessment Process on Civil Airborne Systems and Equipment," 1996.
- [11] T. Aven and E. Zio, *Knowledge in risk assessment and management*. Chichester: Wiley, 2018. <https://doi.org/10.1002/9781119317906>
- [12] T. A. Kletz, "Searchlights from the past," *J. Hazard. Mater.*, vol. 159, no. 1, pp. 130–134, 2008, doi: <https://doi.org/10.1016/j.jhazmat.2007.09.119>.
- [13] J. Briggs, "Retrieval Enhanced Generative Question Answering with OpenAI," 2023. [Online]. Available: <https://github.com/pinecone-io/examples/blob/master/generation/generative-qa/openai/gen-qa-openai/gen-qa-openai.ipynb>. [Accessed: 06-Apr-2023].
- [14] E. Hoole, "ASRS Aviation Reports Dataset," 2022. [Online]. Available: <https://huggingface.co/datasets/elihoole/asrs-aviation-reports>. [Accessed: 13-Apr-2023].
- [15] Aviation Safety Network, "Boeing 737." [Online]. Available: <https://aviation-safety.net/database/types/Boeing-737-series/index>. [Accessed: 13-Apr-2023].
- [16] W. X. Zhao et al., "A Survey of Large Language Models," *arXiv Prepr.*, pp. 1–51, 2023.