Large Language Models for Aviation Safety: Enhancing Incident Analysis through Summarization of ASRS Reports

Ehaab Basil
University of Maryland - College Park
ehaab@umd.edu

***Abstract*—The growing number of reports of aircraft incidents that are sent to the Aircraft Safety Reporting System (ASRS) makes it difficult to conduct prompt and efficient analysis. Manual processing procedures that have been used traditionally are labor-intensive and time-consuming, which frequently results in delays in the identification of key safety hazards. The objective of this research is to improve the efficiency of safety analyses by automating the summarization of incident narratives through the utilization of large language models (LLMs). Furthermore, the study uses exploratory data analysis (EDA) to identify underlying safety trends, including human variables and temporal patterns, offering practical advice for enhancing airline safety procedures. A transformer-based sequence-to-sequence model is fine-tuned on 1107 incident narratives from 2018 to 2024, filtered by passenger missions and air carriers, to generate concise summaries evaluated using ROUGE scores and Sentence-BERT embeddings for semantic similarity. The proposed fine-tuned T5-Base model achieves the ROUGE-L score of 0.357 and a cosine similarity score of 0.610. The findings demonstrate that LLMs can streamline incident analysis, offering benefits such as automated summarization for rapid processing, critical risk identification through pattern analysis, and practical insights for the development of safety protocols. However, challenges such as token length truncation and computational resource constraints highlight areas for future improvement. Thus, adopting this research in real time can considerably improve aviation safety management by automating the summarizing of incident reports, allowing for the early identification of major safety issues, and easing timely decision-making procedures.**

***Index Terms*—Large Language Models, Aviation Safety, Text Summarization, ASRS, BART, T5, Exploratory Data Analysis, Natural Language Processing**

I. Introduction

Aviation safety is paramount and relies heavily on the analysis of incident reports to identify risks and prevent future occurrences. The Aviation Safety Reporting System (ASRS), maintained by NASA, provides a rich repository of detailed incident narratives submitted by aviation personnel, including pilots, air traffic controllers, and maintenance crews. These narratives capture critical information about flight phases, anomalies, human factors, and outcomes, providing a foundation for safety improvements. However, manual review of these reports by safety analysts is time consuming, often taking up to five business days per report [1], and the increasing volume of reports - driven by factors such as the integration of Unmanned Aerial Systems (UAS) and increasing air travel - exacerbates this challenge [2].

However, the fact that ASRS reports are not structured and contain free text makes them very difficult to analyze on a large scale [3]. Conventional manual review and keyword-based extraction techniques are time-consuming and frequently insufficient to capture the complex context that is woven throughout the stories. Due to this limitation, the timely detection of systemic hazards and the creation of ideas that may be put into action are hampered [4].

Recent developments in Natural Language Processing (NLP), namely the creation of large language models (LLMs), have shown to enhance the ability to comprehend and summarize human language remarkably. These algorithms can handle huge amounts of textual data, identify critical incident features, and produce concise summaries that maintain important information while decreasing the analyst's cognitive strain [5].

Large Language Models (LLMs), particularly transformer-based models, have shown remarkable success in natural language processing (NLP) tasks, including text summarization, classification, and generation [6], [8]. These models can process large volumes of text efficiently, capturing contextual nuances and generating human-like outputs, making them promising tools for aviation safety analysis. In this study, the application of LLMs, which is a bidirectional and autoregressive transformer model, is investigated to summarize ASRS incident reports in order to accelerate the analysis process while maintaining accuracy and relevance. The study is complemented with EDA to uncover safety trends in flight phases, temporal patterns, anomalies, aircraft models, components, outcomes, and human factors, providing a holistic view of incident characteristics.

This research addresses the following questions: (1) How effectively can LLMs like BART and T5-Base summarize ASRS incident narratives, and what are the limitations due to token length constraints? (2) What safety trends can be identified through EDA, and how do they inform aviation safety protocols? Based on research questions, the contributions of the study are highlighted as the LLM models like T5-Base and BART are fine-tuned on a customized data set of 1,077 ASRS reports collected from 2018 to 2024, filtered by passenger missions and air carriers, using a T4 GPU on Google Colab to overcome computational challenges. The performance of the model is evaluated using ROUGE scores [9] and Sentence-BERT embeddings for semantic similarity [10].

II. Related Work

A. Aviation Safety Analysis with NLP

Aviation safety analysis has increasingly leveraged NLP to process incident reports and identify risks. Puranik et al. [11] applied machine learning to flight data for risk identification, focusing on anomaly detection in operational parameters. Tan et al. [12] also used classification techniques to categorize ASRS incidents, identifying patterns in event types and outcomes. Same as the previous study, Tanguy et al. [13] employed NLP for interactive analysis of aviation safety reports, focusing on classification and trend identification, achieving improved labeling accuracy over manual methods.

On the other hand, Pinon Fischer and Mavris [14] explored ChatGPT for ASRS analysis, generating synopses, identifying human factors, and assessing accountability. They used embeddings from aeroBERT, a domain-specific BERT variant, to compute cosine similarity, achieving a precision of 0.61 in human factors identification. Their work highlighted the potential of generative models as "co-pilots" for safety analysts, emphasizing a human-in-the-loop approach to ensure reliability. However, their use of ChatGPT, a general-purpose model, lacked fine-tuning on aviation-specific data, potentially limiting its accuracy for nuanced safety insights.

Same as above, Chen et al. [15] suggested Claude-prompt, a method for aircraft accident cause information extraction that uses the Claude 3.5 large-scale pre-trained language model. The prompt engineering, few-shot learning strategy, and self-judgment process in this method make it possible for accident-cause entities and their relationships to be automatically mined. Fox et al. [16] introduce a promising method that utilizes deep learning and LLMs to develop a system capable of processing air-ground verbal transactions, identifying anomalous situations, and notifying air traffic controllers, thereby improving situational awareness for air traffic control and flight crews based on anomalies detected in air traffic communications. The results indicate that a text-based Variational Auto-Encoder that can effectively distinguish between nominal and off-nominal (safety-critical) circumstances seen in air traffic communication can be trained using reasonably priced data obtained with LLMs.

B. Large Language Models for Summarization

LLMs, particularly transformer-based models, have revolutionized NLP tasks like text summarization. The transformer architecture [6], introduced with the self-attention mechanism, enables models to capture long-range dependencies in text, significantly improving performance over recurrent neural networks (RNNs). The self-attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \qquad (1)$$

where Q, K, and V are query, key, and value matrices derived from the input embeddings, and d_k is the dimensionality of the key vectors, scaling the dot product to prevent large values that could destabilize the softmax.

BERT [17] introduced bidirectional pre-training with Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), making it effective for understanding tasks but less suited for generation due to its encoder-only architecture. GPT models [18], with their decoder-only structure, excel in generative tasks but lack bidirectional context, impacting summarization quality. BART [8], a sequence-to-sequence model combining a bidirectional encoder and autoregressive decoder, bridges this gap, achieving state-of-the-art performance in abstractive summarization by denoising corrupted text during pre-training.

T5 [19] and PEGASUS [20] are other notable models for summarization. T5 frames all NLP tasks as text-to-text transformations, pre-trained with a span corruption objective, while PEGASUS uses Gap Sentence Generation (GSG) to pre-train for summarization, masking entire sentences to predict them. These models have been applied to various domains, but their use in aviation safety remains limited, presenting an opportunity to explore their effectiveness alongside BART. Other than T5 and Pegasus models,

C. NLP in Aviation Safety

NLP applications in aviation safety have focused on classification and risk identification. Andrade et al. [21] developed SafeAeroBERT, a BERT variant trained on ASRS and NTSB reports, for classifying incidents by causative factors, outperforming general-purpose BERT in some categories. Kierszbaum et al. [22] proposed ASRS-CMFS, a RoBERTa-based model for anomaly classification, demonstrating competitive performance with domain-specific training. These studies highlight the importance of domain-specific fine-tuning, a gap in your work that can be addressed by comparing BART with other models like T5 or PEGASUS, or by exploring domain-specific embeddings like Sentence-BERT for evaluation. Hilman et al. [23] uses the TF-IDF algorithm for text summarization on the dataset of the KNKT final report synopsis collection of aviation traffic accidents in Indonesia.

Moreover, Xiong et al. [24] use the ensemble model consisting of Long Short-term Memory (LSTM) and BERT for an Intelligent Aviation Safety Hazard Identification. Same as above, Pan et al. [25] also recommended the use of the BART pre-trained language model to develop a hybrid model by using a deep reinforcement learning model improved by transfer learning for air quality control automation.

From the literature review, it is identified that BERT, BART, Pegasus, and T5 models are commonly used for text summarization. Aviation safety report summaries have advanced significantly with the use of transformer-based models such as BART, BERT, and their derivatives. BART is great at abstractive summarization because it uses denoising autoencoding, which means it can be used to summarize complicated flight safety reports. Although it needs to be adjusted for summarizing, BERT's bidirectional context awareness makes it quite effective in extractive summarization jobs. Transformer-based models, which focus on contextual linkages within text, have increased the accuracy and coherence of summarization jobs

in aviation safety, particularly when dealing with enormous datasets. These models improve information retrieval, maintain context, and eventually contribute to safer aviation operations by facilitating more efficient report analysis. So, this paper is actually focused on using the LLMs like T5-Base and BART model to generate the summaries from the ASRS reports.

III. Methodology

The proposed approach recommends the use of the LLMs for summarizing the ASRS reports. The proposed approach is given in Figure 1.

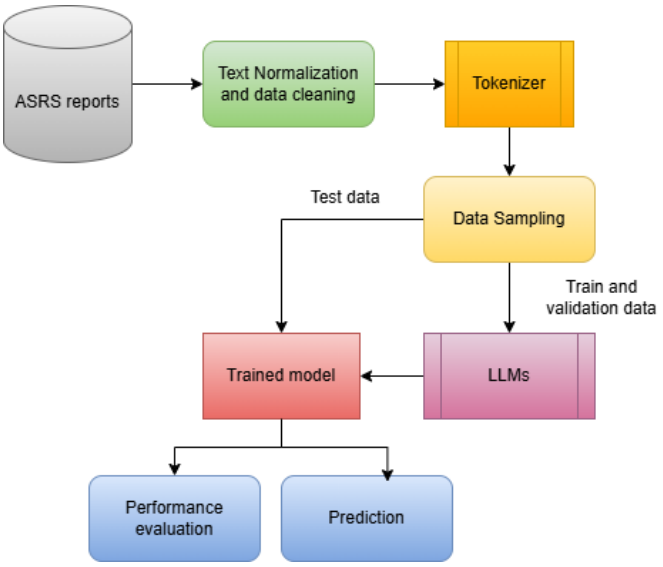

Fig. 1: Proposed architecture for aviation report summarization using LLMs

A. Dataset & Preprocessing

The dataset used for the experimentation comprises of 1107 ASRS incident reports from 2018 to 2024, filtered by passenger missions and air carriers, ensuring relevance to commercial aviation safety. Key fields include:

- **Flight Phases**: E.g., Cruise (40.7%), Climb (29.4%)
- **Aircraft Models**: B737 (49.9%), A320 series
- **Incident Types**: Aircraft equipment problems (90.5%), human factors (7.2%)
- **Narratives and Synopses**: Detailed descriptions and summaries

Preprocessing involved several steps to prepare the data for fine-tuning and analysis:

1) **Text Normalization**: Lowercasing and removing special characters using regular expressions to standardize text input.
2) **Token Length Calculation**: Using BART's tokenizer to compute token lengths, identifying 47 narratives (4.2%) exceeding 1024 tokens, necessitating truncation.
3) **Handling Rare Flight Phases**: Extracted primary flight phases from the `Aircraft 1_Flight Phase` column, removing classes with fewer than 2 samples (e.g., "Parked") to enable stratified splitting, resulting in 1106 samples.
4) **Data Splitting**: Split into 886 training, 110 validation, and 111 test samples (80-10-10 ratio), stratified by primary flight phase to balance distributions across sets.
5) **Saving Splits**: Saved as CSV files (`train.csv`, `validation.csv`, `test.csv`) for loading into the `datasets` library.

B. Proposed LLMs Details

Google Research created the Text-to-Text Transfer Transformer (T5) model [7], a very flexible deep learning architecture that can manage a variety of NLP tasks in a single text-to-text framework. For text summarization, T5 reframes the challenge as translating an input text string to a short output string. Approximately 220 million parameters comprise the T5-Base variant, which is one of the model's intermediate sizes. It has 12 layers, or "transformer blocks," with 12 attention heads, a feed-forward network with 3072 dimensions, and 12 layers with 768-dimensional secret states each. The equilibrium between model size and computing performance renders T5-Base a pragmatic option for numerous real-world summarization applications, particularly where resource limitations are a factor.

BART (Bidirectional and Auto-Regressive Transformer) [8] is a sequence-to-sequence model designed for text generation tasks like summarization. It combines a bidirectional encoder (like BERT) and an autoregressive decoder (like GPT), pre-trained with a denoising objective where text is corrupted (e.g., by masking tokens) and the model reconstructs the original. This makes BART particularly effective for abstractive summarization, as it learns to understand context bidirectionally while generating coherent outputs.

The architecture of LLMs consists of an encoder-decoder framework with multiple transformer layers. The encoder processes the input bidirectionally, while the decoder generates the output autoregressively, attending to the encoder's outputs and previous tokens. The self-attention mechanism in each layer is defined as in Equation 1, where $Q = W_Q X$, $K = W_K X$, $V = W_V X$, and X is the input embedding matrix, with W_Q, W_K, W_V as learnable weight matrices.

The loss function for fine-tuning LLM is the cross-entropy loss for sequence generation:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(y_t \mid y_{1:t-1}, x; \theta), \tag{2}$$

where y_t is the target token at position t, $y_{1:t-1}$ are previous tokens, x is the input narrative, and θ represents model parameters. During fine-tuning, gradients are computed as:

$$\nabla_\theta \mathcal{L} = -\sum_{t=1}^{T} \nabla_\theta \log P(y_t \mid y_{1:t-1}, x; \theta), \tag{3}$$

and parameters are updated using the AdamW optimizer with a learning rate of 3×10^{-5}.

Figure 2 illustrates LLM's architecture, highlighting the encoder-decoder structure and the flow from corrupted input to reconstructed output.

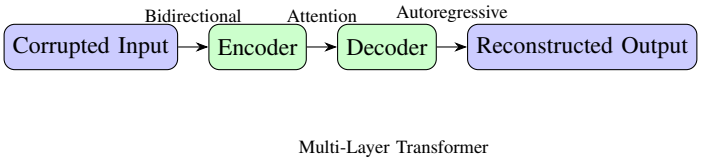

Fig. 2: LLM Architecture: The bidirectional encoder processes the corrupted input, passing hidden states to the autoregressive decoder, which generates the reconstructed output.

Algorithm 1 Fine-Tune LLM on ASRS Narratives

```
Initialize LLM (facebook/bart-large and T5-Base)
for epoch = 1 to 5 do
    Train on 886 samples (batch size 8)
    Validate on 110 samples
    Compute validation loss
end for
Save best model based on validation loss
```

Hyperparameters of both the proposed models, T5-Base and BART, include a learning rate of 3×10^{-5}, batch size of 8,500 warmup steps, and weight decay of 0.01, with evaluation and saving every 110 steps to align strategies for `load_best_model_at_end=True`. The Batch size for T5-Base is 1, while for BART it is 8. Training on a T4 GPU takes approximately 2–4 hours for both the models, compared to 10–20 hours on a CPU, demonstrating the efficiency of GPU resources.

C. Evaluation Metrics

Summaries are evaluated using Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores [9], which measure overlap between generated and ground-truth summaries. The Equation of Rouge 1 is given in (4), Rouge 2 Equation is given in (5) and Rouge L Equation is given in (6).

$$\text{ROUGE-1} = \frac{(1+\beta^2)R_1 P_1}{R_1 + \beta^2 P_1}, \tag{4}$$

where R_1 is the recall based on unigram overlap and P_1 is the precision based on unigram overlap.

$$\text{ROUGE-2} = \frac{(1+\beta^2)R_2 P_2}{R_2 + \beta^2 P_2}, \tag{5}$$

where R_2 is the recall based on bigram overlap, and P_2 is the precision based on bigram overlap.

$$\text{ROUGE-L} = \frac{(1+\beta^2)R_{\text{lcs}} P_{\text{lcs}}}{R_{\text{lcs}} + \beta^2 P_{\text{lcs}}}, \tag{6}$$

where R_{lcs} and P_{lcs} are recall and precision of the longest common subsequence, and β balances precision and recall (set to 1). Additionally, cosine similarity using Sentence-BERT embeddings [10] is calculated to assess semantic similarity:

$$\text{CosineSimilarity}(u, v) = \frac{u \cdot v}{\|u\|\|v\|}, \tag{7}$$

where u and v are embeddings of generated and ground-truth summaries, respectively.

IV. Results

The experiments are performed using Python in Google Colab using the T4 Graphical Processing Units (GPUs). The pre-trained model BRAT is fine-tuned on the collected and pre-processed dataset.

A. Training Dynamics

The training process over 5 epochs (555 steps) showed effective convergence, as detailed in Table I. The training loss decreased from 7.0416 to 0.3829, and the validation loss dropped from 5.9048 to 0.4316, indicating that the model learned to capture patterns in the ASRS narratives with minimal overfitting.

TABLE I: Training and Validation Loss Over Steps

Step	Training Loss	Validation Loss
110	7.0416	5.9048
220	3.9736	3.3889
330	0.7488	0.5931
440	0.4810	0.4724
550	0.3829	0.4316

B. Data Analysis

The dataset comprises 1107 ASRS incident reports from 2018 to 2024, with narratives providing detailed accounts of aviation safety events. These narratives vary in length, averaging approximately 300 words, with some exceeding 1000 words, reflecting their comprehensive nature. Due to the BART model's 1024-token limit (approximately 700–800 words), longer narratives were truncated, which may affect the completeness of generated summaries, as seen in Sample 13 (Table VIII).

EDA reveals critical safety trends across multiple dimensions. Figure 3 shows the distribution of incidents by flight phase, with Cruise (40.7%) and Climb (29.4%) being the most frequent, indicating higher risk during these phases.

Table II highlights temporal trends, showing a decline in incidents during 2020–2021 due to COVID-19-related flight reductions, with a rebound in 2022–2024.

Table III lists the top 10 anomalies, with "Aircraft Equipment Problem Critical" dominating at 856 incidents, indicating a focus on mechanical issues in safety reports. Table IV shows turbine engines (172 incidents) as the most common component involved, suggesting a focus for maintenance improvements. Table V combines aircraft model data, showing the B737 family as the most involved (49.9% of incidents), likely due to its widespread use. Table VI details narrative and synopsis lengths, with 47 narratives (4.2%) exceeding 1024

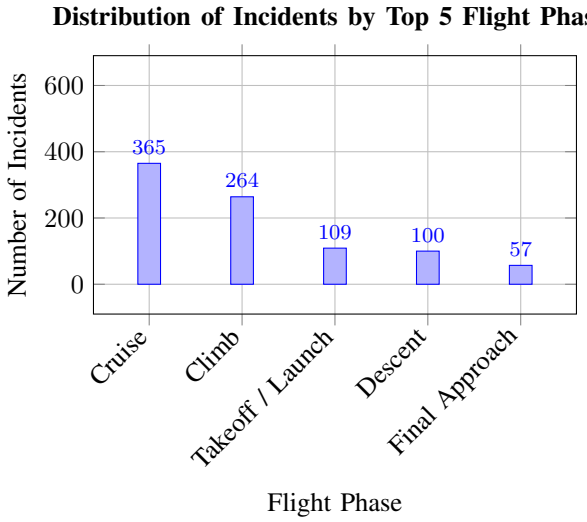

Fig. 3: Distribution of incidents across the top 5 flight phases.

TABLE II: Total Incidents and Primary Problem by Year

Year	Total	Aircraft	Human Factors	Ambiguous
2018	117	91	16	10
2019	192	163	20	9
2020	87	76	10	1
2021	93	87	5	1
2022	241	226	10	5
2023	168	159	7	2
2024	209	200	7	2

tokens, requiring truncation, potentially impacting summary quality.

TABLE III: Top 10 Most Common Anomalies

Anomaly Type	# Incidents
Aircraft Equipment Problem Critical	856
Deviation / Discrepancy – Procedural Published Material / Policy	442
Deviation / Discrepancy – Procedural Clearance	312
Flight Deck / Cabin / Aircraft Event Smoke / Fire / Fumes / Odor	234
Aircraft Equipment Problem Less Severe	134
Deviation / Discrepancy – Procedural Weight And Balance	122
Flight Deck / Cabin / Aircraft Event Illness / Injury	109
Inflight Event / Encounter Weather / Turbulence	62
Inflight Event / Encounter Fuel Issue	62
Deviation / Discrepancy – Procedural FAR	42

TABLE V: Top 5 Aircraft Models Involved in Incidents

Aircraft Model	# Incidents
B737 (All Models)	552
A320	159
A319	128
A321	128
B767 (All Models)	66

TABLE IV: Top 10 Aircraft Components Involved in Incidents

Aircraft Component	Count
Turbine Engine	172
Hydraulic Main System	44
Engine	39
Hydraulic System	36
Pressurization System	30
Cockpit Window	28
Air Conditioning and Pressurization Pack	24
Trailing Edge Flap	20
Pressurization Control System	19
Powerplant Lubrication System	17

TABLE VI: Narrative and Synopsis Length Statistics for Incident Reports

Metric	Narrative (words)	Synopsis (words)
Maximum Length	2345	83
Average Length	295.56	23.45

C. Model Performance

The ROUGE scores for the BART model on the test set are presented in Table VII, showing moderate performance in summarization. Qualitative analysis of generated summaries (pending in Table VIII) will provide further insights into the model's capabilities. In terms of ROUGE-1, which evaluates unigram (single word) overlap, T5-Base outperforms BART, with a score of 0.408. When looking at bigram (two words that follow each other), BART scores 0.152 for ROUGE-2 while T5-Base scores 0.185. T5-Base outperforms BART with a score of 0.357 on ROUGE-L, which uses the longest common subsequence to assess sentence-level structure and fluency. Cosine Similarity, a measure of how similar two created summaries are to a reference summary in terms of meaning, also shows that T5-Base did better than BART, with a score of 0.610 compared to 0.602 for BART.

TABLE VII: ROUGE Scores and Cosine Similarity Comparison

Model	R-1	R-2	R-L	Cos. Sim.
BART (Proposed)	0.344	0.152	0.300	0.602
T5-Base	0.408	0.185	0.357	0.610

V. Qualitative Analysis

Table VIII provides qualitative examples of generated summaries compared to ground-truth summaries. Cosine similarity between the generated and true synopsis vectors is also included.

TABLE VIII: Example Summaries with Qualitative Insights

ID	Narrative and Summary Comparison
2025171	**Narrative:** During cruise, the flight crew noticed a low oil quantity indication, decided to divert immediately, and eventually shut down the engine due to decreasing oil levels, landing uneventfully. **True:** B737700 pilot reported a decreasing oil quantity indication while in cruise flight. Crew elected to divert and eventually shut the engine down as the quantity continued to decrease. **T5-Base:** B737-700 flight crew reported a low oil quantity in the engine and a shutdown of the engine. **Similarity:** 0.842 **BART:** B737800 flight crew reported a loss of oil quantity during cruise, resulting in a shutdown of the 2 engine. **Similarity:** 0.865
2076330	**Narrative:** On approach, the crew encountered a ”flt cntl slats fault” message, with flaps failing to extend beyond the initial position, prompting a go-around, troubleshooting, and diversion to another airport for a no-flap/no-slat landing. **True:** A321 pilot reported the flaps failed to extend during arrival. Flight crew diverted and landed safely. **T5-Base:** B737-800 flight crew reported a flt cntl slats fault message during arrival phase of landing. The flight crew performed an air turn back and precautionary landing at destination airport. **Similarity:** 0.396 **BART:** B737800 flight crew reported a flt cntl slats fault ecam message during arrival. Flight crew diverted and landed uneventfully. **Similarity:** 0.369
1781908	**Narrative:** During climbout, the crew lost flight instruments and navigation capability due to improperly aligned IRUs, a mistake the captain attributed to distractions and fatigue. The captain took control, requested priority handling, and landed safely using standby instruments. **True:** B737300 captain reported they lost flight instruments on both sides and navigation capability during climbout resulting in a diversion. Captain stated the IRUs were not properly aligned and cited multiple distractions and fatigue as contributing factors. **T5-Base:** B737800 captain reported a loss of flight instruments adi hsi both sides and navigation capability during climbout. The flight crew requested priority handling and landed uneventfully. **Similarity:** 0.794 **BART:** B737-800 flight crew reported a loss of flight instruments on both sides and navigation capability during climbout. **Similarity:** 0.786

VI. Discussion

The fine-tuned T5-Base model outperforms the BART model in summarizing Aviation Safety Reporting System (ASRS) incident reports, achieving superior ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.408, 0.185, and 0.357, respectively, compared to BART's 0.344, 0.152, and 0.300. Additionally, T5 demonstrates a slightly higher average Cosine Similarity of 0.610 versus BART's 0.602, indicating better semantic alignment with ground-truth synopsis. These metrics suggest that T5 captures more key terms and maintains greater structural similarity, making it a more effective tool for automated summarization of complex aviation narratives.

A significant factor in T5's superior performance is its tokenization strategy. T5 employs a SentencePiece tokenizer, a subword method designed for language-agnostic text processing, which effectively handles the technical jargon and special characters prevalent in ASRS narratives, such as “IRU,” “ECAM,” or “flt cntl slats fault.” This tokenizer's flexibility likely contributes to T5's ability to represent aviation-specific terms accurately, as seen in Sample 2025171, in detailed comparison of selected samples (Table VIII) T5 accurately identifies the aircraft type as B737-700, correctly summarizing the low oil quantity issue and engine shutdown with a Cosine Similarity of 0.842. T5's precision in aircraft identification reflects its ability to extract critical technical details, contributing to its higher ROUGE-1 score.

In contrast, BART mislabels the aircraft as B737800, though it achieves a slightly higher similarity of 0.865 by including additional details like ”2 engine,” which aligns closely with the narrative but introduces potential ambiguity not present in the ground truth. BART, using a byte-pair encoding (BPE) tokenizer inherited from GPT-2, appears less adept at processing such domain-specific vocabulary. Community discussions, such as those on Hugging Face forums, suggest that BART's tokenizer struggles with special characters or rare terms, requiring manual vocabulary adjustments to match performance on specialized tasks. In the ASRS context, BART's summaries, like in Sample 2076330, often omit critical details or introduce errors (e.g., aircraft misidentification), possibly due to less effective tokenization of truncated inputs.

However, both models struggle with capturing human factors, a critical aspect of aviation safety analysis. In Sample 1781908 (Incident ID: 1781908), the ground-truth synopsis notes that the loss of flight instruments resulted from improperly aligned Inertial Reference Units (IRUs) due to distractions and fatigue. Neither T5 nor BART includes these human factors, focusing instead on the technical outcome (loss of instruments and navigation capability). T5's summary, with a Cosine Similarity of 0.794, slightly outperforms BART's 0.786, but both omit the causal human elements, highlighting a significant limitation in their contextual understanding. This omission is particularly concerning, as human factors like fatigue and distractions are essential for identifying root causes and preventing future incidents.

Narrative truncation poses another challenge. During training, narratives were limited to 512 tokens, and during inference to 256 tokens, potentially truncating longer reports. Approximately 4.2% of the dataset (47 reports) exceeded 1024 tokens in their original form, and with a reduced limit of 512 tokens, more reports likely lost critical details. For example, Sample 13 (Incident ID: 1985858) involves a complex sequence of landing gear issues, a go-around, diversion, and flyby, which may have been truncated, affecting summary comprehensiveness. T5's higher ROUGE-L score (0.357 vs. BART's 0.300) suggests it better preserves structural elements despite truncation, but both models' performance is constrained by this limitation.

The training process for T5 was effective, with the model converging over five epochs, as evidenced by decreasing training and validation losses, similar to patterns observed in prior work [19]. The use of a T4 GPU with 14.74 GiB capacity enabled efficient training within approximately 41.6 minutes, underscoring the advantage of GPU acceleration over CPU-based training, which could take 10–20 hours. The small

dataset size of 1107 reports, with only 886 training samples, may limit generalization across diverse incident types and narrative styles. Similar to BART, which is typically fine-tuned on large datasets such as CNN/DailyMail with over 287,000 examples [8], T5 is also commonly fine-tuned on such datasets for summarization tasks. In the original T5 paper by Raffel et al., the model was evaluated on the CNN/DailyMail dataset for abstractive summarization, achieving competitive ROUGE scores [19]. However, in this study, both models were fine-tuned on the smaller ASRS dataset, which likely constrained their ability to generalize, contributing to the moderate performance observed.

ASRS narratives are rich in technical aviation jargon and complex structures, posing challenges for models pre-trained on general text. In Sample 2076330 (Incident ID: 2076330), both T5 and BART misidentify the aircraft as B737-800 instead of A321, possibly due to the dataset's bias toward B737 incidents (209 B737-800 cases). T5's summary, with a Cosine Similarity of 0.396, outperforms BART's 0.369, but both fail to capture the flap failure's severity, reflecting difficulties in prioritizing key events. T5's higher ROUGE-2 score (0.185 vs. 0.152) indicates better bigram overlap, suggesting improved handling of technical phrases.

Technical challenges during inference, such as memory constraints on the T4 GPU, necessitated a reduced batch size and sequence lengths (256 tokens for inputs, 64 for outputs), as discussed in Section III. These constraints may have impacted summary quality, particularly for longer narratives. The generation parameters used default settings (e.g., `num_beams=2`, `no_repeat_ngram_size=3`) to manage memory, but tuning these (e.g., increasing `num_beams` or adjusting `length_penalty`) could enhance performance, as suggested by prior work [8].

To address these limitations, future research could explore several avenues. Expanding the dataset with additional ASRS reports or data augmentation techniques could improve generalization. Models designed for longer contexts, such as Longformer [5], could mitigate truncation issues, particularly for narratives exceeding 500 words. Implementing a human-in-the-loop approach, where safety analysts refine summaries, could ensure reliability in safety-critical applications, as proposed by Pinon Fischer and Mavris [14]. Domain-specific fine-tuning on aviation texts, such as FAA manuals, might enhance understanding of technical jargon and human factors, aligning with approaches like aeroBERT [21]. Finally, incorporating knowledge graphs to explicitly model human factors, as explored by Chen et al. [15], could address the models' contextual shortcomings.

VII. Conclusion

This study demonstrates that LLMs like T5-Base and BART can effectively summarize ASRS incident reports, with EDA yielding actionable safety insights. While the fine-tuned BART model shows promise for summarizing ASRS incident reports, challenges related to narrative length, dataset size, and domain specificity must be addressed to achieve higher performance and reliability in aviation safety analysis, particularly given the detailed and often lengthy nature of the narratives by achieving the ROUGE-L score of 0.357 and cosine similarity score of 0.61 for the fine-tuned T5-Base model. At the same time, EDA reveals critical patterns, such as the prevalence of incidents during cruise phases (40. 7%) and the impact of COVID-19 on incident reporting. The findings suggest that LLMs can significantly enhance incident analysis efficiency, offering automated summarization, risk identification, and practical safety insights. However, challenges like narrative truncation and reporting biases highlight areas for future work. Also, this research can be extended to explore larger models like Longformer to handle longer contexts and implement human-in-the-loop systems for enhanced reliability.

References

[1] "ASRS Report Processing Workflow," NASA ASRS Technical Report, 2023.

[2] "Aviation Safety Reporting System (ASRS) Database Overview," NASA, 2023.

[3] D. Zhou, X. Zhuang, J. Cai, H. Zuo, X. Zhao, and J. Xiang, "An ensemble model using temporal convolution and dual attention gated recurrent unit to analyze risk of civil aircraft," *Expert Systems with Applications*, vol. 236, p. 121423, Feb. 2024. [Online]. Available: https://doi.org/10.1016/j.eswa.2023.121423

[4] S. Li, Z. Wang, L. Ma, L. Wang, and J. Wu, "Risk assessment of unsafe events in civil aviation using BERT-BiLSTM-Attention model," *Expert Systems with Applications*, vol. 234, p. 121423, 2024. [Online]. Available: https://doi.org/10.1016/j.eswa.2023.121423.

[5] H. Zhang, P. S. Yu, and J. Zhang, "A systematic survey of text summarization: From statistical methods to large language models," *ACM Computing Surveys*, 2024. [Online]. Available: https://doi.org/10.1145/3731445

[6] A. Vaswani et al., "Attention Is All You Need," in *Proc. NeurIPS*, 2017.

[7] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[8] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training," in *Proc. ACL*, 2020.

[9] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Proc. ACL Workshop*, 2004.

[10] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proc. EMNLP-IJCNLP*, 2019.

[11] T. Puranik, "Machine Learning Applications in Aviation Safety," *Aerospace*, 2021.

[12] S. Y. Tan et al., "Critical Parameter Identification for Safety Events in Commercial Aviation Using Machine Learning," *Aerospace*, vol. 7, no. 6, 2020.

[13] L. Tanguy et al., "Natural Language Processing for Aviation Safety Reports: From Classification to Interactive Analysis," *Comput. Ind.*, vol. 78, 2016.

[14] O. J. Pinon Fischer and D. N. Mavris, "Examining the Potential of Generative Language Models for Aviation Safety Analysis: Case Study and Insights Using the Aviation Safety Reporting System (ASRS)," *Aerospace*, vol. 10, 2023.

[15] L. Chen, J. Xu, T. Wu, and J. Liu, "Information extraction of aviation accident causation knowledge graph: An LLM-based approach," *Electronics*, vol. 13, no. 19, p. 3936, Oct. 2024. [Online]. Available: https://doi.org/10.3390/electronics13193936

[16] K. L. Fox, K. R. Niewoehner, M. Rahmes, J. Wong, and R. Razdan, "Leverage large language models for enhanced aviation safety," in *Proc. 2024 Integrated Communications, Navigation and Surveillance Conf. (ICNS)*, Herndon, VA, USA, Apr. 2024, doi: 10.1109/ICNS60906.2024.10550651.

[17] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL*, 2019.

[18] A. Radford et al., "Language Models are Unsupervised Multitask Learners," *OpenAI*, 2019.

[19] C. Raffel et al., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” *JMLR*, vol. 21, 2020.
[20] J. Zhang et al., “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” in *Proc. ICML*, 2020.
[21] S. Andrade et al., “SafeAeroBERT: Towards a Safety-Informed Language Model for Aviation,” *Aerospace*, 2021.
[22] S. Kierszbaum et al., “ASRS-CMFS vs. RoBERTa: Comparing Pre-Trained Language Models for Anomaly Prediction in Aviation Reports,” *Aerospace*, vol. 9, 2022.
[23] M. F. Hilman, R. Passarella, D. Kurniawan, and S. Sutarno, ”Automatic Text Summarization on Aviation Traffic Accident Report Synopsis Using Term Frequency-Inverse Document Frequency (TF-IDF) Algorithm,” Mar. 14, 2024. [Online]. Available: https://ssrn.com/abstract=4758981 or http://dx.doi.org/10.2139/ssrn.4758981
[24] M. Xiong, H. Wang, Y. D. Wong, and Z. Hou, ”Enhancing aviation safety and mitigating accidents: A study on aviation safety hazard identification,” *Advances in Engineering Informatics*, vol. 2024, p. 102732, 2024. [Online]. Available: https://doi.org/10.1016/j.aei.2024.102732.
[25] W. Pan, B. Han, and P. Jiang, ”Study on the standardization method of radiotelephony communication in low-altitude airspace based on BART,” *Frontiers in Neurorobotics*, vol. 19, 2025, Art. no. 1482327, Apr. 2, 2025. [Online]. Available: https://doi.org/10.3389/fnbot.2025.1482327.