Automatic Job Safety Report Generation using RAG-based LLMs

Mario Luca Bernardi
Department of Engineering
University of Sannio
Benevento, Italy
bernardi@unisannio.it

Marta Cimitile
Department of Law and Digital Society
Unitelma Sapienza University
Rome, Italy
marta.cimitile@unitelmasapienza.it

Riccardo Pecori
SMARTEST Research Centre & Institute of Materials for Electronics and Magnetism
eCampus University & National Research Council of Italy
Novedrate (CO) & Parma, Italy
riccardo.pecori@uniecampus.it

2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10651320

Abstract—**This study introduces an innovative approach to safety report generation using a Retrieval-Augmented Generation (RAG) framework, tailored to synthesize comprehensive reports from descriptions and logs of work sessions. The core contribution of our study is the comparison and optimization of various Large Language Model variants (based on LLaMA) and embedding models, aiming to identify the most effective combination for accurately capturing and reflecting the intricacies of safety-related data in a given domain. Our RAG-based system leverages the strengths of different LLaMA models and embedding techniques to process and contextualize the input data, which include detailed session descriptions and operational logs. By integrating these models, we aim to automate the generation of safety reports that are not only coherent and contextually relevant, but also adhere to the stringent requirements of safety documentation in professional environments. The validation of our approach is performed using an aviation safety dataset and classic metrics in the field, such as Recall@5, GLEU, METEOR, and BERTscore. Our findings demonstrate the potential of RAG-based systems in streamlining the process of safety report generation, offering significant improvements in efficiency and accuracy over traditional methods and non domain-specific tailored models.**

Index Terms—**Job safety, LLMs, RAG, reporting, decision support systems, risk assessment**

I. Introduction

In modern times, safety regulations require continuous monitoring of the workers' activities and their workplace, with the aim of improving the employees' production and ensuring their safety. Indeed, some studies show the positive impact of the continuous monitoring of occupational safety on the health of workers [16]. In this context, human-centered Artificial Intelligence (AI) represents a new challenge for the organizations that research new tools and solutions to automate safety management procedures and perform accident prevention [9]. According to this, several studies have been conducted, in the last years, to evaluate AI-assistance systems based upon sensors [9] and of AI-based approaches for the analysis and generation of incident reports in different industrial domains (e.g., agriculture, aviation, medicine, construction, and railroad industry) [4], [9]. More recently, Large Language Models (LLMs) seem to enhance the efficiency of safety analyses and reduce the time required to process incident reports [4]. LLMs, trained on large datasets, can perform several language-related tasks without any specialized fine-tuning. This makes them potentially useful for the automatic analysis of safety reports, for the identification of patterns and anomalies, and for the detection of safety issues, risks, and strategies (also on the base of historical data). Finally, LLMs can be useful to complement and sometimes replace human expertise. They can also perform training activities and safety tests, since they can simulate several scenarios or generate synthetic data. Despite the great potentialities of LLMs in safety management, very few studies propose and evaluate these models in real scenarios.

This paper proposes an implementation and evaluation of a customizable LLM-based architecture tailored for job safety management and risk assessment, combining a Retrieval-Augmented Generation (RAG) approach with LLaMA Large Language Models (LLM) [25] and leveraging domain-specific embedding models to feed the retrieval stage. The proposed architecture is aimed at:

- generating automatically safety reports from the accidents descriptions;
- comparing the faithfulness of the generated safety reports to human-written safety reports;
- identifying the root causes contributing to the accidents themselves in order to perform a proper risk assessment.

The proposed LLM-based architecture is the kernel of a preliminary prototype of a virtual assistant, preparatory and fundamental for the objectives of the project co-funding the present paper (SWILSS project

Prof. Riccardo Pecori is a member of the INdAM GNCS research group. This work has been done to support the preliminary phase of the SWILSS project, CUP: D83C22002220005, in the framework of the BRIC INAIL 2022 project call.

in the BRIC INAIL 2022 call), which aims at answering the following research questions, investigated in the remainder of this work:

RQ1: *Is RAG-based LLaMA effective in generating security reports from work session logs, and which variant of RAG-based LLaMA provides the optimal performance for this task?*

RQ2: *How do domain-specific embedding models compare to generalist models in the context of a retrieval-augmented generation (RAG) framework, and what is their impact on the overall performance?*

RQ3: *To what extent does a RAG-based LLaMA model accurately identify the root causes of accidents and risky situations in its responses?*

The empirical validation of the proposed approach is preliminary performed on a publicly available dataset, called Aviation Safety Reporting System (ASRS) dataset[1], used also in [24], because of the scarcity of injury data in the hybrid working context, which would be the proper scenario of the project co-funding this paper.

The rest of the document is structured as follows: Section II describes the related work on LLMs applied to job safety management, while Section III concerns the background concepts about LLMs, with more particulars about LLaMA and RAG. Section IV describes the used approach for the automatic generation of safety reports in a detailed manner. Section V details the empirical validation setting, encompassing the exploited dataset and the considered metrics, while the obtained results and their discussion are reported in Section VI. Section VII discusses the threats to the validity of the empirical validation we performed, while Section VIII seals up the paper with some conclusions of the study and future research directions.

II. Related Work

In this section, we summarize some of the most recent works about safety in LLMs in general and then on the specific topic of job safety.

Safety in LLMs is currently mostly tackled from the ethical and responsibility point of view [26], with several recent pre-prints [1], [2], [13], [20], [28] dealing with either the analysis of potential harmful and impolite answers or suggestions provided by the LLMs themselves or the possible leakage of personal data in providing answers.

As specifically regards LLMs usage in the job safety context, very few researches have been conducted till now. In an editorial from 2023 [22], Shutske focuses on safety in a particular scenario, i.e., agriculture, by analyzing the usage of LLMs, specifically ChatGPT, to answer safety and health questions in five use cases. The questions regard the main hazards in a livestock or dairy farm, safety-related pieces of advice for elderly people who have just bought a farm or a tractor, healthcare suggestions for farm workers after specific health conditions, e.g., lung transplant, etc. The paper underlines that ChatGPT 3.5 and 4.0 only considers training data collected till late 2021, but also that there exist free plugins to add specific information, also in the form of audio or video files, to them. Moreover, it pinpoints that this specific information can be used to feed the LLM provided that it is publicly available and there are no copyright infringements. Finally, the author highlights the risks of misleading, inaccurate, as well as biased output from the LLM itself.

Another paper dealing with job safety can be found in [24]. It deals with aviation safety and exploits the same dataset (ASRS dataset) we used in the experiments carried out in our paper. The authors perform different tasks by using ChatGPT in order to i) create incident synopses from narratives, ii) identify different human factors and the entities involved in incidents, obtaining not so good results, with an average weighted precision of only 0.61, and iii) execute an evaluation of accountability by providing explanatory rationales for the LLMs' decisions. The synopses, provided by human analysts and ChatGPT, have been evaluated by comparing the cosine similarity values using three different LLM embeddings, while the human factor issues, detected by human beings and ChatGPT, have been compared with the help of a normalized confusion matrix. Finally, the paper also reports the human entities involved in the accidents, together with the rationale provided by ChatGPT for associating a certain event to a particular entity. Even though we took inspiration from the paper in [24], this is only a first study to assess the applicability of LLMs to the job safety scenario, while our paper is more extensive, comparing, on the same dataset, different and customizable LLMs, not based on GPT, and obtaining better results, thanks also to the application of RAG.

III. Background

A. Large Language Models and RAG

Large Language Models (LLMs) are deep neural networks exploiting several data in order to understand patterns and relationships in natural language corpora of text [18]. In this way, LLMs can mirror very well human beings in performing language-related tasks, such as language translation, summarizing, and question answering, with high accuracy [11]. However, despite the grammatical correctness of the texts generated by LLMs, they often do not result to be extremely accurate or suitable for a certain particular context [18], [7]. This limit of LLMs can be due to biases in the input data or in the fact that they are not correctly updated [10]. As a matter of fact, given that LLMs need many input data

[1] https://huggingface.co/datasets/archanatikayatray/ASRS-ChatGPT

for their training, the quality of their output depend very much on the input data themselves; therefore, biases or mistakes in the training set can undergo enormous amplification in flowing through the LLMs [6]. Retrieval-Augmented Generation (RAG) [14] can be regarded as a promising solution for the aforementioned issues, mainly for knowledge-intensive tasks, and it is getting more and more popular in systems using LLMs. The rationale on the basis of the RAG approach entails the merging of specialized and dynamic external repositories with the original knowledge base on which the LLM has been trained [14]. This is done by using In-Context Learning (ICL) [5], which allows the external retrieval of relevant information using proper search algorithms then returned to the LLM that furnishes further context-related data [6]. The exploited external repository is usually enhanced on the basis of the needs of certain specific domains and can be further updated with novel knowledge leading to an increase in the precision of the model output [6]. RAG usually encompasses both a retriever and a generator [14] and it is normally made up of three main steps, i.e., retrieve, augment, and generate. In the first one, a user query x is used to retrieve relevant context text documents (z) from an external knowledge base with parameters η. Then, the query is embedded into a vector space, exploiting a proper embedding model, and included as additional context in the vector database. On the basis of the similarities between vectors and query, the k closest documents from the vector database are extracted. In the second step, a combination between the initial query and the additional context takes place into a prompt template. In the generate step, the LLM is fed through the retrieval-augmented prompt and then it can generate an answer to a question according to the contextual data coming from the retrieved chunks.

B. LLaMA

LLaMA [25] is a set of founding LLMs, developed by Meta, which features from 7 billions to 70 billions parameters. These LLMs are trained on a very big set of tokens, coming from public datasets, and they can achieve higher performance compared with other similar state-of-the-art models. Considering as an example the 70-billion-parameter model, a recent work [25] demonstrates that it is very competitive with Chinchilla [8] or PaLM-540B [3], some of the best LLMs. The training phase exploits several transformers and a standard optimizer; moreover, its stability is increased using a proper normalization of the input of each transformer sub-layer. The training sets are very heterogeneous, spanning several domains and including all publicly available datasets employed for LLM training. The speed of training is also enhanced by means of the causal multi-head attention mechanism, suitable for lessening memory and run-time use. Furthermore, the training speed is also increased by decreasing the activations, which are recomputed in the backward step with check-pointing. Other modifications [25], compared with traditional architectures, encompass the adoption of the SwiGLU activation function [21], of rotary positional embeddings [23], at each layer of the network, and of the Adam optimizer [15].

C. The embedding models

For Retrieval-Augmented Generation (RAG) approaches, especially in specialized contexts like workplace safety and aerospace, the effectiveness of an embedding model largely depends on its ability to capture domain-specific nuances and terminology. The best embedding models for such applications are typically those that are either pre-trained on a broad range of texts and then fine-tuned on domain-specific data, or models that have been specifically designed for technical and industry-specific contexts.

In this work, we experimented with the following embedding models:

- **Plain BERT and its variants (e.g., RoBERTa, ALBERT)**: BERT models, known for their deep understanding of context and language nuances, can be very effective, especially when fine-tuned on safety-related or aerospace-specific datasets. Variants like RoBERTa or ALBERT offer improvements in certain aspects like training efficiency or parameter count.
- **Domain-specific BERT model (SciBERT, AeroBERT)**: these are BERT variants that have been pre-trained on specific domains. The first one we want to test is **SciBERT** that has been trained on a large corpus of scientific texts and could be beneficial for technical and scientific fields like aerospace. However, in the realm of domain-specific applications like aerospace, custom BERT models like "AeroBERT" are developed by fine-tuning the original BERT or its variants on specialized datasets. This process involves training the model on text data that are highly representative of the aerospace domain, including technical manuals, research papers, safety regulations, and other relevant documents. The goal of such fine-tuning is to adapt the model to better understand and process the specialized vocabulary, jargon, and linguistic structures common in aerospace texts. Since "**AeroBERT**" is a custom model designed for aerospace applications, it is important to test its effectiveness as embedding generation model in the frame of our Retrieval-Augmented Generation (RAG) framework with respect to less domain-specific models.

IV. Automatic generation of safety reports

The proposed architecture is depicted in Figure 1 and is made of various steps. In the training step, the descriptions of incidents are input for an embedding model.

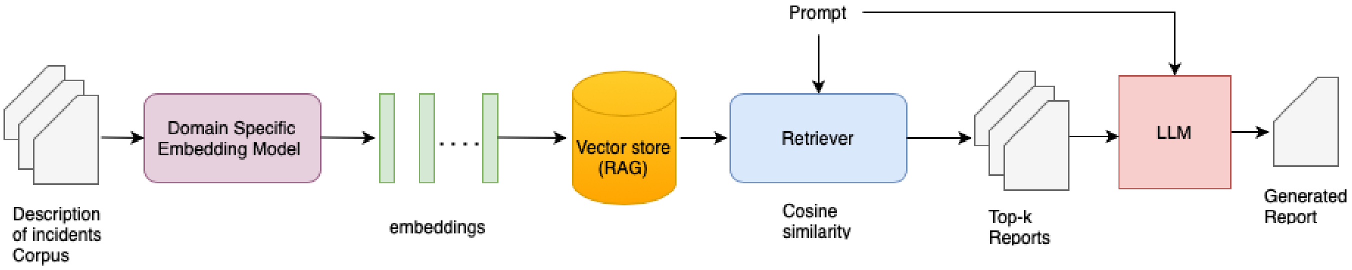


Fig. 1. The proposed RAG-based LLM architecture.

The proposed architecture is based on a domain specific model that allows one to generate the text embeddings according to the domain requirements. The generated embeddings are then sent to the RAG vector store. This RAG store is a collection of vectors that characterize the explored domain of interest according to the starting incident corpus. Afterwards, the embeddings are sent to the retriever component that also receives the user prompt. The retriever computes the cosine similarity between the user prompt and the vectors generating the top k-reports on the base of the highest similarity scores. The top k-reports are used by the LLM as context information useful to enrich the generative model and ensure a more accurate generated report as an answer to the user prompt which it received as input as well. Summarizing, the standard procedure for executing a question-answering task utilizing a localized knowledge base is as follows:

- **Creation of a vector database**: this involves generating embedding vectors for every document in the local knowledge base. These documents are then organized in a vector database, with the embedding vectors serving as indexing mechanism.
- **Context retrieval**: the question is embedded using the same technique used for the documents. This embedded question is then utilized to identify and retrieve the most pertinent documents, with respect to the adopted embedding model, from the vector database, providing the necessary context.
- **Feeding the LLM**: the selected relevant documents, along with the input question, are fed into the LLM. This integration of the question and the contextual information from the local knowledge base enables the LLM to generate a response that is specifically tailored to the information contained within the local knowledge base.

V. Experiment Description

Our experimental objective is to answer to the research questions highlighted in Section I. For all the embedding models described in Section III, we validated the performance of a RAG-based platform designed for generating safety reports in an aerospace context. We aimed to compare the efficacy of different models, namely, plain LLaMA, and LLaMA with different RAG embeddings (BERT, SciBERT, and AeroBERT). The RAG-based platform is set up in order to retrieve relevant information using the embeddings generated by the aforementioned models and then generate safety reports. The retrieval component is optimized to query a database of indexed safety reports, while the generation component is tasked with synthesizing this information into coherent and comprehensive safety reports.

A. Dataset

In this study, an extract of the ASRS dataset as proposed in [24] was adopted to test the aforementioned plain and RAG-based LLMs. The initial ASRS database includes incident reports covering the time period between January 2009 and July 2022. From this initial dataset, the incident records describing only incidents primarily provoked by human factors were selected. The obtained dataset is composed of $9,984$ records that contain a corresponding unique identifier, the incident narrative as described by a human reporter, the synopsis of the incident, and the human factor issues that contributed to the incident. Both the synopsis and the factors are written by a human safety analyst.

B. Experimental setting and validation metrics

The metrics we considered to evaluate the proposed RAG-based LLM architecture on the considered job safety dataset are some of those commonly employed in natural language processing. They are the following:

- **Recall@5**, which is the proportion of the test samples on which five candidate documents contain the ground truth;
- **GLEU** [17], which is a variant of BLEU [19] in order to evaluate syntactic error correction through an n-gram overlap with some reference sentences;
- **METEOR** [12], which means Metric for Evaluation of Translation with Explicit ORdering and is exploited to evaluate machine translations by addressing some issues in BLEU, but it can be also used for the semantic and syntactic correspondence with a reference text;
- **BERTscore** [27], which is employed to evaluate automated text simplification by comparing predicted

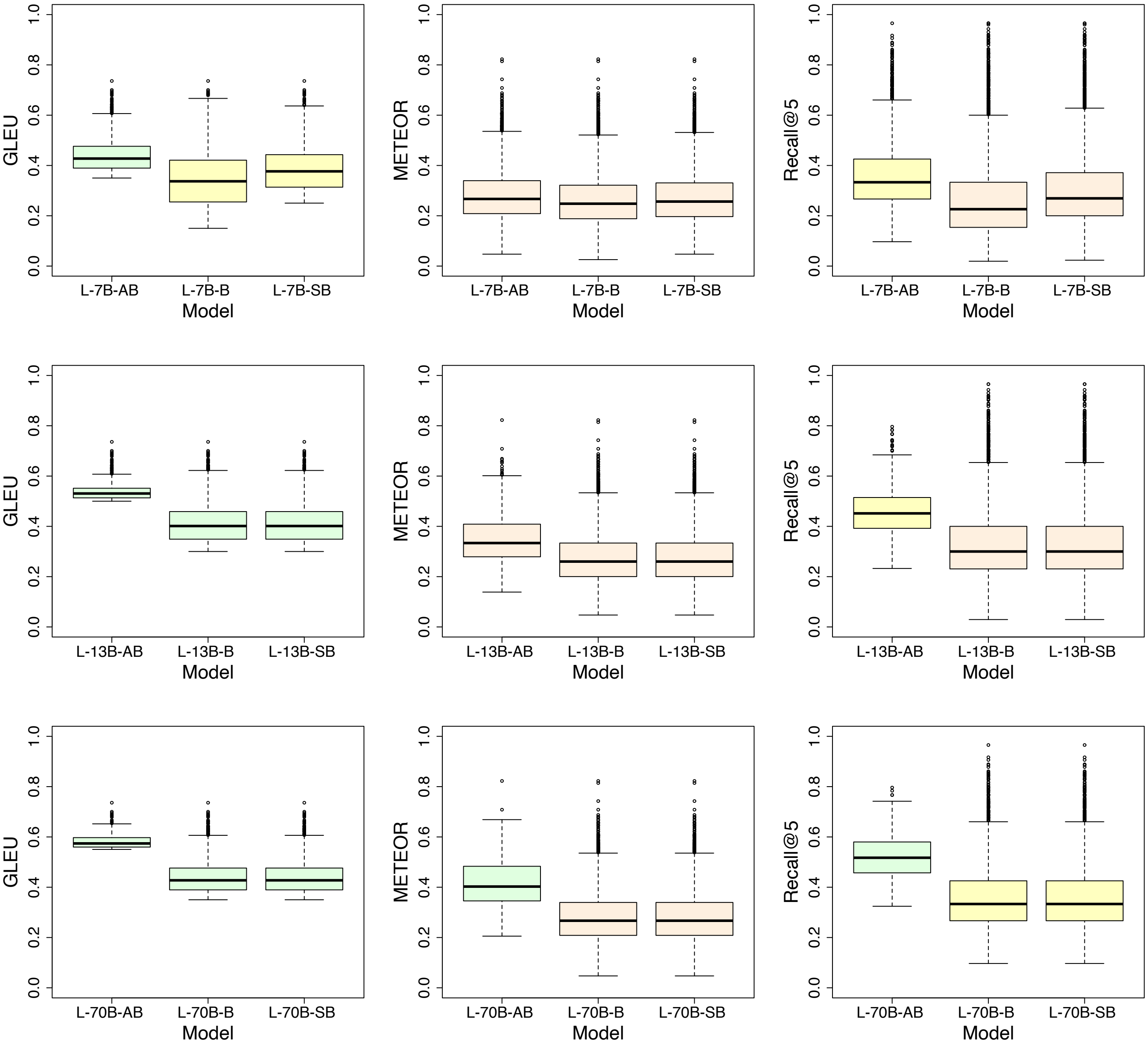


Fig. 2. Distribution of text quality metrics for the generated reports using the considered LLaMa models.

documents with both ground truth and old documents.

In addition, an analysis was conducted to compare the identification of the root causes of the incidents by LLaMA with those determined by human safety analysts. The analysis took into account the frequency with which LLaMA linked specific root causes to accidents logs. For this aspect, a confusion matrix was constructed to demonstrate the degree of agreement between LLaMA and the human safety analysts in attributing root causes to safety incidents. Each row of the matrix corresponds to root causes of accidents as identified by the analysts. Moreover, the analysis employed also standard performance metrics, including precision, recall, and F1 score, to evaluate the accuracy of LLaMA's identifications compared to those made by human experts.

VI. Results and Discussion

This section presents and discusses the main results of the study, both quantitative and qualitative, with the experimental findings for each research question reported in detail.

TABLE I
COMPARISON OF THE DIFFERENT MODELS USING BERTSCORE.

Model	$\mathbf{P_{BERT}}$	$\mathbf{R_{BERT}}$	$\mathbf{F1_{BERT}}$
LLaMA-7B-BERT	0.879	0.871	0.875
LLaMA-7B-SciBERT	0.879	0.876	0.877
LLaMA-7B-AeroBERT	0.879	0.882	0.880
LLaMA-13B-BERT	0.879	0.879	0.879
LLaMA-13B-SciBERT	0.879	0.879	0.879
LLaMA-13B-AeroBERT	0.889	0.894	0.892
LLaMA-70B-BERT	0.879	0.882	0.880
LLaMA-70B-SciBERT	0.879	0.882	0.880
LLaMA-70B-AeroBERT	**0.900**	**0.905**	**0.902**

A. Quantitative analysis

RQ1: *Is RAG-based LLaMA effective in generating security reports from work session logs, and which variant of RAG-based LLaMA provides the optimal performance for this task?*
This question aims at investigating the effectiveness of various LLaMA variants in generating security reports starting from the initial working session logs. To this aim, the distribution of the GLEU, METEOR, and Recall@5 for the evaluated LLaMA models combined with the adopted embedding models have been evaluated. The box-plots of these distributions are shown in Figure 2. The figure shows for all the combinations of the LLM (i.e., LLaMA of different sizes - 7B, 13B, 70B) and of the embeddings model (i.e., BERT as B , SciBERT as SB and AeroBERT as AB) a good distribution (according to the ranges provided in [17]) of the GLEU metric since it is always more than 0.3. Similar considerations can be made for the METEOR and Recall@5 metrics that are always more to 0.2.

RQ2: *How do domain-specific embedding models compare to generalist models in the context of a retrieval-augmented generation (RAG) framework, and what is their impact on the overall performance?*
This question explores the difference between generalist models and domain-specific models. Looking at the first column of figure 2, i.e., the GLEU metric, the best distribution is obtained by the combination of LLAMA-70B with AeroBERT embedding model and in general we can notice that the adoption of a domain specific model improves the performance of the overall report generation. This consideration can be also extended to the Recall@5 distribution, the third column in Figure 2. As for the METEOR metric, the distributions are lower and more similar among each other. However, the lower performance is due to the fact that METEOR is a lexical based metric and it is not able to grasp the semantics and the mining of the text.

In Table I, we show precision, recall, and f-measure of all considered models, with 7, 13, and 70 billions (B) of parameters. As one can see the best performing model is LLaMA-70B-AeroBERT, even if the performance of LLaMA-13B-AeroBERT are not so lower and its complexity is for sure much less.

RQ3: *To what extent does a RAG-based LLaMA model accurately identify the root causes of accidents and risky situations in its responses?*
This question aims to explore the capability of the considered LLaMA models to accurately identify the root causes of the accidents in their responses. Starting from the initial dataset we extracted the root causes for failures. Subsequently, we verified if the root causes of failures were included in the generated report of our model. The confusion matrix for the root causes is generated using the best model (LLaMA-70B with AeroBERT) in the columns with respect to the ground truth in the rows and it is reported in Table II. The matrix shows good capability of the model to extract the root causes of the accidents, given that the matrix resembles very much a diagonal matrix.

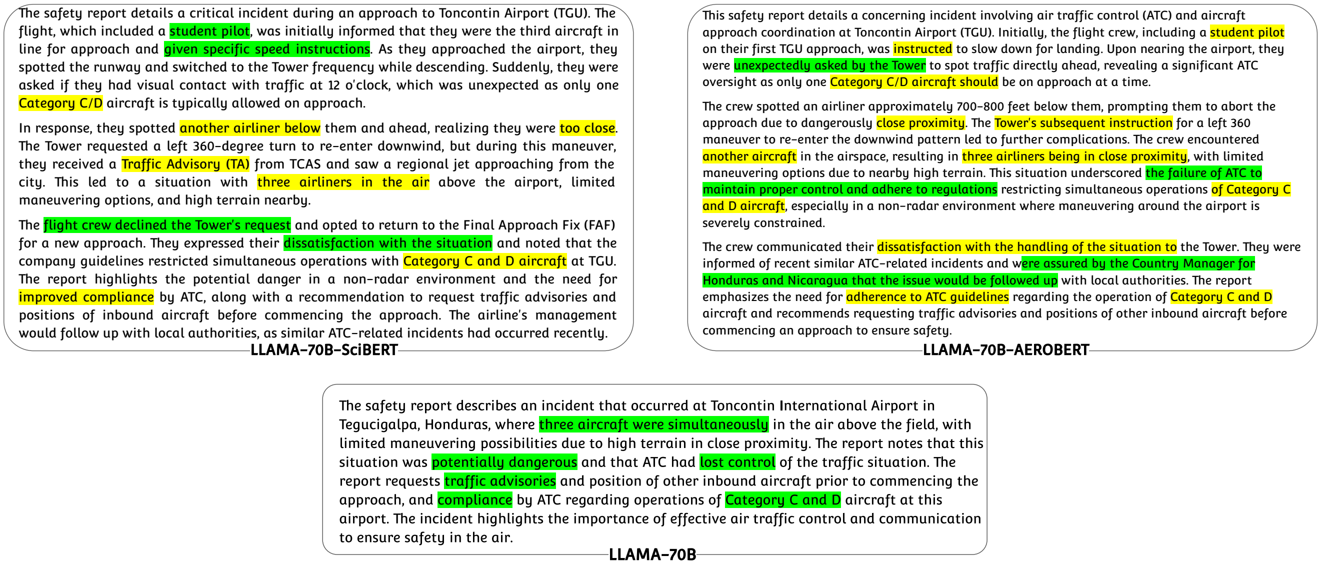
B. Qualitative analysis

In this subsection, we performed a qualitative analysis of three different reports, shown in Figure 3: one generated by plain LLaMA-70B, without RAG pipeline, and the two versions obtained adding a RAG using SciBERT and AeroBERT. The original narrative describes an incident during an approach to the Toncontin Airport involving a student pilot and two other aircrafts, highlighting concerns with the Air Traffic Control (ATC) management and the importance of the compliance with operational guidelines for Category C and D aircrafts in a non-radar environment. We can highlight the following qualitative aspects:

- **Plain LLaMA-70B**: this synthesis focuses on the critical elements of the incident, emphasizing the dangerous situation created by the simultaneous presence of three aircrafts in a restricted airspace. It succinctly captures the key points, i.e., i) the potential danger, ii) the loss of control by ATC, and iii) the need for traffic advisories and compliance with Category C and D aircraft operations. It is concise and highlights the safety concerns effectively.
- **LLaMA-70B-SciBERT**: this version provides a more detailed account of the event, including: i) the involvement of a student pilot, ii) specific instructions given to the flight crew, and iii) the sequence of events leading to the decision to abort the approach. It also mentions the crew's dissatisfaction and the broader context of the company's guidelines as well as the need for improved ATC compliance. It offers a more comprehensive understanding of the incident, including the procedural context and the response of the flight crew.
- **LLaMA-70B-AeroBERT**: this synthesis also gives a detailed account, similar to LLaMA-70B-SciBERT, but with a slightly different emphasis on the ATC

TABLE II
CONFUSION MATRIX FOR ROOT CAUSES OF ACCIDENTS WITH RESPECT TO HUMAN-BASED GROUND TRUTH.

				LLaMa 70B aeroBERT Generated								
		Accidents Root Causes		1	2	3	4	5	6	7	8	9
Ground Truth	System or Component Error	Pressurization System Error	1	0.73	0.00	0.02	0.02	0.01	0.07	0.00	0.02	0.01
		Autopilot Error	2	0.01	0.88	0.02	0.00	0.00	0.02	0.03	0.00	0.00
		Communication Error	3	0.10	0.05	0.69	0.02	0.00	0.18	0.00	0.17	0.06
		GPS System Malfunctioning	4	0.01	0.01	0.01	0.87	0.02	0.02	0.06	0.04	0.00
		Oxygen masks problems	5	0.12	0.00	0.00	0.00	0.61	0.02	0.01	0.00	0.00
		Malfunctioning Landing Gear	6	0.00	0.00	0.00	0.01	0.00	0.53	0.02	0.01	0.03
		Hydraulic Pump Error	7	0.00	0.03	0.03	0.05	0.08	0.03	0.70	0.05	0.00
		Software Errors	8	0.01	0.01	0.20	0.01	0.03	0.01	0.04	0.68	0.01
		Electrical/Mechanical Error	9	0.02	0.02	0.03	0.02	0.25	0.12	0.14	0.03	0.89



The safety report details a critical incident during an approach to Toncontin Airport (TGU). The flight, which included a student pilot, was initially informed that they were the third aircraft in line for approach and given specific speed instructions. As they approached the airport, they spotted the runway and switched to the Tower frequency while descending. Suddenly, they were asked if they had visual contact with traffic at 12 o'clock, which was unexpected as only one Category C/D aircraft is typically allowed on approach.

In response, they spotted another airliner below them and ahead, realizing they were too close. The Tower requested a left 360-degree turn to re-enter downwind, but during this maneuver, they received a Traffic Advisory (TA) from TCAS and saw a regional jet approaching from the city. This led to a situation with three airliners in the air above the airport, limited maneuvering options, and high terrain nearby.

The flight crew declined the Tower's request and opted to return to the Final Approach Fix (FAF) for a new approach. They expressed their dissatisfaction with the situation and noted that the company guidelines restricted simultaneous operations with Category C and D aircraft at TGU. The report highlights the potential danger in a non-radar environment and the need for improved compliance by ATC, along with a recommendation to request traffic advisories and positions of inbound aircraft before commencing the approach. The airline's management would follow up with local authorities, as similar ATC-related incidents had occurred recently.

LLAMA-70B-SciBERT

This safety report details a concerning incident involving air traffic control (ATC) and aircraft approach coordination at Toncontin Airport (TGU). Initially, the flight crew, including a student pilot on their first TGU approach, was instructed to slow down for landing. Upon nearing the airport, they were unexpectedly asked by the Tower to spot traffic directly ahead, revealing a significant ATC oversight as only one Category C/D aircraft should be on approach at a time.

The crew spotted an airliner approximately 700-800 feet below them, prompting them to abort the approach due to dangerously close proximity. The Tower's subsequent instruction for a left 360 maneuver to re-enter the downwind pattern led to further complications. The crew encountered another aircraft in the airspace, resulting in three airliners being in close proximity, with limited maneuvering options due to nearby high terrain. This situation underscored the failure of ATC to maintain proper control and adhere to regulations restricting simultaneous operations of Category C and D aircraft, especially in a non-radar environment where maneuvering around the airport is severely constrained.

The crew communicated their dissatisfaction with the handling of the situation to the Tower. They were informed of recent similar ATC-related incidents and were assured by the Country Manager for Honduras and Nicaragua that the issue would be followed up with local authorities. The report emphasizes the need for adherence to ATC guidelines regarding the operation of Category C and D aircraft and recommends requesting traffic advisories and positions of other inbound aircraft before commencing an approach to ensure safety.

LLAMA-70B-AEROBERT

The safety report describes an incident that occurred at Toncontin International Airport in Tegucigalpa, Honduras, where three aircraft were simultaneously in the air above the field, with limited maneuvering possibilities due to high terrain in close proximity. The report notes that this situation was potentially dangerous and that ATC had lost control of the traffic situation. The report requests traffic advisories and position of other inbound aircraft prior to commencing the approach, and compliance by ATC regarding operations of Category C and D aircraft at this airport. The incident highlights the importance of effective air traffic control and communication to ensure safety in the air.

LLAMA-70B

Fig. 3. Comparison of a safety report as generated by the considered LLaMA models. The salient parts added by each model are highlighted in green while those found also by other models are in yellow.

failure to maintain control and adhere to regulations. It highlights the unexpected instruction from the Control Tower and the subsequent complications that led to the dangerous situation. This version effectively communicates the severity of the ATC oversight and the crew's response, including the follow-up actions taken by the airline's management.

Qualitatively, we can say that the outcome of the plain LLM is very short and does not provide a clear and complete insight in what really happened. Indeed, LLaMA-70B-SciBERT is sufficiently detailed to clearly grasp all notch details of what took place; while LLaMA-70B-AeroBERT is more detailed, but some not essential information could be omitted without loosing the main concepts. From this perspective, and considered both the performance shown in Table I and the complexity of the models, we can state that even if LLaMA-70B-AeroBERT can lead to some more fractions of performance, RAG-based models with lower complexity can achieve, on the considered dataset, a very good trade-off between quantitative performance, complexity of the model, and qualitative output.

VII. THREATS TO THE VALIDITY

In this section, we discuss some threats to the validity related to the proposed study.

A first threat regards the adoption of open-source LLM models even if, currently, closed-source models are prevalent. However, the adoption of close-source models, such as ChatGPT, is not possible in a research context given the high costs and the lack of control over the evolution of the models themselves.

The rapid evolution of LLM models also represents an internal threat to the validity. The rapid changes that are characterizing LLMs can potentially make the obtained results obsolete in a short time. In order to mitigate this threat, we have compared in this paper various recent LLM models.

As regards the external validity threats, a critical aspect regards the hard repeatability of the proposed experiments. This is due to the fact that the same LLMs sometimes exhibit different answers to the same prompt. To mitigate this threat, we have used an archiving system, that is available to the external parties on request, to reproduce the proposed experiments. Finally, another threat to external validity, concerns the generalizability of the described results. Indeed, in this study, we have focused on a specific dataset, regarding the aviation context, and we are aware that the findings described in this study could be only consolidated and generalized in our future work.

VIII. Conclusions

In this paper, we evaluated the effectiveness of different LLaMA models to perform automatic generation of safety reports. Various LLaMA variants have been assessed and compared to optimize the performance of the evaluated models. Moreover, we compared, in the context of a RAG framework, the performance of plain models with respect to embedding models. Finally, an evaluation of the capability of LLaMA models to identify the root causes of accidents in their responses was performed. The experiments, conducted on the ASRS dataset show that the RAG-based LLaMA models provide good performance in generating security reports obtaining optimal results when the LLaMA-70B combined with AeroBERT is used. This highlights that the adoption of RAG-based domain-specific models improves the overall performance. Finally, the considered LLaMA models show very good capability to include the root causes of accidents in the generated security reports. In future work, new experiments will be conducted including new datasets on different job safety domains, by considering more validation metrics, and additional LLM models, also not using LLaMA, to find out if one is better than all others in different job safety contexts or each context requires a particular LLM model to get optimal outcomes.

References

[1] Rishabh Bhardwaj and Soujanya Poria. Red-teaming large language models using chain of utterances for safety-alignment, 2023.

[2] Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions, 2023.

[3] Aakanksha Chowdhery, Sharan Narang, et al. Palm: Scaling language modeling with pathways, 2022.

[4] Katherine Darveau, Daniel Hannon, and Chad Foster. A comparison of rule-based and machine learning models for classification of human factors aviation safety event reports. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 64(1):129–133, 2020.

[5] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A survey on in-context learning, 2023.

[6] Yunfan Gao, Yun Xiong, et al. Retrieval-augmented generation for large language models: A survey, 2024.

[7] Philipp Hacker, Andreas Engel, and Marco Mauer. Regulating chatgpt and other large generative ai models, 2023.

[8] Jordan Hoffmann, Sebastian Borgeaud, et al. Training compute-optimal large language models, 2022.

[9] Jaroslava Huber, Michael Haslgrübler, Martin Schobesberger, Alois Ferscha, Viktorijo Malisa, and Georg Effenberger. Addressing worker safety and accident prevention with ai. In *Proceedings of the 11th International Conference on the Internet of Things*, IoT '21, page 150–157, New York, NY, USA, 2022. Association for Computing Machinery.

[10] Susmit Jha, Sumit Kumar Jha, et al. Dehallucinating Large Language Models Using Formal Methods Guided Iterative Prompting. In *2023 IEEE International Conference on Assured Autonomy (ICAA)*, pages 149–152, 2023.

[11] Enkelejda Kasneci, Kathrin Sessler, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

[12] Alon Lavie and Abhaya Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz, editors, *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[13] Sharon Levy, Emily Allaway, et al. Safetext: A benchmark for exploring physical safety in language models, 2022.

[14] Patrick Lewis, Ethan Perez, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[16] Phoebe V. Moore. Osh and the future of work: Benefits and risks of artificial intelligence tools in workplaces. In Vincent G. Duffy, editor, *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body and Motion*, pages 292–315, Cham, 2019. Springer International Publishing.

[17] Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel R. Tetreault. GLEU without tuning. *CoRR*, abs/1605.02592, 2016.

[18] Ipek Ozkaya. Application of Large Language Models to Software Engineering Tasks: Opportunities, Risks, and Implications. *IEEE Software*, 40(3):4–8, 2023.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[20] Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models, 2023.

[21] Noam Shazeer. Glu variants improve transformer, 2020.

[22] John M. Shutske. Editorial: Harnessing the power of large language models in agricultural safety &amp; health. *Journal of Agricultural Safety and Health*, 29(4):205–224, 2023.

[23] Jianlin Su, Murtadha Ahmed, et al. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[24] Archana Tikayat Ray, Anirudh Prabhakara Bhat, et al. Examining the potential of generative language models for aviation safety analysis: Case study and insights using the aviation safety reporting system (asrs). *Aerospace*, 10(9), 2023.

[25] Hugo Touvron, Thibaut Lavril, et al. Llama: Open and efficient foundation language models, 2023.

[26] Guohai Xu, Jiayi Liu, et al. Cvalues: Measuring the values of chinese large language models from safety to responsibility, 2023.

[27] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[28] Zhexin Zhang, Leqi Lei, et al. Safetybench: Evaluating the safety of large language models with multiple choice questions, 2023.