Improving Industrial Safety by Auto-Generating Case-specific Preventive Recommendations

Sangameshwar Patil, Sumit Koundanya, Shubham Kumbhar, Alok Kumar
TCS Research
{sangameshwar.patil,sumit.koundanya,shubham.kumbhar,k.alok9}@ tcs.com

Abstract

In this paper, we propose a novel application to improve industrial safety by generating preventive recommendations using LLMs. Using a dataset of 275 incidents representing 11 different incident types sampled from real-life OSHA incidents, we compare three different LLMs to evaluate the quality of preventive recommendations generated by them. We also show that LLMs are not a panacea for the preventive recommendation generation task. They have limitations and can produce responses that are incorrect or irrelevant. We found that about 65% of the output from Vicuna model was not acceptable at all at the basic readability and other sanity checks level. Mistral and Phi-3 are better than Vicuna, but not all of their recommendations are of similar quality. We find that for a given safety incident case, the generated recommendations can be categorized as specific, generic, or irrelevant. This helps us to better quantify and compare the performance of the models. This paper is among the initial and novel work for the preventive recommendation generation problem. We believe it will pave way for use of NLP to positively impact the industrial safety.

1 Introduction

In this paper, we propose a novel application of Natural Language Processing (NLP) to improve industrial safety and thereby take a step towards creating positive impact on the society in general. Industrial incidents refer to unplanned events or accidents that occur in industrial settings and frequently lead to injuries, property and material loss, and may also cause loss of life or environmental damage. Industrial accidents continue to be a major global concern. According to the International Labour Organization (ILO), there are millions of work-related deaths and injuries annually (ILO, 2023). Incidents can lead to direct costs such as medical expenses, equipment repairs, and legal fees. Further, they also entail indirect costs like lost productivity and the hidden costs of long term damage to environment (Jayapriyanka J, 2023) as well as reputation of an organization (e.g., the fallout of Boeing 737 MAX (Lampert and Ganapavaram, 2024)). The economic burden of occupational injuries and illnesses in the U.S. alone is estimated to be more than $100 billion annually[1]. These costs have remained high across different geographies (Tompa et al., 2021), and also observed over long duration (Leigh, 2011).

Preventing incidents not only saves invaluable lives and avoids injuries but also is more cost-effective than dealing with their consequences. By systematically identifying and controlling risks, organizations can better protect their assets, personnel, and operations from unforeseen events. Further, many industries are subject to strict regulatory requirements related to occupational health and safety[2]. Implementing preventive recommendations ensures compliance with these regulations. It also helps to avoid penalties and legal liabilities.

A critical component of an overall risk management strategy is to prevent recurrence of similar incidents (Patil et al., 2023). Industrial environments often involve complex processes, machinery, and technologies. Identifying potential hazards (Ramrakhiyani et al., 2021) and developing effective preventive measures requires in-depth knowledge of these operations and specialized expertise. Further, different industries and workplaces face diverse risks and hazards. Developing comprehensive preventive recommendations requires a thorough understanding of industry-specific risks, regulatory requirements, and best practices tailored to each environment. Acquiring and retaining such qualified personnel with the necessary skills can be costly for organizations. Hence, obtaining good preven-

[1] https://injuryfacts.nsc.org/work/costs/work-injury-costs/

[2] https://webapps.dol.gov/elaws/elg/osha.htm

Proceedings of the Third Workshop on NLP for Positive Impact, pages 349–353

tive recommendations for industrial safety can be challenging and costly.

NLP in general and Large Language Models (LLMs) in particular can be valuable tools for generating preventive recommendations for industrial safety. LLMs can analyze vast amounts of data quickly, including past incident reports, safety regulations, and industry best practices. This enables them to identify patterns and insights that might be missed even during manual reviews. LLMs can process and generate recommendations for large datasets of industrial incidents without the need for proportional increases in human resources. Thus, NLP tools can have a positive impact on the society by improving industrial safety.

Contributions: In this paper, we propose a novel application to improve industrial safety by generating preventive recommendations using LLMs. Using a dataset of 275 incidents representing 11 different incident types sampled from real-life OSHA incidents (Zhang et al., 2020), we compare three different LLMs to evaluate the quality of preventive recommendations generated by them. We also demonstrate that while large language models (LLMs) hold significant promise, they are far from being a comprehensive solution for generating preventive recommendations. Despite their capabilities, LLMs have inherent limitations and are prone to producing responses that may be inaccurate or off-topic. Their performance can be inconsistent and they can generate recommendations that do not always align with the intended goals or context. We show that for a given incident case, the generated recommendations may be (i) *specific* and directly relevant to the case, or (ii) *generic*, i.e., are useful for as a broad preventive measure which need not be focused on the current incident for which recommendation is sought from LLM, or (iii) the recommendations may be completely irrelevant for the current incident and it may even be hallucination by the LLM.

Rest of the paper is organized as follows. In Section 2, we first describe the details of the proposed study. Section 3 covers the experimental setup, results and analysis. In Section 4, we give a brief overview of the related work. Limitations and ethical considerations for generating preventive recommendations using LLMs are discussed in Section 5. Finally, we conclude in the Section 5.

Table 1: Sample Industrial Safety Incident

On October 28 2011, Employee #1 used a cutting torch on a 55-gallon sealed drum that had contained a combustible liquid and might have still contained some of that liquid. The drum exploded and Employee #1 was killed.

Table 2: Example of LLM generated safety recommendation for prevention of similar industrial incident

LLM	**Excerpt from safety recommendations generated**
Mistral	Ensure that all drums containing combustibles are properly labeled and that employees are aware of the contents before using torches or other open flames.
Phi-3	Proper storage and handling: Ensure that combustible liquids are stored in appropriate containers and stored in well-ventilated areas away from sources of ignition.

2 Proposed Approach

In this work, we use generative power of Large Language Model (LLM) to generate preventive recommendations for industrial incidents. Figure 1 shows the steps in the recommendation generation and their evaluation process. In this study, we examine three different LLMs: (a) Vicuna-13b-v1.5-16k [3], (b) Mistral-7B-Instruct-v0.2[4], (c) Phi-3-mini-4k-instruct[5]. We chose the Vicuna (Chiang et al., 2023), Mistral (Jiang et al., 2023), and Phi3 (Haider et al., 2024) models for this task because they are representative of recently released open source models. They all give us the important benefit of easy customization and are freely available to the community. They also require less computing power compared to larger or more expensive models. Further, their relatively smaller sizes allow us to experiment with moderate compute resources. The small sized models also are easier for integration in larger solution. All these factor make them an easy-to-use and cost-effective proposition.

We prompt each LLM with the incident report text in special delimited format (triple quotes used as the delimiter) and ask it to generate the preventive recommendations to avoid recurrence of

[3] https://huggingface.co/lmsys/vicuna-13b-v1.5-16k

[4] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[5] https://huggingface.co/microsoft/Phi-3-mini-4k-instruct

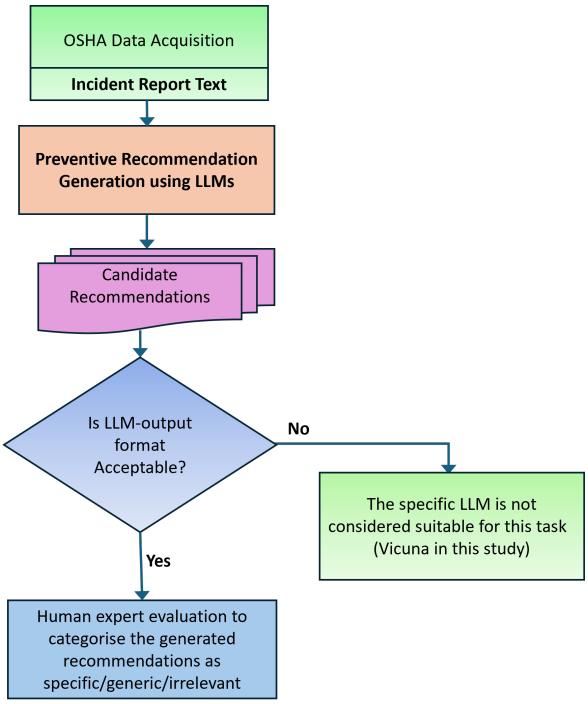

Figure 1: Flowchart of preventive recommendations generation and their evaluation for industrial incidents

similar incidents in future. Table 1 shows an excerpt from a real-life industrial incident report and Table 2 shows excerpts from the preventive recommendations generated using two LLMs.

We also tweaked the parameters to improve the quality of output from the LLMs as follows: (i) $max\_new\_tokens = 2000$: The maximum number of tokens to generate, ignoring the number of tokens in the prompt. (ii) $no\_repeat\_ngram\_size = 3$: This parameter helped to prevent the model from generating repetitive sequences by restricting the repetition of n-grams in the generated text. (iii) $temperature = \{0.0, 0.3, 0.7\}$: The temperature parameter of an LLM helped to regulate the amount of randomness or diversity in the output. Lower temperature helped to reduce the hallucination.

The generated recommendations were evaluated in two stages. First, we check and quantify the acceptability of the generated text. For this purpose, we checked the basic criteria such as: (i) the output is readable by a human, e.g., the words in a sentence should be properly separated by white-space and punctuation etc. The sentences are clearly separated using period or punctuation or bullet points so that further automated analysis at sentence level is possible. (ii) There is no verbatim copy of the input incident text, (iii) to check if the generated text contains hallucinations, i.e., text which is completely unrelated to the input incident report. For this purpose, we sampled output from each LLM for about 20% of the incidents and checked it with respect to the above mentioned basic sanity check criteria. At the end of this stage, we found that about 65% of the output from Vicuna model was not acceptable. Hence, Vicuna was eliminated from further evaluation.

For the remaining models, viz. Mistral and Phi-3, we observed that not all recommendations of same quality. In particular, we noted that some of the recommendations are very specific and directly useful for the given incident. Other recommendations were general suggestions and a few were irrelevant. Hence, in the second stage of evaluation, we soliciated help from human experts to categorize the generated recommendations into one of the following three categories, viz. $\{specific, generic, irrelevant\}$. This categorization helps us to quantitatively benchmark the utility of the LLMs for the task of preventive recommendation generation.

3 Experimental Evaluation

3.1 Dataset Overview

We collect a dataset of 1863 Occupational Safety and Health Administration (OSHA) incidents report summaries originally compiled by (Zhang et al., 2020). The dataset is distributed into 11 different classes, viz., *asphyxiation, caught in/between objects, collapse of object, electrocution, exposure to chemical substances, exposure to extreme temperatures, falls, fires and explosion, struck by moving objects, struck by falling object, traffic*. We sample 25 incident summaries from each class to generate a subset of 275 incident report summaries. This sampling strategy aims to achieve a balanced distribution across all classes, thereby ensuring that the dataset used for analysis and experimentation is representative and unbiased towards any specific class label. We perform our experiments and analysis on this subset.

3.2 Analysis

To evaluate whether the generated safety recommendations are conforming to basic readability as well as they are free from hallucination, we choose a sample of 55 incidents (20% of the actual data) and analyze the experimental outputs. We observe that only 19 out of 55 incident summaries are ac-

Table 3: Evaluation of model generated preventive recommendations by human experts

Model	Rec. Sentences	Specific	Generic	Irrelevant
Phi-3	2826	1395 (49.36%)	1370 (48.47%)	61 (2.16%)
Mistral	2427	1397 **(57.56%)**	933 (38.44%)	97 (3.99%)

ceptable in case of recommendations generated using Vicuna model. This roughly translates to 34.5%. Consequently, 65.5% of the recommendations generated by Vicuna model are not acceptable due to poor readability, formatting issues, hallucination and vague output.

We observed that Mistral_v0.2 and Phi-3 models fare much better than Vicuna. Most of the recommendations generated by Mistral_v0.2 and Phi-3 do not face tokenization and other basic readability or formatting errors. We further categorize them into (i) *specific* and directly relevant to the case, or (ii) *generic*, (iii) *irrelevant*. Human annotators evaluated categorization of recommendation sentences are shown in Table 3. The human annotators have experience of working with the industrial safety data and half of them had real-life manufacturing industry experience as well. We note that 49.36% of recommendation generated using Phi-3 are *specific* to the incident text. Recommendations categorized into *specific* using Mistral_v0.2 are 57.56%. We conclude that Mistral_v0.2 is better than Phi-3 in terms of generating specific recommendations to incident text.

4 Related Work

This section describes the related work to prevent incidents occurring in the construction industry. Chinniah (2015) discuss about manual analysis of report to identify cause of the incident and suggests preventive actions based on the findings and on the literature. The work in (Leung et al., 2012) surveys 395 construction workers (CWs) and identifies different organizational stressors, personal and emotional stress, and safety behaviors using factor analysis to prevent injury incidents and enhance safety behaviors of CWs. Manual analysis of reports or surveys to identify cause of the incident is a cost intensive and time-consuming activity. To solve the issue, Cheng et al. (2013) study the cause of incident using data mining techniques but provides suggestions for a few specific cases. Nielsen et al. (2006) examine whether the introduction of an incident reporting scheme with feedback in two industrial plants had an effect on the number of major incidents. Kasirossafar and Shahbodaghlou (2015) talks about incidents prevention through design (PTD)/ Design for safety (DFS) concept. Such techniques require collaboration of all stakeholders, development of new design standards and regulations, and improved availability of PTD/DFS tools. More importantly, these technologies are country specific and may not be available in other countries. To the best of our knowledge there is no prior work on providing recommendations and categorizing them for a large-scale dataset such as Occupational Safety and Health Administration (OSHA) using large language models (LLMs).

5 Conclusion

Industrial incidents can disrupt operations and production schedules, causing downtime and delays. Preventive recommendations help maintain continuity by minimizing disruptions and ensuring that work can proceed safely and efficiently. In this work, we proposed to use LLMs to improve industrial safety. Based on the comparative analysis of different LLMs, we identified their strengths and weaknesses. We found that Vicuna model is not suitable for this task. Phi-3 and Mistral models perform much better than Vicuna. Even with these two models, only about half of the recommendations generated are specific to the particular safety incident. Other recommendations tend to be generic and a small fraction of the recommendations is irrelevant. We also highlight that the LLM technology still needs significant improvement for this task and the preventive recommendations from LLMs need to be reviewed by safety professionals before actual implementation to ensure they are valid, practical, and aligned with industry standards. As part of future work, we plan to explore how investigation of causes and generating questions to probe relevant temporal aspects (Bedi et al., 2021; Hingmire et al., 2020) can be used to improve the recommendations. Further, we plan to improve alignment of the models so that the hallucinations and irrelevant recommendations are reduced and the fraction of specific recommendations in the output generated is improved.

Limitations and Ethical considerations

Relying on LLMs for safety recommendations without proper oversight could lead to ethical concerns, especially if the recommendations result in unintended negative consequences. LLMs might provide general recommendations that may not account for unique aspects of a specific industrial setting, such as particular operational constraints or site-specific hazards. The nature of industrial risks can change rapidly due to new technologies, processes, or regulations. LLMs might not always be up-to-date with the latest developments unless regularly updated and fine-tuned. LLMs generate recommendations based on patterns in data, not actual expertise. They might produce recommendations that are technically correct but impractical or unsafe without expert validation. The effectiveness of LLM-generated recommendations heavily relies on the quality of the input data. Inaccurate, outdated, or incomplete data can lead to misleading or suboptimal recommendations.

References

Harsimran Bedi, Sangameshwar Patil, and Girish Palshikar. 2021. Temporal question generation from history text. In *Proc. of the 18th International Conference on Natural Language Processing (ICON)*.

Ching-Wu Cheng, Hong-Qing Yao, and Tsung-Chih Wu. 2013. Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. *Journal of Loss Prevention in the Process Industries*, 26(6).

Wei-Lin Chiang et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Yuvin Chinniah. 2015. Analysis and prevention of serious and fatal accidents related to moving parts of machinery. *Safety science*, 75.

Emman Haider et al. 2024. Phi-3 safety post-training: Aligning language models with a "break-fix" cycle. *Preprint*, arXiv:2407.13833.

Swapnil Hingmire, Nitin Ramrakhiyani, Avinash Kumar Singh, Sangameshwar Patil, Girish Palshikar, Pushpak Bhattacharyya, and Vasudeva Varma. 2020. Extracting message sequence charts from hindi narrative text. In *Proc. of the First Joint Workshop on Narrative Understanding, Storylines, and Events (in conjunction with ACL)*.

International Labour Organization ILO. 2023. A call for safer and healthier working environments. https://www.ilo.org/publications/call-safer-and-healthier-working-environments.

HBL Jayapriyanka J. 2023. Unreported industrial accidents: Hidden environmental consequences in india. https://www.thehindubusinessline.com/business-tech/unreported-industrial-accidents-hidden-environmental-consequences-in-india/article67352516.ece.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mohammad Kasirossafar and Farzad Shahbodaghlou. 2015. Construction design: Its role in incident prevention. *Professional Safety*, 60(08).

Allison Lampert and Abhijith Ganapavaram. 2024. Boeing's safety culture under fire at us senate hearings. https://www.reuters.com/business/aerospace-defense/boeings-safety-culture-spotlight-us-senate-hearings-2024-04-17/.

J Paul Leigh. 2011. Economic burden of occupational injury and illness in the united states. *The Milbank Quarterly*, 89(4).

Mei-yung Leung, Isabelle Yee Shan Chan, and Jingyu Yu. 2012. Preventing construction worker injury incidents through the management of personal stress and organizational stressors. *Accident Analysis & Prevention*, 48.

Kent J Nielsen, Ole Carstensen, and Kurt Rasmussen. 2006. The prevention of occupational injuries in two industrial plants using an incident reporting scheme. *Journal of safety research*, 37(5).

Sangameshwar Patil, Nitin Ramrakhiyani, Swapnil Hingmire, Alok Kumar, Girish K Palshikar, and Harsimran Bedi. 2023. Timeline as a knowledge representation for retrieving similar safety incidents from industrial repositories. In *Workshop on Knowledge Augmented Methods for Natural Language Processing, in conjunction with AAAI*.

Nitin Ramrakhiyani, Swapnil Hingmire, Sangameshwar Patil, Alok Kumar, and Girish Palshikar. 2021. Extracting events from industrial incident reports. In *Proc. of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (in conjunction with ACL)*.

Emile Tompa, Amirabbas Mofidi, Swenneke van den Heuvel, Thijmen van Bree, Frithjof Michaelsen, Young Jung, Lukas Porsch, and Martijn van Emmerik. 2021. Economic burden of work injuries and diseases: a framework and application in five european union countries. *BMC Public Health*, 21.

Jinyue Zhang, Lijun Zi, Yuexian Hou, Da Deng, Wenting Jiang, and Mingen Wang. 2020. A c-bilstm approach to classify construction accident reports. *Applied Sciences*, 10(17).