MEDINFO 2025 — Healthcare Smart × Medicine Deep
M.S. Househ et al. (Eds.)

doi:10.3233/SHTI250867

Evaluating Large Language Models for Analysing Safety Risks in Healthcare Incident Reports

Kerstin DENECKE[a,1] and Helmut PAULA[b]
[a] *Bern University of Applied Sciences, Bern, Switzerland*
[b] *Stiftung Patientensicherheit Schweiz, Zürich, Switzerland*
ORCiD ID: Kerstin Denecke http://orcid.org/0000-0001-6691-396X

Abstract. Incident reports provide a rich source for analysing safety risks in healthcare systems. To support the timely analysis and interpretation of incident reports, natural language processing (NLP) can be applied. The aim of this paper is to evaluate the potential of large language models (LLMs) in extracting the causes of incidents and identifying contributing factors from incident reports. As dataset, we considered 10,063 messages from CIRRNET®, the Swiss national database for critical incidents in healthcare. We applied the LLM Gemma-2 to extract events, causes and contributing factors and group them along themes. 100 event reports were assessed manually regarding quality of extraction. Events were extracted with 92% accuracy, causes with 84% and contributing factors with 72% accuracy. Extraction of contributing factors fails as the LLM hallucinates or interprets. We conclude that LLMs show potential in analysing incident reports and can improve the efficiency and consistency of incident analysis.

Keywords. Large language model, Critical incident reporting system, Text analysis, Natural language processing, Patient safety event reporting

1. Introduction

Patient safety event reporting is a critical process for improving patient safety and quality of care [1] by encouraging healthcare professionals to report errors or potential errors and incidents. Although reporting is usually voluntary, such systems have proven invaluable in understanding and improving patient safety as they help to identify root causes and inform corrective actions [2,3]. Despite their usefulness, current reporting systems have significant limitations [4]. They are unable to provide accurate information on the incidence of errors in healthcare settings and rely on manual analysis of reports. This manual approach leads to delays in addressing issues and risks compromising the neutrality of the analysis. Text mining and natural language processing (NLP) techniques have already demonstrated their value in analysing patient safety event narratives [5]. NLP has been applied for incident analysis specifically for the analysis of themes and sentiments to describe the content of incident reports [6]. Ong et al. used statistical text classifiers (e.g., Naïve Bayes and Support Vector Machines) to detect extreme risk events

[1] Corresponding Author: Kerstin Denecke, Bern University of Applied Sciences, Institute Patient-centered Digital Health, Quellgasse 21, Bern, Switzerland, E-Mail: kerstin.denecke@bfh.ch

in incident reports [7]. Wang et al. used binary classifier ensembles [8], while Tabaie et al. used NLP techniques (e.g. bag-of-words and semantically rich features) to categorise contributing factors in reported events [9]. Young et al. provided a review of NLP classification tasks related to incident reporting and adverse event analysis [10]. In recent years, Large Language Models (LLMs) have emerged as powerful tools in a wide range of healthcare applications [11]. The aims of this paper are to 1) evaluate the potential of LLMs in extracting the causes and identifying contributing factors from incident reports, and 2) evaluate the performance of an LLM that could be run by anyone, even with limited hardware resources.

2. Method

Our dataset consists of 10,063 messages from CIRRNET® [12], covering the period from 21 May 2006 to 23 September 2023. CIRRNET®, the Critical Incident Reporting & Reacting Network (https://patientensicherheit.ch/cirrnet/), serves as a supra-regional platform that centralises the networking of local incident reporting systems in Switzerland. Incident reports often contain sensitive data, so we decided to test a locally run LLM that could be run by anyone, even with limited hardware resources. To meet these requirements, the experiments were carried out on a MacBook Air with an Apple M2 chip and 24 GB of RAM. We used the Gemma-2 language model with 9 billion parameters, implemented in Python using the Ollama framework API (https://ollama.com). Gemma-2 was applied to the CIRRNET® reports with the aim of identifying causes of critical events, critical events, and contributing factors. With 'causes', we refer to the primary classification of the underlying reason for an event. 'Contributing factors' are secondary elements that increased the likelihood or severity of the incident. The following prompt was used together with the event description from the CIRRNET® database: *"You are a language model specialized in analysing cause-effect relationships in German-language reports from CIRS (Critical Incident Reporting Systems) in hospitals. Your task is to analyse the reports, identify events and their causes, and categorize them thematically. Present each relation in the following format: [Event category]; [Cause category]; [Contributing factors]. Answer in English. Avoid explanations or comments. Only use information from the text."*. Event and cause categories were aggregated into specific time periods: 2006-2009, 2010-2014, 2015-2019, 2020-2021 and 2022-2023. This segmentation was designed to distinguish the initial phase of CIRRNET® from its regular use, and to analyse the pandemic and post-pandemic periods separately. For 100 texts, one author manually validated the results. Each extracted event, cause and contributing factor was assigned one of the labels 'correct', 'incorrect', 'incomplete but correct' or 'information not provided in text'. A second author approved 20 of the 100 texts to calculate the inter-annotator agreement and ensure the quality of the validation.

3. Results

3.1. Quality of extraction

90% of the 100 manually validated results were labelled as correct, while 7% were classified as incorrect and 3% were considered incomplete. Extracted causes were

considered correct for 84% of the 100 test examples and incorrect for 6% of the examples. Causes were reported for three events although none were described, and causes were incomplete for seven events. The extraction of contributing factors was less accurate. Contributing factors were correctly extracted for only 72% of the events (n=100), while 17% were incorrectly identified. Contributing factors were reported for nine events, although none were described, and two were incomplete. Inter-annotator agreement was high for 20% of the assessed entities, with a Cohen's Kappa of 1 for events and causes, and 0.91 for contributing factors.

3.2. Analysis of causes, events, and contributing factors

Table 1 outlines the identified 10 types of events and their associated causes. The causes of medication errors are predominantly human factors such as misidentification, miscommunication, oversight and inattention. Patient falls are often caused by improper handling, inadequate supervision or monitoring, and unsteady gait. Environmental hazards, such as obstacles or unsafe conditions, also contribute significantly to the risk of falls. Surgical and procedural errors often result from miscommunication, unclear documentation or errors in information transfer. Patient identification errors result from human error, miscommunication, mislabelling and data entry errors. Equipment malfunctions are caused by improper handling, manufacturing defects, inadequate maintenance or incorrect setup. Inaccurate documentation is primarily related to human error, while communication errors result from information gaps or misinterpretation. Delays in diagnosis and treatment are typically due to miscommunication and lack of coordination. Inadequate infection control is due to non-adherence to protocols or inadequate training. Finally, patient transfer problems often result from miscommunication or incomplete information.

Table 1. Types of failure events (in bold) and their causes.

2006-2009	2010-2014	2015-2019	2020-2021	2022-2023
Medication errors: Human error, miscommunication, similar packaging, incorrect dosage calculation, and misidentification.	**Medication errors:** Human error (misidentification, miscommunication, oversight	**Medication errors:** Human Error (Misidentification, Miscommunication, Lack of Attention, Incorrect Dosage Calculation)	**Medication errors:** Miscommunication, human error (misidentification, miscalculation, misinterpretation), lack of information sharing, incorrect prescription, and inadequate documentation.	**Medication errors:** Miscommunication, human error, incorrect prescriptions, lack of documentation.
Patient falls: Improper handling, lack of supervision, unsteady gait, and environmental obstacles.	**Patient falls** Inadequate patient monitoring, improper patient handling, environmental hazards		**Inadequate infection control:** Non-compliance with hygiene protocols, insufficient PPE supply, lack of training, and delayed isolation of patients.	**Inadequate infection control:** Lack of adherence to protocols, inadequate awareness/training, improper infection control measures.
Surgical and procedural errors: Miscommunication, human error, incorrect documentation, and lack of adherence to protocols.	**Communication error:** Miscommunication, lack of communication, inadequate handover	**Surgical and procedural errors:** Miscommunication, Lack of Proper Documentation, Human Error (Misidentification)	**Surgical and procedural errors:** Human error, miscommunication, misunderstanding of equipment operation, and lack of clarity regarding procedures.	**Patient transfer issues:** Inadequate communication, lack of timely intervention, missing information/orders
Patient identification errors: Human error, miscommunication, similar patient names, and inadequate verification procedures.	**Incorrect Documentation:** Human error (misidentification, miscommunication, oversight)	**Patient identification errors:** Mislabeling, Data Entry Errors, Lack of Verification	**Patient identification errors:** Human error during data entry, scanning wrong patient labels, mislabeling, and miscommunication.	**Communication error:** Inadequate communication, lack of clear documentation, misinterpretation of information.
Equipment malfunction /Technical failures: Improper handling, manufacturing defects, lack of maintenance, and incorrect setup.	**Equipment malfunction /Technical failures:** Equipment failure, improper maintenance, human error	**Equipment malfunction /Technical failures:** Equipment Malfunction, Communication System Failure	**Delayed Diagnosis and treatment:** Miscommunication, lack of coordination, failure to review laboratory results, and inadequate process handling	**Equipment malfunction /Technical failures:** Improper storage, maintenance issues, human error during assembly/disconnection
			Communication error: Miscommunication, lack of information sharing, and inadequate documentation.	

Several contributing factors have been consistently identified over the years. Communication breakdowns remain a major issue, including unclear orders, incomplete

handovers and misinterpretations. Lack of attention to detail is another important factor, often linked to human error, distraction and rushing. Inadequate training and supervision also play a crucial role, with inadequate training, lack of experience and poor supervision often contributing to critical incidents.

Documentation problems, such as incomplete, inaccurate or out-of-date records, are common contributors to incidents. Equipment and technology problems, including faulty equipment, outdated technology and confusing systems or system designs, are also common. In addition, fatigue and workload are significant human factors, with high workloads, staff shortages and fatigue affecting quality of care. Finally, the lack of double-checking procedures is a recurring theme.

4. Discussion and conclusion

The results show that an LLM can support the analysis of incident reports. Interesting to note that in the period after 2020, the issue of 'inadequate infection control' increases in importance. This shows that the CIRS messages really reflect current issues that can be identified using LLMs. However, the LLM extracts aspects although they are not explicitly described in the text. We recognized this particularly for contributing factors. Example are shown in Table 2 (example 1 and 2). The second example shows that the event information can be incomplete. To avoid hallucinations, the prompt could be adapted or examples could be provided. A rule-based postprocessing step could also help to eliminate hallucinations. Complex cases were also often source of errors and were not analysed correctly, as example 3 shows. Another observation is that the LLM well recognizes the health hazard, but the actual event remains unconsidered. The examples in the table demonstrate this observation. Future work could assess whether LLMs can identify which barriers have prevented serious patient harm. This could help to learn more about well-functioning prevention interventions that contribute to patient safety.

Compared to a previous analysis of the same data where BERTopic and word clouds were used to describe its content [6], we can recognize that BERTopic results were to a certain extent more specific as the top-n words can be used to describe the topic while LLMs provide a theme for events and causes which may result in generalizations.

Table 2. Types of failure events and their causes.

Incident report (translated from German)	Cause, event, contributing factor
Oxycodone 10/5 mg was administered instead of Oxynorm 10 mg Monitoring of the patient	**Event:** Medication error **Cause**: Human error (Medication administration) **Contributing factor:** Lack of Double-Check System, Inadequate Patient Monitoring
Patient brought into the operating theatre with jewellery, underwear and lab slip from another patient. Time out sheet was signed by department that everything was done correctly.	**Event:** Patient brought to surgery with incorrect items **Cause:** Human error/Lack of attention to detail **Contributing factor:** Incorrect items present (jewellery, underwear, another patient's lab results)
Patient admitted to the ward with abdominal pain due to diverticulitis and analgesia with Optifen 600mg 3/day. The Baisis GFR was already 40, i.e. reduced. Acute renal insufficiency was diagnosed during the course of the disease. Tubular necrosis was diagnosed microscopically, which is consistent with the nephrotoxic effect of Optifen. The medication was discontinued	**Event:** Acute kidney insufficiency **Cause:** Nephrotoxic medication (Optifen) **Contributing factor:** Pre-existing reduced glomerular filtration rate (GFR of 40), Diverticulitis;

In contrast to a classifier-based approach as introduced by Tabaie et al. for contributing factors [9], using LLMs provides the benefits, that contributing factors do not have to be known in advance: we do not need to train a classified to identify factors.

A limitation of this work is that we only assessed the quality of 100 randomly selected reports and their results which provides a first insight into the quality, but cannot be considered representative. Our aim was to gain insights into the analysis results from an LLM and its usefulness instead of conducting a well-designed assessment of the quality of extraction. We studied only the quality of one LLM, which is Gemma 2 9B. As mentioned earlier, this LLM was chosen to study a realistic set up for quality management departments in hospitals without any resources for running huge LLMs. Our prompt was following the zero-shot principle. Few-shot or fine-tuned approaches could be tested in future. The incident reports are in German and contain many abbreviations. We did not assess in detail whether the LLM interprets these abbreviations correctly. This remains open for future analysis as this may impact the quality of extraction.

In conclusion, LLMs show potential in analysing critical incident reports. While limitations such as hallucinations and errors remain, LLMs provide a basis for improving the efficiency and consistency of incident analysis. We envision a future hybrid approach where several automated methods analyse and aggregate information from incident reports, complementing human expertise in drawing actionable conclusions.

References

[1] Mitchell I, Schuster A et al. Patient safety incident reporting: a qualitative study of thoughts and perceptions of experts 15 years after '*To Err is Human*'. *BMJ Qual Saf*. 2016;**25:**92–99.

[2] Fernando GHS, Bandara T, and Purva M. Are Incident Reporting Systems in Healthcare Systems a Requirement for Improving Patient Safety? A Review. *INJHSR*. 2023. doi:10.51595/INJHSR22/019.

[3] Scott J, Dawson P. Content Analysis of Patient Safety Incident Reports for Older Adult Patient Transfers, Handovers, and Discharges: Do They Serve Organizations, Staff, or Patients? *J Patient Saf*. 2021;17: e1744–e1758. doi:10.1097/PTS.0000000000000654.

[4] Kowalski A et al. 17 years of the critical incident reporting and learning system "jeder-fehler-zaehlt.de" for primary care: Analysis of reports. *Z Evid Fortbild Qual Gesundhwes*. 2024;185: 10–16.

[5] Fong A. Realizing the Power of Text Mining and Natural Language Processing for Analyzing Patient Safety Event Narratives: The Challenges and Path Forward. *J Patient Saf*. 2021;17:e834–e836.

[6] Denecke K and Paula H. Analysis of Critical Incident Reports Using Natural Language Processing. *Stud Health Technol Inform*. 2024;313:1–6. doi:10.3233/SHTI240002.

[7] Ong M-S, Magrabi F and Coiera E. Automated identification of extreme-risk events in clinical incident reports. *Journal of the American Medical Informatics Association*. 2012;19:e110–e118.

[8] Wang Y, Coiera E, Runciman W, Magrabi F. Using multiclass classification to automate the identification of patient safety incident reports by type and severity *BMC Med Inform Decis Mak*. 2017;**17**: 84. doi:10.1186/s12911-017-0483-8.

[9] Tabaie A, Sengupta S, Pruitt ZM, Fong A. A natural language processing approach to categorise contributing factors from patient safety event reports. *BMJ Health Care Inform*. 2023;30:e100731.

[10] Young IYB, Luz S, and Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *International Journal of Medical Informatics*. 2019;132:103971. doi:10.1016/j.ijmedinf.2019.103971.

[11] Denecke K, May R, Rivera-Romero O. Transformer Models in Healthcare: A Survey and Thematic Analysis of Potentials, Shortcomings and Risks. *J Med Syst*. 2024;48:23.

[12] Frank O, Hochreutener M, Wiederkehr P, and Staender S. CIRRNET® - learning from errors, a success story. *Ther Umsch*. 2012;69:341–346. doi:10.1024/0040-5930/a000295.