Human Factors and Systems Interaction, Vol. 192, 2025, 76–86
https://doi.org/10.54941/ahfe1006630

Is LLM a Reliable Risk Detector? An Evaluation of Large Language Models in EMR-Related Medical Incident Detection

Siyuan Zhang and Xiuzhu Gu

Department of Industrial Engineering and Economics, School of Engineering, Institute of Science Tokyo, Tokyo, Japan

ABSTRACT

Medical institutions typically rely on manual analysis of adverse medical events, which requires significant human resources, time, and specialized knowledge and expertise. These requirements reduce the effectiveness of identifying potential risks. Can large language models (LLMs) leverage their powerful natural language processing capabilities to function as reliable risk detectors? In this pilot study, we aim to evaluate the effectiveness of LLMs in identifying electronic medical record system (EMR)-related medical incident risks. We first curated a dataset comprising 573 medical incident reports that had been manually analyzed. Then, using a few-shot prompting approach, we designed instructions to evaluate five LLMs, including GPT-4o, Claude 3.5 Sonnet, DeepSeek V3, Nova Pro, and Llama 3.1-405b. The results indicated that the best-performing LLMs could accurately extract more than half of the risk factors and generate reasonable explanations grounded in real-world case contexts. While general-purpose LLMs can provide some assistance, further optimization tailored to specific medical scenarios is necessary to enhance their capability in handling complex cases.

Keywords: Healthcare safety, Large language models, Medical incidents, Prompt engineering, Risk factors

INTRODUCTION

Large Language Models (LLMs) possess powerful natural language processing capabilities, entailing the potential to acquire extensive medical knowledge and complete complex tasks, which have garnered widespread attention. However, as an essential aspect of healthcare, the application of LLMs in healthcare safety remains underexplored. Compared to standardized medical knowledge assessments, healthcare safety analysis involves processing complex medical records and lacks fixed answers, making it more challenging. There is limited empirical evidence regarding the capabilities of LLMs in facilitating healthcare safety management, specifically their effectiveness in detecting and analyzing risks within complex medical systems.

The widespread adoption of electronic medical records (EMRs) has improved healthcare efficiency and safety (Lloyd et al., 2024). Meanwhile,

it has also brought significant risks. These risks stem from human factors among healthcare workers (HCWs), such as physical condition, professional experience, and work habits (Bisrat et al., 2021), as well as technical defects in the EMR system, such as hardware failures and power outages (Jo et al., 2024). The complexity and diversity of information in EMR-related medical incidents make these records ideal samples for studying the performance of LLMs in the field of healthcare safety.

Comprehensive review and management of medical information play a crucial role in healthcare risk identification and quality improvement, predominantly relying on manual operations by HCWs or administrative staffs. However, the continuously growing volume of medical data, encompassing patient records and nursing documentation, exceeds manual processing capacity, thereby increasing the risk of oversight in potential clinical hazards (Meeks et al., 2014). Meanwhile, the quality of manual reviews is constrained by individual experience and judgment, making the analysis prone to subjective bias and lacking consistency. LLMs have demonstrated significant potential in processing unstructured textual information and might help mitigate these challenges. However, there is currently no evaluation that reflects the performance of LLMs in handling the real complexity of healthcare safety records and associated analytical tasks.

By designing precise instructions and providing examples, the performance of LLMs can be significantly enhanced without requiring fine-tuning. This method not only optimizes the model's performance in specific tasks but also plays an important role in healthcare safety and quality management (Wang et al., 2023). With well-crafted instructions, the model can more efficiently identify and address potential risks within healthcare systems, thereby supporting the structuring and analysis of medical records.

Based on this context, this study aims to evaluate how effectively LLMs could identify risk factors in EMR-related incidents. Thus, we present three new insights:

1. Identifying the risk recognition capabilities of LLMs in healthcare safety and evaluating the rationality of the critical information and explanations provided by LLMs.
2. Developing instructions suitable for risk identification and interpretation generation task using a few-shot prompting approach.
3. Proposing practical applications and directions for future optimization.

METHODS

Dataset Preparation

We utilized 573 EMR-related medical incident reports from the Project to Collect Medical Near-Miss/Adverse Event Information system, managed by the Japan Council for Quality Health Care. The reports are written in Japanese and cover incidents from 2010 to 2023. Besides the time, severity, patient information, and involved personnel, the reports include detailed descriptions of medical incidents.

The extracted data from the reports initially had a complex format, with diverse types of information distributed across multiple free-text paragraphs, making subsequent processing challenging. To enhance data usability, we summarized the distribution of key information and the correlations between different information categories, converting unstructured text into a structured data format suitable for LLM tasks such as information extraction and explanation generation. Through text parsing and field extraction, we retained or merged key fields (e.g., patient information, incident details, background summaries) while removing irrelevant content to ensure a consistent field structure for each record.

Using the framework of established principles for information technology safety (Magrabi et al., 2015) and methods for analyzing human errors in healthcare (Itoh et al., 2009), we manually classified risk factors of each report. The taxonomy framework primarily included machine errors and contributing factors (such as human errors and contributing factors). The taxonomy process was initially conducted independently by two researchers with expertise in healthcare safety, with disagreements resolved by a third expert. The original classification system comprised four major categories and 79 specific subcategories to support multidimensional analysis of medical incidents. Due to significant differences in the frequency of various risk factors and to avoid difficulties in identifying rare cases, we finally selected 21 high-risk factors as the gold standard.

To facilitate subsequent instruction generation and testing, considering the potential impact of the scarcity of medical incident report data on LLM performance, we divided the dataset into two groups: a training set, 458 cases were used for tuning instructions during pre-testing and serving as a reference for the LLM; an evaluation set, 115 cases were used to assess the LLM's performance during formal testing. Figure 1 outlines the workflow from building dataset to evaluation: (1) making structured dataset, and (2) evaluating LLMs' responses against gold standard analyzed by professionals.

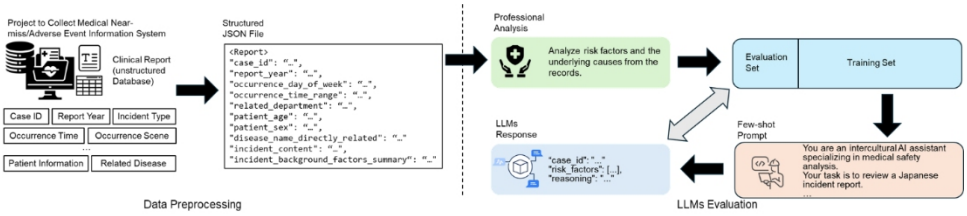

Figure 1: Workflow for dataset preparation to evaluation.

Instruction Design

To enable the model to identify risk factors from Japanese medical incident reports and generate corresponding reasoning, we designed standardized instructions to clarify task objectives, input formats, and output requirements. Ahmed et al. (2024) indicates that few-shot prompting, by

providing a small number of relevant examples, effectively guides the model to perform tasks more accurately. Unlike zero-shot prompting, which only includes task descriptions, few-shot prompting incorporates highly relevant contextual examples, providing the model with a clear reference framework. Therefore, we adopted the few-shot prompting approach in the instruction design.

To validate the effectiveness of few-shot prompting, we conducted comparative experiments using GPT-4o to evaluate both zero-shot and few-shot prompts. The results showed that while zero-shot prompting could produce some reasonable outputs, it exhibited significant variability, such as identifying risk factors outside the task scope or generating unrelated reasoning. In contrast, few-shot prompting had significantly higher output quality, ensuring that results aligned better with task objectives and remained consistent.

During optimization, few-shot prompting was further refined by selecting diverse and representative examples, therefore simplifying instruction phrasing and adding task constraints. These improvements enhanced the model's understanding of the task and improved the consistency of outputs. The optimized few-shot examples covered the diversity and representativeness of task objectives, while clear instructions reduced interpretation errors. Defining task boundaries minimized the generation of irrelevant content. After five consecutive test rounds, the model's performance stabilized, confirming the advantages of few-shot prompting in handling complex tasks.

Model Selection and Configuration

The selection of models is based on two core criteria: first, the model must support rapid deployment or direct application to meet the efficiency, and practicability demands of healthcare environments; second, the model must consistently generate task-compliant content to ensure reliable output in safety-critical scenarios. During the pre-test phase, we systematically tested models with varying parameter scales and types, encompassing both open-source and commercial models. Open-source models included versions with parameter scales ranging from 7b to 405b, such as Qwen 7b, Llama 3.1 8b, and Phi-4 14b. Commercial models included leading options like GPT-4o, Claude 3.5 Sonnet, and Amazon Titan. The results indicated that compared to models with smaller or medium scales high-parameter open-source models and most commercial models demonstrated superior performance by stably producing high-quality output. Additionally, these models showed potential for rapid deployment and adaptation to real-world scenarios. Therefore, five models were selected in this study: GPT-4o, Claude 3.5 Sonnet, Amazon Nova Pro, Llama 3.1 405b, and DeepSeek V3.

To ensure consistency and operability in the experiments, this study integrated five large language models into a unified Open WebUI interface through APIs and the Amazon Bedrock service. This interface serves as a standardized experimental platform, enabling consistent input and invocation across different models. All models were invoked using default

parameters to evaluate their baseline capabilities, thereby reflecting their most authentic performance. The temperature setting was configured to the default value provided by each model. The context length was set to the maximum range supported by the models. The output was structured in a standardized JSON format, which included a list of identified risk factors and a reasoning field.

Experiment Design

We used scripts to load the evaluation set and populate the instructions, inputting them one by one into the API of Open WebUI for testing with the corresponding models. The input instructions included the task description, few-shot examples, the current medical incident report text, and output requirements. To address the potential decline in recognition performance caused by the scarcity of medical incident report samples, the training set was converted into Markdown format and added to the Retrieval-Augmented Generation (RAG) knowledge base of Open WebUI to enhance the model's background knowledge. The embedding model was sentence-transformers/all-MiniLM-L6-v2.

We also employed scripts to store the model-generated results in JSON format. The stored data included the categories of risk factors identified by the model and the causal reasoning generated based on the input report. The outputs were categorized and stored according to the task ID of the input, ensuring precise matching between the outputs and the corresponding inputs.

Evaluation Metrics

To comprehensively reflect model performance, we applied quantitative and manual evaluation methods. For the task of extracting risk factors, four classic classification indicators were used as quantitative evaluation: hamming accuracy, precision, recall, and F1 score. These metrics were calculated by matching the risk factor lists generated by the models with manually annotated data.

Manual evaluation was employed for the task of generating reasoning and key information. The evaluation considered four metrics: coherence, relevance, factual consistency, and usefulness. Coherence focused on the logical clarity and organization of the generated content. Relevance assessed the alignment between the outputs and the input cases. Factual consistency measured the extent to which the generated content matched real case (Sedlakova et al., 2023). Usefulness evaluates the practical application value of the reasoning content. Manual evaluation employed a five-point Likert scale ranging from "1: completely non-compliant" to "5: completely compliant".

To ensure scientific rigor and comprehensiveness of the manual evaluation, stratified sampling was used to cover all identified risk factors and assess the corresponding reasoning capabilities. Specifically, the risk factors identified in the outputs of the five models were categorized, and stratified sampling was conducted based on the distribution characteristics of different risk types. This ensured that the sampled cases represented the full range of risk factors.

The same 35 cases (approximately 30% of evaluation set) were selected as evaluation samples for all five models. The outputs generated by each model were anonymized, and the specific source models were not disclosed during the evaluation process. Evaluators scored the outputs for each model across the four dimensions individually. This approach ensured a fair comparison of different models under identical task conditions while covering reasoning content corresponding to a variety of risk factors. The methodology also minimized the potential impact of sampling bias on evaluation results. Finally, a composite score was calculated by assigning equal weights (25% each) to all the four manual evaluation metrics, determining the overall performance.

RESULTS

Table 1 presents the performance of five LLMs in the task of extracting risk factors. GPT-4o achieved a hamming accuracy of 0.809 as the highest among all models, indicating its ability to comprehensively cover risk factors while effectively reducing incorrect predictions. Claude 3.5 Sonnet, DeepSeek V3, and Nova Pro demonstrated similar performance, demonstrating strong adaptability in multi-risk factor identification tasks. In contrast, Llama 3.1-405B had an accuracy of 0.269, highlighting its substantial shortcomings in overall consistency.

Table 1: Quantitative performance evaluation of LLMs in identifying risk factors.

Models	Hamming Accuracy	Precision	Recall	F1
GPT-4o	0.809	0.511	0.629	0.564
Claude 3.5 Sonnet	0.778	0.491	0.754	0.594
DeepSeek V3	0.782	0.495	0.690	0.577
Nova Pro	0.789	0.474	0.642	0.545
Llama 3.1-405b	0.269	0.431	0.417	0.424

GPT-4o showed the best performance of 0.511 in precision, indicating its ability to minimize false positives. However, its recall of 0.629 was lower than Claude 3.5 Sonnet's 0.754, suggesting it was slightly less effective at covering a broader range of potential risk factors. Similar to Claude 3.5 Sonnet, DeepSeek V3 offered balanced coverage and predictive accuracy by a recall of 0.690 and a precision of 0.495. Again, Llama 3.1-405b performed the poorest in both precision and recall, demonstrating its inability to meet task requirements in terms of risk factor identification accuracy and coverage.

Considering the combined performance of precision and recall, Claude 3.5 Sonnet achieved the highest F1 score of 0.594, indicating a strong balance between comprehensive coverage and predictive accuracy. DeepSeek V3 ranked second, showing competitive performance in balancing recall and precision. GPT-4o's F1 score was 0.564, slightly lower than that of Claude 3.5 Sonnet, but its superior precision highlighted its preference for high accuracy.

Figure 2 illustrates the overall performance of LLMs in providing reasoning for the identified risk factors. GPT-4o achieved the highest composite score of 4.28, reflecting its ability to generate content that

serves as effective references. Claude 3.5 Sonnet's score was close to GPT-4o, indicating good quality in its generated explanations, though some shortcomings existed. DeepSeek V3 and Nova Pro demonstrated stable mid-level performance. Llama 3.1-405b scored 2.52, significantly lower than the other models, indicating substantial weaknesses across multiple dimensions.

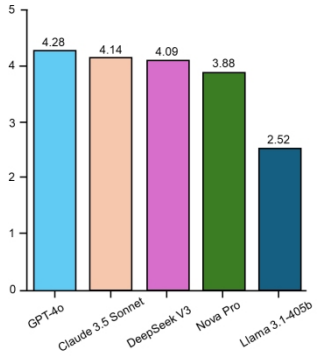

Figure 2: Overview of composite scores by manual evaluation.

Figure 3 provides a detailed depiction of the performance of each model across four metrics. Overall, the ranking of LLMs in the four metrics aligns with their total scores. GPT-4o demonstrated the most outstanding coherence, with a mean of 4.65 and the smallest standard deviation. This result indicates that GPT-4o consistently maintained strong logical abilities across diverse and complex cases. Claude 3.5 Sonnet and DeepSeek V3 respectively achieved similar means of 4.46 and 4.37. However, DeepSeek V3 exhibited greater variance, suggesting that while its logical content was generally acceptable, its stability varied across cases. Llama 3.1-405b had the lowest mean of 2.86 and the largest standard deviation, revealing significant shortcomings in coherence.

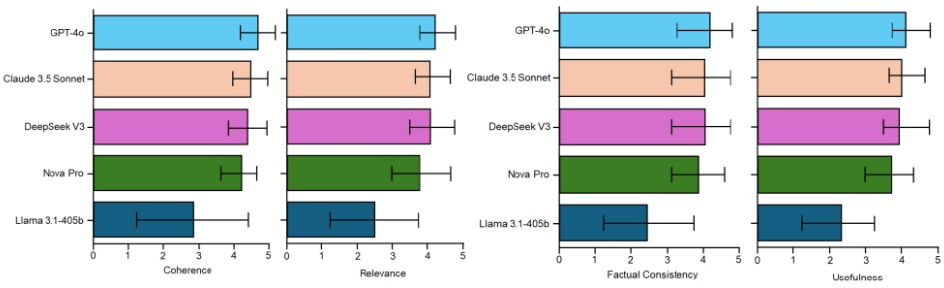

Figure 3: Mean and standard deviation of four manual evaluation metrics.

In terms of relevance, GPT-4o achieved a mean of 4.20, demonstrating its ability to generate content closely aligned with risk factors and case details. Claude 3.5 Sonnet and DeepSeek V3 each achieved a mean of 4.06, indicating

their capability to understand case content and align with the relevant risk factors. Llama 3.1-405b had a mean of 2.49 and exhibited high variability, reflecting its limited ability to adapt to reports with varying complexity and its weaker alignment with risk factors.

GPT-4o also achieved the highest mean of 4.17 in factual consistency. This result reflects its reliability in generating content consistent with the facts in the reports, with fewer hallucinations. Nova Pro achieved a mean of 3.86, with a lower standard deviation compared to GPT-4o, suggesting stable performance but some room for improvement in accurately reflecting case facts. Llama 3.1-405b had the lowest mean and the largest variance, indicating frequent factual errors and hallucinations, which made it difficult to generate content aligned with the facts.

In terms of usefulness, GPT-4o achieved a mean of 4.09, indicating its high practical value as a reference tool, even though it may not yet be ready for direct application. Claude 3.5 Sonnet achieved a mean of 4.00 with the smallest variance, highlighting its stable performance and unique strengths in generating practically useful explanations. Nova Pro achieved a mean of 3.69, reflecting moderate practical value but showing limitations in terms of detail and depth.

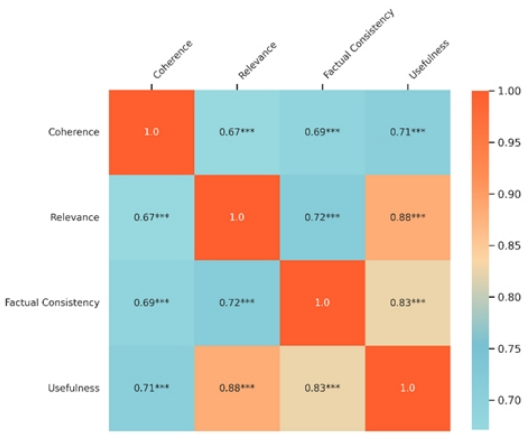

Figure 4: Correlations of four manual evaluation metrics by Spearman's rho.

Figure 4 depicts that all metrics are positively correlated, with all correlations reaching a significant level ($p < 0.001$). The correlation coefficients range from 0.67 to 0.88. The highest correlation was observed between relevance and usefulness ($\rho = 0.88$), indicating that in tasks involving the generation of explanations for risk factors, content that is more closely aligned with the input information tends to have greater practical application value. The correlation between factual consistency and usefulness was also high ($\rho = 0.83$), suggesting that the alignment of generated content with the facts in medical incident reports is a critical factor influencing its practical utility. Similarly, coherence shows a slightly lower correlation with usefulness ($\rho = 0.71$) but still underscores the importance of logical rigor in

enhancing the practical value of outputs. The lowest correlation is between coherence and relevance ($\rho = 0.67$), which possibly indicates that coherence is not a direct determinant of relevance in the generated reasons.

DISCUSSION

This study evaluated the effectiveness of five mainstream LLMs in extracting risk factors and generating explanations within the context of healthcare safety. The results showed that GPT-4o achieved the highest accuracy in predicting risk factors, while Claude 3.5 Sonnet demonstrated superior coverage of risk factors. DeepSeek V3 demonstrates a good balance between accuracy and coverage, showcasing unique advantages. They also performed well in generating explanations. Meanwhile, the high relevance between the model's output and the input cases can provide more practically meaningful references.

The differences in model performance can be attributed to several potential factors. Data scale significantly impacts the effectiveness of LLMs in identifying risk factors. The strong performance of GPT-4o and Claude 3.5 Sonnet is largely due to their large training datasets and broad coverage, which include public and third-party data sources (Wu et al., 2024). Extensive data scales enhance the models' breadth and generalization capabilities, making them more adept at recognizing linguistic patterns in reports compared to models with smaller training datasets. However, data in the medical field is relatively scarce compared to domains such as economics and law, with strict access and usage regulations to protect personal privacy and ethical standards. Regarding healthcare safety-related reports, publicly accessible datasets are limited, constraining LLMs' ability to improve performance in this specific domain.

Our findings suggested that GPT-4o and Claude 3.5 Sonnet generated explanations that were more closely aligned with report details and handled logical relationships effectively, even when processing complex incident reports with interacting factors. The reason could be the parameter scale that is considered a direct factor affecting the quality of explanations generated by LLMs. Larger parameter scales generally result in stronger representational capabilities as they capture more complex patterns and relationships. Thus, when expanding the size of incident-related training data is not feasible, models with larger parameter scales can better leverage the input cases for analysis.

Moreover, specialized fine-tuned models in the medical domain, such as Med-PaLM and ClinicalGPT, have demonstrated strong capabilities in generating medical knowledge. However, no dedicated models have yet emerged for the field of healthcare safety. The findings presented in Maharjan et al. (2024) indicate that general-purpose (base) models can outperform intensively fine-tuned models in performing specialized healthcare tasks when optimized through prompting alone. In this study, structured instructions built with few-shot prompts successfully enabled LLMs to capture the complex elements inherent in real-world medical workflows. This demonstrates the potential of prompt engineering to unlock

emergent properties in large foundational models, laying the groundwork for leveraging LLMs to tackle more complex and diverse healthcare safety tasks in the future.

Additionally, as a knowledge-enhancement module that combines dynamic retrieval and generation, RAG is considered an effective strategy for strengthening LLMs performance on specific tasks without requiring fine-tuning. Li et al. (2024) showed that integrating ChatGPT with RAG outperformed manually designed templates in organizing medical information and generating templates, particularly in reducing hallucinations. This highlights the potential of RAG in improving model accuracy and reducing hallucination-related issues. In this study, a knowledge base composed of a small sample demonstrated the benefits of introducing task-specific external knowledge to address the limitations of the model's internal knowledge. It is worth noting that RAG leverages external knowledge repositories that receive timely updates, enabling LLMs to incorporate the latest changes and ensure the relevance and accuracy of their outputs. However, the exploration of RAG and knowledge bases in the context of safety remains limited in this study. Improving data structuring to facilitate effective retrieval and integrating retrieved information with input instructions could further enhance the quality of generated explanations.

The current limitations of LLMs indicate that they are still far from practical application. Although the tested LLMs provided lists of factors to varying degrees when identifying risks, they still have limitations in omitting critical risk factors and failing to accurately identify risks. The complexity of medical incidents arises from the combined effects of multiple contributing factors whereas the impact of each factor is not uniform. This implies that LLMs need a profound "understanding" of complex medical incident scenarios rather than relying solely on pattern inference (Wang and Shen, 2024). To meet task requirements, LLMs must enhance their causal reasoning capabilities and the ability to recognize interactions among multiple factors.

CONCLUSION

The preliminary evaluation of this study suggests that general-purpose LLMs offer notable support in healthcare safety scenarios due to their extensive training scope and strong representational capabilities. Designing effective instructions enables LLMs to accurately perform tasks such as information extraction and explanation generation. Additionally, the integration of external knowledge bases and targeted optimization can further help LLMs realize their full potential. For practitioners aiming to incorporate LLMs into medical incident analysis in the future, the findings of this study can serve as a reference, aiding them in better understanding the applicability and limitations of LLMs.

ACKNOWLEDGMENT

This work was supported by JST SPRING (Grant Number JPMJSP2180) and Japan Society for the Promotion of Science (Grant Number 24K07926).

REFERENCES

Ahmed, A., Hou, M., Xi, R., Zeng, X. and Shah, S. A. 'Prompt-Eng: Healthcare Prompt Engineering: Revolutionizing Healthcare Applications with Precision Prompts', *Companion Proceedings of the ACM Web Conference 2024*, Singapore, Singapore: Association for Computing Machinery, 1329–1337.

Bisrat, A., Minda, D., Assamnew, B., Abebe, B. and Abegaz, T. (2021) 'Implementation challenges and perception of care providers on Electronic Medical Records at St. Paul's and Ayder Hospitals, Ethiopia', *BMC Med Inform Decis Mak,* 21(1), p. 306. doi: 10.1186/s12911-021-01670-z.

Itoh, K., Omata, N. and Andersen, H. B. (2009) 'A human error taxonomy for analysing healthcare incident reports: assessing reporting culture and its effects on safety performance', *Journal of Risk Research,* 12(3–4), pp. 485–511. doi: 10.1080/13669870903047513.

Jo, I., Kim, W., Lim, Y., Kang, E., Kim, J., Chung, H., Kim, J., Kang, E. and Jung, Y. B. (2024) 'Strategy for scheduled downtime of hospital information system utilizing third-party applications', *BMC Med Inform Decis Mak,* 24(1), p. 300. DOI: 10.1186/s12911-024-02710-0.

Li, A., Shrestha, R., Jegatheeswaran, T., Chan, H. O., Hong, C. and Joshi, R. (2024) 'Mitigating Hallucinations in Large Language Models: A Comparative Study of RAG-enhanced vs. Human-Generated Medical Templates', *medRxiv*, p. 2024.09.27.24314506. doi: 10.1101/2024.09.27.24314506.

Lloyd, S., Long, K., Probst, Y., Di Donato, J., Oshni Alvandi, A., Roach, J. and Bain, C. (2024) 'Medical and nursing clinician perspectives on the usability of the hospital electronic medical record: A qualitative analysis', *Health Inf Manag,* 53(3), pp. 189–197. doi: 10.1177/18333583231154624.

Magrabi, F., Baker, M., Sinha, I., Ong, M. S., Harrison, S., Kidd, M. R., Runciman, W. B. and Coiera, E. (2015) 'Clinical safety of England's national programme for IT: a retrospective analysis of all reported safety events 2005 to 2011', *Int J Med Inform,* 84(3), pp. 198–206. doi: 10.1016/j.ijmedinf.2014.12.003.

Maharjan, J., Garikipati, A., Singh, N. P., Cyrus, L., Sharma, M., Ciobanu, M., Barnes, G., Thapa, R., Mao, Q. and Das, R. (2024) 'OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models', *Scientific Reports,* 14(1), p. 14156. doi: 10.1038/s41598-024-64827-6.

Meeks, D. W., Smith, M. W., Taylor, L., Sittig, D. F., Scott, J. M. and Singh, H. (2014) 'An analysis of electronic health record-related patient safety concerns', *J Am Med Inform Assoc,* 21(6), pp. 1053-9. doi: 10.1136/amiajnl-2013-002578.

Sedlakova, J., Daniore, P., Horn Wintsch, A., Wolf, M., Stanikic, M., Haag, C., Sieber, C., Schneider, G., Staub, K., Alois Ettlin, D., Grübner, O., Rinaldi, F. and von Wyl, V. (2023) 'Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review', *PLOS Digit Health,* 2(10), p. e0000347. doi: 10.1371/journal.pdig.0000347.

Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., Yang, Q., Kang, Y., Wu, J., Hu, H., Yue, C., Zhang, H., Liu, Y.-H., Li, X., Ge, B., Zhu, D., Yuan, Y., Shen, D., Liu, T. and Zhang, S. (2023) 'Prompt Engineering for Healthcare: Methodologies and Applications', *ArXiv,* abs/2304.14670.

Wang, L. and Shen, Y. (2024) 'Evaluating Causal Reasoning Capabilities of Large Language Models: A Systematic Analysis Across Three Scenarios', *Electronics,* 13(23), p. 4584. Available at: https://www.mdpi.com/2079-9292/13/23/4584.

Wu, S., Koo, M., Blum, L., Black, A., Kao, L., Fei, Z., Scalzo, F. and Kurtz, I. (2024) 'Benchmarking Open-Source Large Language Models, GPT-4 and Claude 2 on Multiple-Choice Questions in Nephrology', *NEJM AI,* 1(2), p. AIdbp2300092. doi: 10.1056/AIdbp2300092.