Safety Science 181 (2025) 106695

Contents lists available at ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/safety

How can large language models assist with a FRAM analysis?

M. Sujan [a,b,*], D. Slater [c], E. Crumpton [b]

[a] Department of Computer Science, University of York, UK
[b] Human Factors Everywhere Ltd., UK
[c] School of Engineering, Cardiff University, UK

ARTICLE INFO

Keywords:
Large Language Model
Artificial Intelligence
FRAM
Healthcare
Runway Incursion

ABSTRACT

Large Language Models (LLMs) are transforming the way in which people interact with artificial intelligence. In this paper we explore how safety professionals might use LLMs for a FRAM analysis. We use interactive prompting with Google Bard / Gemini and ChatGPT to do a FRAM analysis on examples from healthcare and aviation. Our exploratory findings suggest that LLMs afford safety analysts the opportunity to enhance the FRAM analysis by facilitating initial model generation and offering different perspectives. Responsible and effective utilisation of LLMs requires careful consideration of their limitations as well as their abilities. Human expertise is crucial both with regards to validating the output of the LLM as well as in developing meaningful interactive prompting strategies to take advantage of LLM capabilities such as self-critiquing from different perspectives. Further research is required on effective prompting strategies, and to address ethical concerns.

1. Introduction

The development of Large Language Models (LLMs) represents a significant milestone within the recent resurgence of artificial intelligence (AI), the latter having been driven primarily by advancements in Deep Learning techniques and applications, e.g., in high-profile healthcare imaging and diagnostics research studies (McKinney et al., 2020; Topol, 2019). User-facing LLMs, such as OpenAI's ChatGPT and Google Gemini, represent the current peak of public interest in AI, showcasing unprecedented capabilities that have the potential to permeate all facets of society. The rapid adoption of LLMs is evident from the exponential growth in their user base, as seen, for example, with ChatGPT, which is said to have accumulated 100 million users within just two months of its public release at the end of 2022 (Wu et al., 2023).

LLMs have sparked widespread interest and generated considerable hype across many work domains with their potential to transform how people interact with technology and how work is done. In fields like medical decision making and education, LLMs have demonstrated their potential to augment human capabilities, leading to discussions about their integration into everyday practices (Ebrahimian et al., 2023; Roberts et al., 2023; Li et al., 2024; Vaishya et al., 2023). Beyond healthcare, industries spanning finance, manufacturing, and entertainment have also begun exploring the applications of LLMs in enhancing productivity and innovation (Kocoń et al., 2023; Deng et al., 2023; Badini et al., 2023).

Furthermore, the implications of deploying LLMs, especially in safety–critical contexts, have not escaped the attention of policymakers worldwide, prompting the development of numerous standards and guidance documents aimed at ensuring the safe and responsible use of frontier technologies. Notably, initiatives such as the UK-hosted global AI Safety Summit in November 2023 have underscored the urgency of addressing the safety implications associated with the use of LLMs and AI more widely. Numerous standards and guidance documents have been developed, including integrative standards, such as BS 30440 in the healthcare domain (Sujan et al., 2023), which synthesises best practices and guidelines into an auditable validation framework for healthcare AI.

The limitations of LLMs and their potential to contribute to risk in safety–critical domains have rightly received considerable scrutiny, even if further work is required to provide a more comprehensive understanding. For example, "hallucinations", i.e., false or imagined outputs of the LLM, which are not based on source data, are a frequently described phenomenon. Hallucinations have significant safety implications, because they may sound plausible and can be hard to detect (Bruno et al., 2023). A review examining the risks of using GPT-4 as a chatbot in healthcare also found that GPT-4 was prone to hallucinations and could mislead users (Lee et al., 2023). In addition, LLMs are trained

* Corresponding author at: Department of Computer Science, University of York, UK.
E-mail address: mark.sujan@york.ac.uk (M. Sujan).

https://doi.org/10.1016/j.ssci.2024.106695
Received 11 February 2024; Received in revised form 14 September 2024; Accepted 7 October 2024
Available online 16 October 2024

on publicly available data and do not have access to restricted data such as patients' medical records, which can affect their relevance and accuracy. Similar concerns have been raised in a study that investigated the use of ChatGPT (GPT-3.5) to provide information and advice regarding safety-related topics, such as using mobile phones while driving or supervising children around water (Oviedo-Trespalacios et al., 2023). The study provided examples of incorrect and potentially harmful advice given by the tool. In addition, the capabilities of LLMs (and generative AI more generally) could be exploited and put to malicious use thereby increasing security, political and societal risks (Brundage et al., 2018).

With the use and the associated safety-risks of LLMs being studied across many domains, it is timely to also consider how safety analysis and safety practitioners might benefit from employing LLMs. A recent study looked at the ability of ChatGPT to undertake an analysis using Systems Theoretic Process Analysis (STPA) and concluded that without human intervention the results might not be adequate nor suitable (Qi et al., 2023). However, it is worth bearing in mind that interactive LLMs, such as ChatGPT, are designed specifically for human intervention. Therefore, another way of approaching this topic is to ask how, rather than whether, LLMs can support such an analysis. In this paper, we explore this issue with respect to the Functional Resonance Analysis Method (FRAM) (Hollnagel, 2012), i.e., how can LLMs assist a FRAM analysis? As a corollary, we also provide suggestions for how safety practitioners might use LLMs when undertaking a FRAM analysis.

The next section (Section 2) provides basic background on FRAM. Then, we describe how we used two LLMs (ChatGPT using GPT-3.5 and Google Bard) to interactively develop a FRAM analysis of two example scenarios (Section 3). In Section 4 and Section 5, respectively, we describe the analysis performed by using Google Bard and ChatGPT. Section 6 discusses the findings within the wider context of using LLMs as part of the safety practitioner's tool kit. Conclusions are provided in Section 7.

2. FRAM

The Functional Resonance Analysis Method (FRAM) is a widely used systems-based approach that builds on the principles of Resilience Engineering (Hollnagel, 2012; Hollnagel et al., 2006). Initially developed as an accident model, FRAM has subsequently evolved into a flexible analysis method for studying everyday work (work-as-done or WAD). FRAM can be helpful for examining the variability resulting from trade-offs, adaptations and functional interactions inherent in complex socio-technical systems (Hollnagel, 2009).

FRAM uses the concept of "aspects" to analyse and represent interactions between functions. FRAM defines six possible aspects: input, output, control, resource, precondition and timing. When represented graphically, a function in FRAM is visualised as a hexagon. An open-source software tool to support the analysis is available (FRAM Model Visualiser FMV, on a user community GitHub site[1]).

A FRAM analysis typically involves four steps: (1) identification of functions; (2) description of performance variability of each function; (3) analysis of couplings between functions; and (4) monitoring and control of variability.

Over the past decade, numerous publications have described the application of FRAM across a range of industries, including aviation, healthcare, shipping and mining, as summarised in a literature review (Patriarca et al., 2020). Initially, the dominant application area was the aviation industry (Martinie et al., 2013; Herrera and Woltjer, 2010). More recently, however, FRAM has been applied frequently in healthcare settings (McGill et al., 2022; Salehi et al., 2021; Kaya et al., 2019; Schutijser et al., 2019; Sujan et al., 2023), potentially due to the "messy" reality of everyday clinical work, which appears to be particularly well suited to be analysed using FRAM.

3. Methods

This study employs critical reflection as a framework to examine the process of utilising LLMs for conducting a FRAM analysis (Fook et al., 2011). Drawing on Schön's concept of reflection-on-action (Schön, 1983), we use a self-reflective process to explore not only the strengths and weaknesses of using LLMs for FRAM analysis but also to extract valuable lessons for future research and practice in this evolving domain.

We used Google Bard and ChatGPT (GPT-3.5) as freely available tools for this study. ChatGPT was chosen due to its popularity. Google Bard was chosen from the pool of available LLMs, because it is regarded as one of the main freely available competitors to ChatGPT. Google Bard has been updated with a new model since the study was undertaken, and it is now known as Google Gemini. Both Google Bard and ChatGPT enable transparent exploration and replication of the study. Employing two different LLMs ensures that the study findings are not tied to a single tool, thereby enhancing their potential application to a broader range of similar tools within this quickly evolving landscape.

Two scenarios were selected, one from the healthcare domain and one from the aviation domain, as described in detail in sections 4 and 5. Each scenario was analysed using only one of the tools, as the main focus was on exploring how such tools could be used, rather than to compare the outputs of the two tools against one another.

The FRAM analysis using the tools consisted of an initial prompt to describe the scenario and to frame and bound the scope of the analysis. Subsequent prompts were triggered by identified gaps in the preliminary analysis and further questions of importance for the analysis. This interaction was responsive to the output of the tool, e.g., if obvious gaps or issues were identified with the output (such as missing out essential functions), then the tool was prompted to address this.

We critically reflected on the analysis process developed by the LLMs in interaction with the analyst. This reflection is not intended as an evaluation of LLM performance or comparison with human performance. The intention is to explore promising ways of using LLMs to support the FRAM analysis.

4. Google Bard to support FRAM analysis of emergency CT pathway

This section describes the FRAM analysis of an emergency CT pathway with the help of Google Bard. The example was chosen because one of the authors has familiarity with the emergency CT pathway.

4.1. Scenario description

The pathway for this scenario is the emergency CT pathway in an English NHS hospital. The pathway starts in the emergency department, where an electronic referral for an emergency CT is made. The referral is received by the radiology department, assessed and triaged, and then the patient is transferred to the radiology department for their scan. Scan images are recorded electronically and are reviewed and reported by the radiologist. The referring emergency department physician can access the radiology report electronically to inform their clinical decision making.

4.2. Identifying functions and couplings

Google Bard was instructed to begin the FRAM analysis with the prompt shown in Table 1 (prompt 1). As can be seen, the tool was supposed to assume the role of a human factors expert. The start and end functions were provided.

At this stage, the tool only identified and briefly described the functions as instructed. The tool identified eight functions as shown in

[1] https://github.com/functionalresonance/FMV_Community_Edition/wiki.

Table 1
Prompts used to interact with Google Bard.

Prompt ID	Prompt content
1. Initial prompt for the FRAM analysis of the emergency CT pathway.	You are a human factors professional. You want to use the Functional Resonance Analysis Method to analyse the process of doing an emergency CT scan in a large hospital in England. First, identify the main functions involved starting with "To request an emergency CT scan" originating in the emergency department and finishing with "To report on the CT scan".
2. Prompt to add aspects and couplings to previously identified functions.	In FRAM, every function has six aspects, which can be used to identify couplings between the functions. For the above functions, develop their aspects and couplings. The six aspects are: input, output, control, resource, precondition, timing.
3. Prompt to reduce the number of orphans.	Your FRAM model is missing background functions. Ensure that there are no "orphans" in the FRAM model. This means that every aspect needs to be created as an output of some other function.
4. Prompt to consider variability and to create recommendations for improvement.	In the above FRAM analysis of emergency CT, identify the main types of output variability and describe how this variability could potentially propagate or reinforce other variability. Then, consider ways of monitoring and controlling output variability in practice.
5. Prompt to generate recommendations from a resilience engineering perspective.	How can Resilience Engineering principles help monitor and control variability in the emergency CT process? Give specific examples.
6. Prompt to self-critique from the perspective of an experienced radiographer.	Act as an experienced radiographer working in an NHS radiology department. Critique the FRAM model from this perspective and identify gaps and oversimplifications where appropriate.
7. Prompt to self-critique from the perspective of a safety and quality improvement professional.	Now act as a patient safety professional working in a quality improvement department in an NHS hospital. Review and critique the recommendations made on the back of the FRAM analysis.
8. Prompt to self-critique from the perspective of an independent safety expert.	You are an independent patient safety expert. The patient safety team at an NHS hospital have used FRAM to analyse the variability and safety of the emergency CT pathway. Review and critique their recommendations for improving resilience from an independent perspective.

Table 2 (functions 1 – 3 and 7 – 11). Upon inspection, the functions appeared to describe the pathway reasonably meaningfully, albeit generically. This is expected, because the tool does not consider the specifics of one environment and instead draws on a range of data sources. On the other hand, there was an obvious gap in the functional description, because the set of functions does not describe how the patient gets transferred from their assessment in the emergency department to the scan in the radiology department (gap between functions 3 and 7). This was communicated to the tool in a subsequent prompt, and the tool generated three additional functions around patient transfer, labelled 4, 5 and 6 in Table 2. In addition, the analyst could consider on a function-by-function basis if there is value in prompting the tool to break down specific functions into several functions based on the description already provided by the tool.

The next step in the FRAM analysis is to think about the aspects of the functions and the couplings between different functions via their aspects. The tool was prompted to develop each function with its aspects and couplings, see Table 1 (prompt 2).

The textual output of Google Bard's response was manually imported

Table 2
Functions identified through initial and follow-up prompts.

Function	Description
1. To request an emergency CT scan	**Initiating the request**: This could involve a doctor, nurse, or other healthcare professional in the emergency department recognising the need for a CT scan based on the patient's presentation and symptoms. **Gathering clinical information**: The requesting clinician would need to gather relevant clinical information about the patient, such as medical history, current symptoms, and vital signs. **Completing the scan request form**: This form would typically include information about the clinical indication for the scan, the body region to be scanned, and any specific concerns or priorities.
2. To triage the request and determine clinical urgency	**Assessing the clinical urgency**: A designated individual, such as a radiologist or senior clinician, would assess the clinical information and scan request to determine the urgency of the scan. This might involve using scoring systems or clinical guidelines. **Prioritising the scan queue:** Based on the urgency assessment, the scan request would be placed in the appropriate queue for scheduling. Emergency scans would take priority over less urgent cases.
3. To schedule the CT scan	**Identifying available scanner time:** The radiology department would need to identify a CT scanner that is available and suitable for the requested scan type. This might involve considering factors like scanner capabilities, patient compatibility, and ongoing procedures. **Communicating with the patient and referring clinician:** The radiology department would inform the patient and referring clinician about the estimated wait time for the scan.
4. To request patient transfer	**Initiating the request:** Once the CT scan is scheduled, a designated staff member (nurse, clerk, or transporter) initiates a transfer request to move the patient from their current location to the radiology department. **Gathering patient information**: The staff member gathers essential details, including the patient's name, current location, scheduled scan time, any special needs or mobility restrictions, and the urgency of the transfer. **Communicating the request**: The request is typically communicated to a transport team or coordinating centre using an electronic system, phone call, or pager, ensuring clear and concise information.
5. To assign a porter	**Receiving and reviewing transfer requests:** The transport team or coordinator receives and reviews incoming transfer requests, evaluating patient needs, urgency, and available resources. **Identifying available porters:** They check for available porters with appropriate skills and capacity to handle the transfer, considering patient characteristics and any special requirements. **Assigning the task**: Once a suitable porter is identified, the transfer request is assigned to them, providing relevant patient information and instructions. **Notifying the porter:** The assigned porter is notified of the task, usually via a pager, phone call, or electronic notification system, along with essential details about the patient and transfer.
6. To transport the patient	**Receiving assignment and reviewing details:** The assigned porter receives the transfer assignment and reviews the patient's information, understanding their location, destination, any special handling needs, and safety concerns. **Locating the patient**: The porter proceeds to the patient's current location, using appropriate equipment like wheelchairs or stretchers if

(*continued on next page*)

Table 2 (*continued*)

Function	Description
	necessary, and identifies the patient correctly. **Communicating with the patient**: The porter introduces themselves, explains the transfer process, and addresses any patient questions or concerns to ensure understanding and cooperation. **Safely transporting the patient**: They follow hospital transport guidelines and safety protocols, ensuring patient comfort and security throughout the journey, and navigating any obstacles or delays effectively. **Arriving at the radiology department**: The porter delivers the patient to the radiology department on time for their scheduled scan, communicates their arrival to radiology staff, and provides any pertinent information about the patient's condition or needs.
7. To prepare the patient for the CT scan	**Obtaining informed consent**: The patient would need to provide informed consent for the CT scan procedure, being informed about the risks and benefits involved. **Addressing patient needs and concerns**: The radiology staff would address any questions or concerns the patient may have about the scan and ensure they are comfortable and prepared for the procedure. This might involve providing instructions on holding breath, removing metal objects, or taking contrast medication.
8. To perform the CT scan	**Positioning the patient in the scanner:** The radiology technologist would position the patient correctly in the CT scanner based on the body region being scanned. **Operating the CT scanner:** The technologist would operate the scanner controls to acquire the necessary images. This might involve adjusting scan parameters based on the patient's anatomy and clinical indication.
9. To analyse the CT scan images	**Image review and interpretation:** A radiologist would review the CT scan images to identify any abnormalities or findings relevant to the clinical indication. This might involve using specialised software and image manipulation techniques. **Preliminary reporting:** The radiologist would generate a preliminary report summarising the findings of the CT scan.
10. To report on the CT scan findings	**Finalising the report:** The radiologist would finalize the report, including interpretations, diagnoses, and recommendations for further management. **Communicating the results:** The radiology department would communicate the CT scan results to the referring clinician, typically through the electronic medical record system or verbally.
11. To document the CT scan procedure and findings	**Documenting scan parameters and patient information:** The radiology department would document the technical details of the CT scan procedure, as well as relevant patient information. **Archiving the scan images and report:** The CT scan images and report would be securely archived for future reference and potential re-evaluation.

into the FRAM Model Visualiser to produce a graphical representation. Upon inspection, it was found that there were large numbers of orphans, i.e., aspects that were defined for only one function. To address this, Google Bard was prompted to develop background functions to reduce the number of orphans, see Table 1 (prompt 3). The tool created accordingly a number of background functions. It is interesting to note that this did not result in no orphans as instructed, but still left a considerable number of orphans. Instead, Google Bard suggested that the model might benefit from further refinement because in complex systems there could be a significant number of background functions. Arguably with further prompting and focusing of the tool, it would have been possible to eliminate the remaining orphans. Further, because Google Bard included aspects for the initial starting function ("To request emergency CT"), a new starting function was manually created ("To identify need for CT") to keep the model in line with FRAM syntax. By convention, start and end functions in FRAM can only have outputs (start function) and inputs (end function), but not couplings via other aspects. The graphical representation in FMV is shown in Fig. 1. Remaining orphans are indicated with a red circle.

4.3. Analysing variability and developing recommendations

The next steps in the FRAM analysis are to consider variability and then to suggests ways to monitor and control variability. Google Bard was instructed as shown in Table 1 (prompt 4). Examples of the output of the tool are shown in Table 3. The identified examples of variability are, arguably, all meaningful. However, on closer inspection they tend to be focused on potential errors, misunderstandings and non-compliance with guidelines and protocols. In practice, a FRAM analysis is often used to highlight the trade-offs people need to make, which is not as evident in the Google Bard analysis. For example, triaging the referral request (function 2) is likely to have significant variability in terms of timing. This is because the radiologist needs to balance their triaging work with their reporting work, so the radiologist might focus on reducing the reporting backlog or they might prioritise triaging new referrals. In addition, when radiologists are reporting a complex case, they might prefer to do so uninterrupted and only look at new referrals once the reporting has been completed. Similarly, patient preparation for CT is highly variable, because it depends on the clinical condition of the patient (e.g., are they conscious, are they arriving in a bed or walking by themselves) and the amount of preparation work already done in the emergency department (e.g., has a cannula been placed, have items containing metal been removed).

Equally, the suggested recommendations are all meaningful, but often include a narrow focus on having clear protocols and standards, and monitoring and ensuring adherence to these. It is laudable that the tool recognises the need for equity and meeting the needs of a culturally diverse population. However, an analyst using FRAM with a resilience engineering background, might wish to focus more strongly on systems interventions that strengthen resilience. This was fed back to Google Bard, and the tool was prompted to develop recommendations from a resilience engineering perspective, see Table 1 (prompt 5). The tool responded with a set of four generic recommendations, along with a set of examples specific to the emergency CT pathway as illustrated in Table 4. These recommendations and specific examples are probably the weakest part of the AI-supported analysis because the recommendations are so generic that, at best, they serve as a reminder of key resilience engineering concepts. Similarly, while the specific examples are valuable, they are not sufficiently detailed and not linked transparently to the preceding analysis.

4.4. Google Bard self-critique

LLMs have the interesting ability to self-critique. This can be a useful exercise in addition to review and critique by a human analyst. Google Bard was prompted to critique its own analysis from different perspectives: (1) an experienced radiographer working in an NHS hospital; (2) a patient safety professional working in a quality improvement department; and (3) an independent safety expert, see Table 1 (prompts 6 – 8).

The radiographer perspective provides useful insights into gaps and oversimplifications. For example, the self-critique from this perspective resulted in highlighting the diversity in the patient population, which can include issues such as anxiety, mobility limitations and language barriers, which can all affect the variability observed in practice. Similarly, the tool reminds us that staffing levels fluctuate and that there is a significant shortage of radiographers, which can create capacity issues and bottlenecks. Further, the self-critique reveals that scheduling in practice is not as linear as represented in the model, that triaging is a

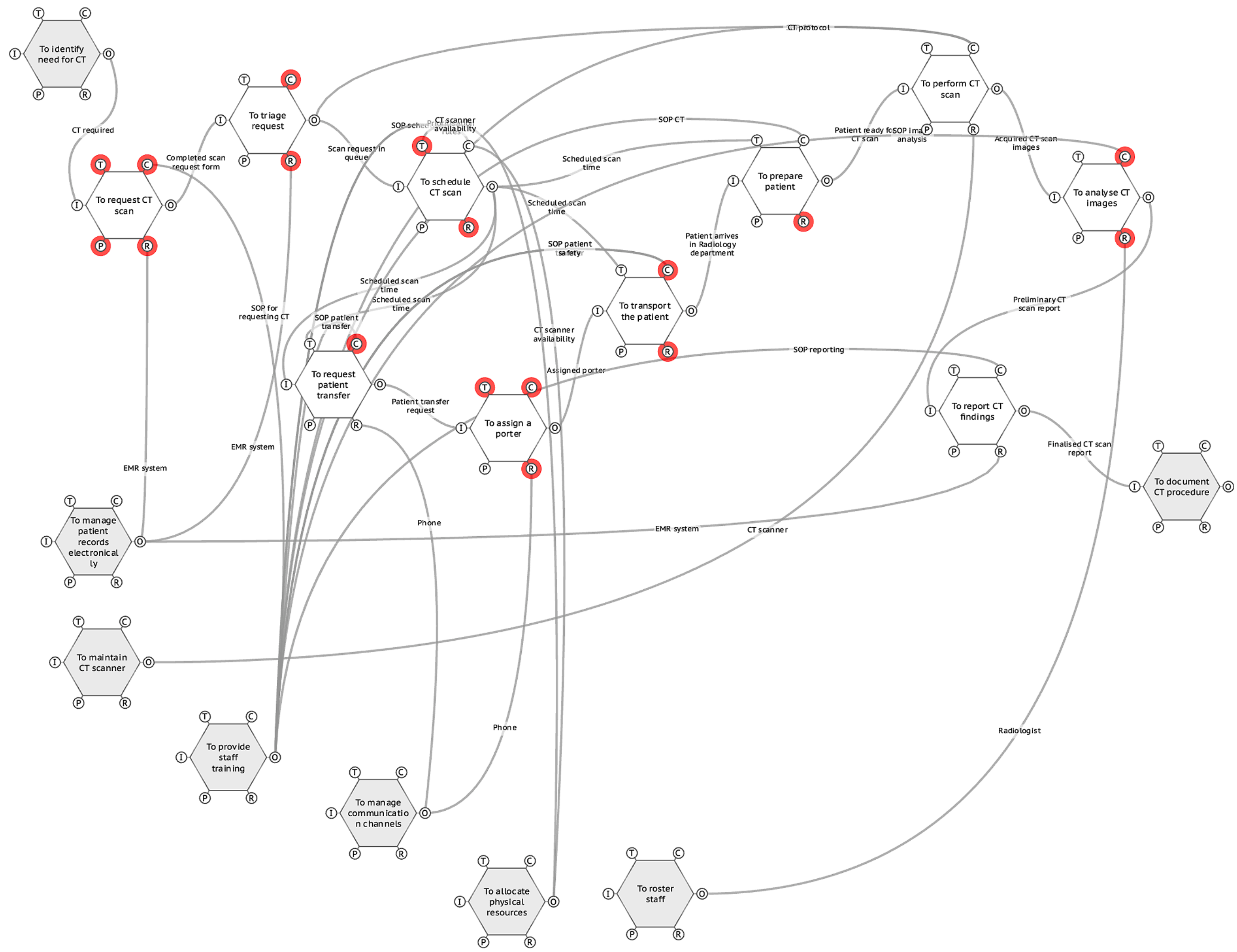

Fig. 1. Functions and couplings identified by Google Bard represented in FMV.

complex activity where multiple factors have to be balanced, and that frequent adjustments to scan protocols are necessary. All of these issues are, indeed, important aspects, which need to be captured in a FRAM analysis of work-as-done.

From the perspective of a patient safety professional working in a quality improvement department the tool critiques in a constructive way the lack of practical focus, e.g., around lack of prioritisation of recommendations, the need for developing measures to accompany interventions, and the need to consider the impact on staff and patients. From this perspective, the tool helpfully points to implementation issues, which also include the need to collect feedback on the interventions in practice, to monitor their impact, and to involve patients.

Lastly, assuming the role of an independent safety the expert the tool provides further, albeit overly generic suggestions. These include, for example, human factors integration, utilisation of technology, systems-based analysis and regular review and update. The independent safety expert also appears to lean towards a Safety-I perspective and recommends the use of root cause analysis to complement the focus on variability. This might be indicative of a certain bias in the tool to frequency of described approaches and interventions rather than their adequacy, i.e., in this case traditional thinking (Safety-I) has a much larger publication base than more recent approaches, which might bias the tool.

5. ChatGPT to support FRAM analysis of runway incursions

This section describes the FRAM analysis of a runway incursion scenario with the help of ChatGPT (GPT-3.5). The example was chosen due to the recurring and persistent nature of runway incursions as a major aviation safety incident.

5.1. Scenario description

The recent tragic accident where an Airbus A350 operated by Japanese Airlines (JAL 516) collided with a Japan Coast Guard aircraft at Haneda airport on 2nd January 2024 was caused by a runway incursion. This is not an unusual occurrence, and statistics from the Federal Aviation Administration indicate that there were 1,760 runway incursions in the US in 2023, up by around 1.7 % compared with 2022.[2] Investigations of such incidents and accidents often focus on the consequences of the incidents, but the actions of those involved frequently tend to be similar, whether near misses, or tragically disastrous. Therefore, it seemed appropriate to apply a systems analysis of the issues involved. The FRAM method seems appropriate to analyse how the system functions normally to identify any critical points, interactions or interdependencies which are needed to make the system work safely.

As the authors had only limited experience and domain expertise in Air Traffic Control systems, it was decided to explore whether an AI tool could be employed to supply the essential background information.

[2] https://www.faa.gov/airports/runway_safety/statistics/year/?fy1=2023&fy2=2022.

Table 3
Examples of variability and suggested improvements developed by Google Bard.

Function	Variability and Impact	Recommendation
2. To triage the request and determine clinical urgency	**Accuracy of urgency assessment:** Misjudgements can lead to inappropriate prioritisation and delays in urgent cases. **Consistency of decision-making:** Variations in triage criteria or interpretation can create inequities in access to scans. **Clarity of communication:** Unclear or delayed communication of triage decisions can disrupt workflow and create uncertainty for staff.	Regularly review triage criteria and guidelines, ensuring alignment with clinical evidence and best practices. Implement decision support tools or algorithms to guide triage and reduce variability. Establish clear communication channels and protocols for timely notification of triage decisions.
6. To transport the patient	**Time to transport the patient:** Delays in transport can impact scan schedules and patient care. **Patient comfort and safety:** Uncomfortable or unsafe transport can cause distress or injuries. **Communication with radiology staff:** Incomplete or delayed arrival notifications can disrupt workflow.	Track transport times, identify common causes of delays, and implement strategies to address them. Regularly review and update transport guidelines and safety protocols. Use real-time tracking systems to monitor patient transport and provide accurate arrival notifications.
7. To prepare the patient for the CT scan	**Patient adherence to instructions:** Non-compliance with preparation guidelines (e.g., fasting, contrast intake) can affect scan quality or necessitate rescheduling. **Timely completion of preparation:** Delays in preparation can disrupt scan schedules and impact patient flow. **Effectiveness of communication:** Misunderstandings or incomplete information can lead to patient anxiety, errors in preparation, or suboptimal scan outcomes.	Provide clear, concise, and culturally sensitive preparation instructions in multiple formats (verbal, written, visual aids). Implement reminder systems and track adherence rates to identify areas for improvement. Use standardised communication protocols and tools (e.g., checklists) to ensure comprehensive information exchange.

Table 4
Google Bard recommendations for improving resilience.

Generic recommendations	Specific examples
1. Monitor for early signs of variability 2. Build flexibility and adaptive capacity 3. Enhance monitoring and feedback systems 4. Facilitate learning and adaptation	**Flexible scheduling:** If a CT scanner malfunctions, cross-trained staff can quickly reassign patients to other available scanners, minimizing delays. **Adaptable protocols:** If a patient has difficulty following breath-hold instructions, the radiologist can adjust the scan protocol to obtain diagnostic images. **Real-time feedback:** If scan completion times are consistently exceeding targets, staff can immediately investigate the causes and implement corrective actions. **Learning from incidents:** If a patient receives the wrong contrast agent, a thorough analysis can identify system vulnerabilities and prevent similar errors in the future.

Further, the opportunity was taken to see whether the tool chosen (ChatGPT), was indeed capable of contributing not only that background information, but also provide the initial framework of functions necessary to start using the FRAM.

5.2. Developing an initial FRAM model

In order to develop the initial FRAM model, a single, longer prompt was used as shown in Table 5 (prompt 1). This prompt anchors the LLM as a professional systems engineer using FRAM. The LLM is instructed to use a task analysis (HTA) (Stanton, 2006) to identify functions, which are then used as the basis for the FRAM. This is one possible way of identifying functions as described in the FRAM handbook (Hollnagel, 2012), but a FRAM analysis does not necessarily have to start with a task analysis.

ChatGPT identified 14 functions as shown in Table 6. This output was then imported manually into the FMV software to produce a graphical representation (Fig. 2). As in the previous example (Section 4), entry and exit functions ("To schedule Flights" and "To Take Off"), were added and the aspect labels were standardised to ensure the correct linkages.

5.3. Identifying hazards and potential safety issues

Using a second prompt (Table 5, prompt 2), the LLM was then instructed to identify hazards and potential safety issues, and to describe how the variability of functions relates to these. ChatGPT suggested accordingly four key safety issues, associated functions, variability, and potential mitigations, as shown in Table 7.

Both the FRAM model and the suggested hazard list were useful in understanding the way this particular, unfortunate "instantiation" of the system played out. ChatGPT warns of the importance of communication

Table 5
Prompts used to interact with ChatGPT.

Prompt ID	Prompt content
1. Initial prompt for the FRAM analysis	Act as a professional, qualified systems engineer. You will analyse and model a complex sociotechnical system using the Functional Resonance Analysis Method. Start by analysing the functions utilised in the process of controlling the ground movements of aircraft at a major airport. In doing this you should identify the tasks that are needed to be carried out to deliver the output of the process the system is designed to facilitate. Use the language of a Hierarchical Task Analysis representation of how and in what sequence these tasks are carried out. Identify each of these tasks as achieved by a function, which you will label as "To do something". Set out a list of these functions and any interactions or interdependencies with other functions. Use a Business Process Modelling approach to identify the sequence of these functions in achieving the intended outcomes of the process from an incoming plane landing to a subsequent flight taking off. Set out your reasoning for your assignments. Then build a FRAM model of the system using all the functions identified in the HTA. List all the functions and subfunctions identified in the Hierarchical Task Analysis as FRAM functions with labelled inputs, outputs, preconditions, resources, control, and timing interdependencies with other function outputs as necessary, or discovered. Continue to do the full process for all of the functions. Do not use the continue as before short cut.
2. Prompt to suggest hazards and potential safety issues.	Identify any known issues and hazards with its operations and any functions needed to address these. Identify the functions and their variabilities involved in these unintended behaviours.

Table 6
Functions identified through initial prompt.

Function	Description
1.To receive landing clearance	**Inputs:** Communication from air traffic control **Outputs:** Cleared for landing signal **Preconditions:** Aircraft in the landing zone **Resources:** Communication equipment, air traffic control personnel **Control:** Air traffic control protocol **Timing:** Synchronisation with air traffic control communication
2. To Prepare for landing	**Inputs:** Landing clearance **Outputs:** Aircraft prepared for landing **Preconditions:** Cleared landing zone Resources: Aircraft systems, ground crew **Control:** Standard operating procedures **Timing:** Prior to landing clearance
3. To Communicate with air traffic control	**Inputs:** Aircraft status **Outputs:** Communication to air traffic control **Preconditions:** Landing process initiated **Resources:** Communication equipment, air traffic control frequency **Control:** Air traffic control communication protocol **Timing:** Throughout the landing process
4. To follow taxiway guidance.	**Inputs:** Ground movement clearance **Outputs:** Correct taxiing path **Preconditions:** Ready for taxiing **Resources:** Taxiway signage, ground control communication **Control:** Taxiing protocol **Timing:** Throughout taxiing process
5. To communicate with ground control	**Inputs:** Aircraft position **Outputs:** Ground control communication **Preconditions:** Taxiing in progress **Resources:** Communication equipment, ground control frequency **Control:** Ground control communication protocol **Timing:** As needed during taxiing
6. To Navigate to designated gate	**Inputs:** Gate assignment **Outputs:** Aircraft at designated gate **Preconditions:** Taxiing completed **Resources:** Navigation systems, ground crew **Control:** Gate assignment protocol **Timing:** End of taxiing process
7. To Prepare aircraft for boarding	**Inputs:** Boarding signal **Outputs:** Aircraft ready for boarding **Preconditions:** Aircraft at designated gate **Resources:** Ground crew, aircraft systems
8. To Check-in passengers	**Inputs:** Passenger information **Outputs:** Checked-in passengers **Preconditions:** Aircraft ready for boarding **Resources:** Check-in counters, personnel **Control:** Check-in procedures **Timing:** Before passenger boarding.
9. To Board passengers	**Inputs:** Boarding clearance **Outputs:** Passengers on board **Preconditions:** Aircraft ready for boarding, passengers checked in **Resources:** Boarding gates, personnel **Control:** Boarding procedures **Timing:** Passenger boarding process
10. To Conduct safety checks	**Inputs:** Safety checklist **Outputs:** Aircraft cleared for flight Preconditions: Passengers on board **Resources:** Inspection teams, safety protocols **Control:** Safety check procedures **Timing:** Before takeoff
11. To Verify cargo and baggage	**Inputs:** Cargo and baggage manifest **Outputs:** Verified cargo and baggage **Preconditions:** Aircraft cleared for flight **Resources:** Ground crew, cargo handling equipment **Control:** Verification procedures **Timing:** Before takeoff
12. To Receive takeoff clearance	**Inputs:** Communication from air traffic control **Outputs:** Cleared for takeoff signal **Preconditions:** Aircraft cleared for flight **Resources:** Communication equipment, air traffic control personnel **Control:** Air traffic control protocol **Timing:** Synchronization with air traffic control communication.
13. To Taxi to the runway	**Inputs:** Takeoff clearance **Outputs:** Aircraft at the runway **Preconditions:** Cleared for takeoff **Resources:** Navigation systems, ground control communication **Control:** Taxi to the runway protocol **Timing:** Before takeoff.
14. To Perform pre-takeoff checks	**Inputs:** Pre-takeoff checklist **Outputs:** Aircraft ready for takeoff **Preconditions:** Taxiing to the runway **Resources:** Flight crew, aircraft systems **Control:** Pre-takeoff check procedures **Timing:** Before takeoff

links, safety checks of key equipment such as runway lights, stop lights and taxiway confusion.

As was evident in the previous example, the LLM does not easily transition from the mindset underpinning a FRAM analysis towards applying the same mindset for the subsequent development of improvement suggestions and mitigations. In this example, all of the mitigation suggestions are very much rooted in traditional thinking and are not indicative of systems thinking. This would need to be addressed through a revised and iterative prompting strategy, for example as was done in the previous case of the emergency CT pathway.

6. Discussion

This study explored the potential of LLMs as tools for assisting with FRAM analyses. The application of two commonly used free LLMs to two representative safety–critical scenarios provided valuable insights into their strengths and limitations. Overall, these exploratory findings suggest that LLMs can assist with the creation of initial FRAM models, especially when guided by a human analyst through iterative prompting and critical evaluation. As demonstrated in the first example, LLMs possess the fascinating ability to self-critique, which represents a potentially transformative opportunity for how a FRAM analysis is done. Taken together, this suggests that the use of LLMs can be a valuable addition to the FRAM analyst's tool kit for different purposes, such as building an initial FRAM model, providing different perspectives, and supporting novices in getting started.

With the help of both tools, it was possible to develop initial FRAM models in the two examples. A FRAM model typically describes the potential variability in the system, whereas corresponding instantiations of the FRAM model represent actual variability in a given situation. The AI tools offered, in the first instance, a FRAM model based on a common, but single instantiation. Such a model is limited in terms of its robustness and ability to explore system variability. Further prompting needs to interrogate the system variability in order to develop a more comprehensive FRAM model that can be used to describe additional instantiations. In the first example, the analyst attempted to address this issue by prompting the system in such a way that a broader range of situations could be considered. For example, the initial prompt asked Google Bard to identify functions in the emergency CT pathway without providing a specific situation. Then, the third prompt asked to identify variability in each function and to consider different ways in which this could affect other functions. This type of interrogation can be the starting point for developing instantiations of actual variability. On the other hand, the second example did not use such a prompting approach, and the model remains more limited.

The first example demonstrated that the analyst identified gaps in the initial FRAM model, which were easily bridged through further guidance and prompting by the analyst. In the second example the analyst did not have detailed background knowledge, but in this instance the initial FRAM model could be useful as a starting point for subsequent deeper as well as broader exploration by the analyst. However, the difference between the first example and the second example strongly suggests that, inevitably, the LLM will leave gaps, and human expertise is required to identify and bridge these. When the analyst does not possess the necessary domain and system knowledge, it is essential that further input is sought because otherwise there is a risk that the gaps remain undetected.

LLMs are intended to be used as interactive tools, not static resources. Therefore, developing appropriate prompting strategies and critically evaluating the output of the LLM with targeted user feedback is crucial. The effectiveness of LLMs for assisting with a FRAM analysis is highly dependent on the quality and suitability of prompts and feedback. This study highlights that LLMs require specific direction to access diverse areas of their knowledge base, as illustrated, for example, by feedback on gaps in the initial model of the emergency CT pathway. There is a growing body of research on the development of effective

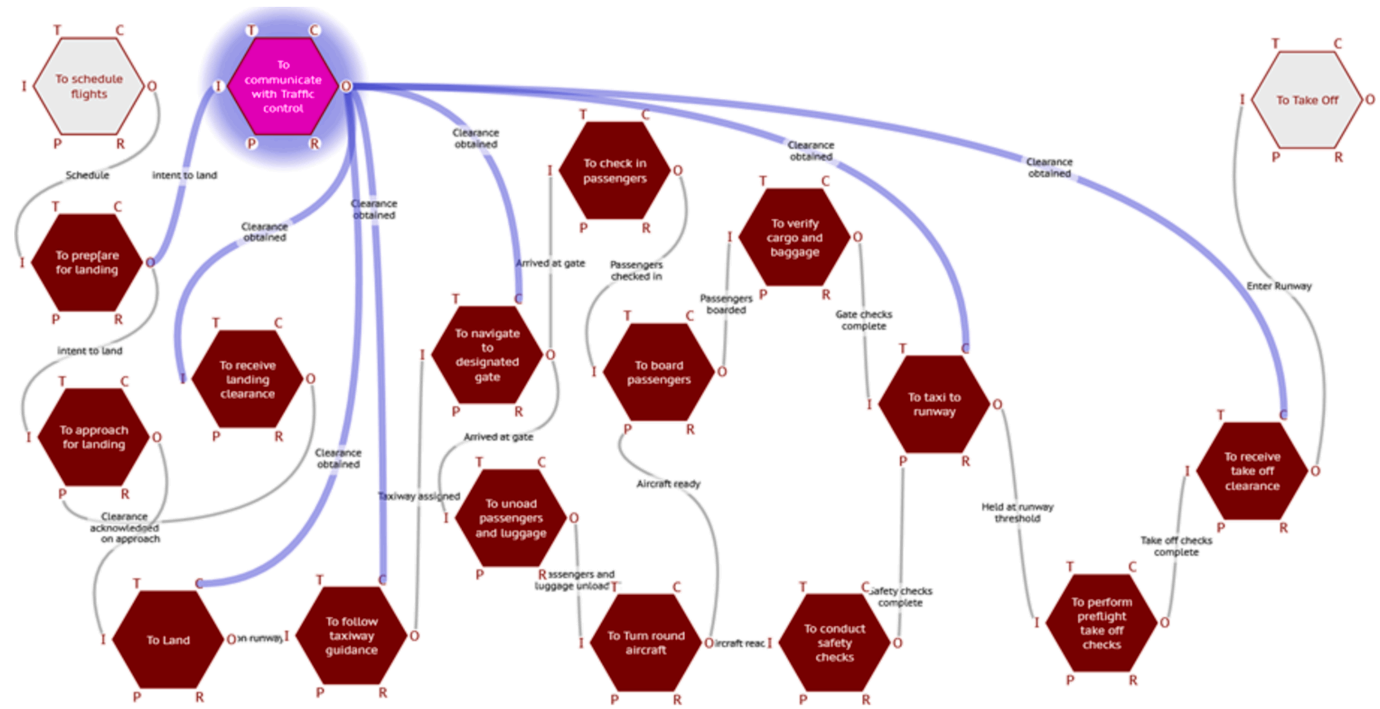

Fig. 2. Functions and couplings identified by ChatGPT in FMV.

Table 7
Safety issues identified by ChatGPT.

Safety issue	Functions	Variability	Mitigation
Communication Failures	Communicate with air traffic control, Communicate with ground control, Board passengers, Communicate with ground control during taxiing.	Technical failures, miscommunication	Redundant communication systems, rigorous training
Safety Checks Failure	Conduct safety checks, Verify cargo and baggage	Human error, equipment malfunction	Double-check procedures, regular equipment maintenance
Taxiing Protocol Deviation	Follow taxiway guidance, Navigate to designated gate.	Incorrect taxiing path, gate assignment issues	Improved signage, better ground control communication
Boarding Procedures Issues	Prepare aircraft for boarding, Check-in passengers, Board passengers	Boarding delays, passenger check-in issues	Efficient ground crew management, advanced passenger check-in systems

prompting strategies (Meskó, 2023; Wang et al., 2023). We used a hybrid prompting approach consisting of initial zero-shot prompting followed by feedback and refinement. Exploring other methods, such as one-shot or multi-shot prompting, where the LLM receives one or multiple examples, self-supervised prompting where the LLM asks questions to enhance its ability to perform the task, or even automated prompting techniques, could hold promise for further improving the capability of the LLM to support a FRAM analysis.

Related to this is the LLM ability to offer valuable self-critique (Weng et al., 2022) depending on how the analyst frames the context for the tool. It is fascinating to observe that in the first scenario the LLM was able to supply important additional insights about the variability of functions when prompted to do so from a clinical perspective. This underscores again the importance of prompting strategies, which need to harness this ability to self-critique and to integrate multiple perspectives into a model. However, this is an ability, which is as yet poorly understood and caution is required, especially when comparing LLM performance based on improvement through self-critique versus improvement through external expert validation and feedback (Valmeekam et al., 2023; Stechly et al., 2023; Luo et al., 2023).

The prompting strategy in the two examples was very different, and the outputs illustrate the importance of the approach to prompting. In the first example, the analyst interacted with the tool to identify and to bridge gap, and to push the LLM to consider different perspectives. The prompting in the second example was more simplistic. While, in part, this is explained by the modest aim in the second example of getting started with the analysis of a somewhat unfamiliar (to the analyst) system, the outputs reveal significant biases and weaknesses. For example, the initial prompt to start with a task analysis resulted in fewer organisational and background functions compared with the first example. In addition, the prompt to consider known issues and hazards resulting in undesired system behaviours apparently nudged the LLM to come up with what could be regarded a very traditional style of analysis, which, arguably, does not utilise and exploit the potential of FRAM. This illustrates the importance of articulating prompts in such a way that they are consistent with the mindset and principles underpinning FRAM, i.e., in this case resilience engineering thinking.

The findings of this study suggest that integrating LLMs into the analyst's tool kit holds promise. By offering diverse perspectives and prompting the identification of potential gaps or avenues for further exploration, LLMs can act as a valuable springboard for FRAM analysis. While domain-specific knowledge is crucial in order to be able to identify gaps and to sense-check the outputs of the LLM, it is interesting to reflect on the extent to which scenario-specific knowledge of the setting and system under consideration is required. The LLM has access to generalist knowledge documented across a potentially large number of studies undertaken in, for example, the emergency CT pathway (i.e., multiple descriptions of work-as-done across different settings). In the first example, this aggregated WAD knowledge was reasonably detailed to encompass the main variabilities of the scenario in a specific setting.

6.1. Limitations and future work

In this study, only two scenarios and the use of two tools were considered, and prompts were used intuitively rather than systematically. The findings are suggestive rather than definitive. However, they can serve as a starting point for further exploration.

The development of effective prompting strategies for optimal use of LLMs in FRAM analyses and in safety analyses more broadly requires further study. This should extend to the development of prompting patterns for interacting with an LLM (White et al., 2023) for assisting with a FRAM analysis (or any other kind of safety analysis). In addition, future research should continue to explore promising ways of integrating LLMs into the tool kit of safety analysts. This could involve, for example, the application of different LLMs to a wider range of scenarios, assessing the effect of different prompting strategies, and, eventually, empirical studies of the effect of LLM-assisted analysis on safety outcomes.

Alongside this, further research is required to address potential ethical concerns surrounding the use of LLMs in safety–critical domains, including issues such as bias (Lucy and Bamman, 2021), transparency and the ripple impact of the quality of data available to LLMs when more and more potentially sub-standard and wrong AI produced content is published online.

7. Conclusions

The use of LLMs affords safety analysts the opportunity to enhance the FRAM analysis by facilitating initial model generation and offering different perspectives. Responsible and effective utilisation of LLMs requires careful consideration of their limitations as well as their abilities. Human expertise is crucial both with regards to validating the output of the LLM as well as in developing meaningful interactive prompting strategies to take advantage of LLM capabilities such as self-critiquing from different perspectives.

Funding

This work was funded in part by Health Education England.

CRediT authorship contribution statement

M. Sujan: Writing – review & editing, Writing – original draft, Methodology, Investigation, Funding acquisition, Conceptualization. **D. Slater:** Writing – review & editing, Methodology, Investigation, Conceptualization. **E. Crumpton:** Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [**Mark Sujan**: is managing director of Human Factors Everywhere. The company offers commercial training in FRAM. **David Slater**: works with a not-for-profit group, FRAMsynt, which provides support to FRAM users. **Emma Crumpton**: none declared].

References

Badini, S., Regondi, S., Frontoni, E., Pugliese, R., 2023. Assessing the capabilities of ChatGPT to improve additive manufacturing troubleshooting. Advanced Industrial and Engineering Polymer Research. 6, 278–287.
Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:180207228. 2018.
Bruno A, Mazzeo PL, Chetouani A, Tliba M, Kerkouri MA. Insights into Classifying and Mitigating LLMs' Hallucinations. arXiv preprint arXiv:231108117. 2023.
Deng, X., Bashlovkina, V., Han, F., Baumgartner, S., Bendersky, M., 2023. In: What Do LLMs Know about Financial Markets? A Case Study on Reddit Market Sentiment Analysis. Association for Computing Machinery, Austin, TX, pp. 107–110.
Ebrahimian, M., Behnam, B., Ghayebi, N., Sobhrakhshankhah, E., 2023. ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model. BMJ Health & Care Informatics. 30, e100815.
Fook, J., 2011. Developing Critical Reflection as a Research Method. In: Higgs, J., Titchen, A., Horsfall, D., Bridges, D. (Eds.), Creative Spaces for Qualitative Researching: Living Research. SensePublishers, Rotterdam, pp. 55–64.
Herrera, I.A., Woltjer, R., 2010. Comparing a multi-linear (STEP) and systemic (FRAM) method for accident analysis. Reliab. Eng. Syst. Saf. 95, 1269–1275.
Hollnagel, E., 2009. The ETTO principle: efficiency-thoroughness trade-off. Ashgate, Farnham.
Hollnagel, E.F.R.A.M., 2012. the functional resonance analysis method: modelling complex socio-technical systems. Ashgate Publishing, Ltd.
Hollnagel, E., Woods, D.D., Leveson, N., 2006. Resilience engineering: concepts and precepts. Ashgate, Aldershot.
Kaya, G.K., Ovali, H.F., Ozturk, F., 2019. Using the functional resonance analysis method on the drug administration process to assess performance variability. Saf. Sci. 118, 835–840.
Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., et al., 2023. ChatGPT: Jack of all trades, master of none. Information Fusion. 99, 101861.
Lee, P., Bubeck, S., Petro, J., 2023. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N. Engl. J. Med. 388, 1233–1239.
Li, J., Dada, A., Puladi, B., Kleesiek, J., Egger, J., 2024. ChatGPT in healthcare: a taxonomy and systematic review. Comput. Methods Programs Biomed. 245, 108013.
Lucy, L., Bamman, D., 2021. Gender and representation bias in GPT-3 generated stories. In: Proceedings of the Third Workshop on Narrative Understanding, pp. 48–55.
Luo L, Lin Z, Liu Y, Shu L, Zhu Y, Shang J, et al. Critique ability of large language models. arXiv preprint arXiv:231004815. 2023.
Martinie, C., Palanque, P., Ragosta, M., Sujan, M.A., Navarre, D., Pasquini, A., 2013. Understanding functional resonance through a federation of models: preliminary findings of an avionics case study. In: Bitsch, F., Guiochet, J., Kaâniche, M. (Eds.), Computer Safety, Reliability, and Security: 32nd International Conference, SAFECOMP 2013, Toulouse, France, September 24–27, 2013 Proceedings. Berlin, Heidelberg, Springer, Berlin Heidelberg, pp. 216–227.
McGill, A., Smith, D., McCloskey, R., Morris, P., Goudreau, A., Veitch, B., 2022. The functional resonance analysis method as a health care research methodology: a scoping review. JBI Evidence Synthesis. 20.
McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., et al., 2020. International evaluation of an AI system for breast cancer screening. Nature 577, 89–94.
Meskó, B., 2023. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res. 25, e50638.
Oviedo-Trespalacios, O., Peden, A.E., Cole-Hunter, T., Costantini, A., Haghani, M., Rod, J.E., et al., 2023. The risks of using ChatGPT to obtain common safety-related information and advice. Saf. Sci. 167, 106244.
Patriarca, R., Di Gravio, G., Woltjer, R., Costantino, F., Praetorius, G., Ferreira, P., et al., 2020. Framing the FRAM: a literature review on the functional resonance analysis method. Saf. Sci. 129, 104827.
Qi Y, Zhao X, Huang X. safety analysis in the era of large language models: a case study of STPA using ChatGPT. arXiv preprint arXiv:230401246. 2023.
Roberts, R.H., Ali, S.R., Hutchings, H.A., Dobbs, T.D., Whitaker, I.S., 2023. Comparative study of ChatGPT and human evaluators on the assessment of medical literature according to recognised reporting standards. BMJ Health & Care Informatics. 30, e100830.
Salehi, V., Veitch, B., Smith, D., 2021. Modeling complex socio-technical systems using the FRAM: a literature review. Hum. Factors Ergon. Manuf. Serv. Ind. 31, 118–142.
Schön, D.A., 1983. The reflective practitioner: how professionals think in action. Basic Books, New York.
Schutijser, B.C.F.M., Jongerden, I.P., Klopotowska, J.E., Portegijs, S., de Bruijne, M.C., Wagner, C., 2019. Double checking injectable medication administration: Does the protocol fit clinical practice? Saf. Sci. 118, 853–860.
Stanton, N., 2006. Hierarchical task analysis: developments, applications, and extensions. Appl Ergon. 37.
Stechly K, Marquez M, Kambhampati S. GPT-4 Doesn't Know It's Wrong: An Analysis of Iterative Prompting for Reasoning Problems. arXiv preprint arXiv:231012397. 2023.
Sujan, M., Smith-Frazer, C., Malamateniou, C., Connor, J., Gardner, A., Unsworth, H., et al., 2023. Validation framework for the use of AI in healthcare: overview of the new British standard BS30440. BMJ Health & Care Informatics. 30, e100749.
Sujan, M., Pickup, L., de Vos, M.S., Patriarca, R., Konwinski, L., Ross, A., et al., 2023. Operationalising FRAM in healthcare: A critical reflection on practice. Saf. Sci. 158, 105994.
Topol, E.J., 2019. High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. 25, 44–56.
Vaishya, R., Misra, A., Vaish, A., 2023. ChatGPT: Is this version good for healthcare and research? Diabetes Metab. Syndr. 17, 102744.
Valmeekam K, Marquez M, Kambhampati S. Can Large Language Models Really Improve by Self-critiquing Their Own Plans? arXiv preprint arXiv:231008118. 2023.

Wang J, Shi E, Yu S, Wu Z, Ma C, Dai H, et al. Prompt engineering for healthcare: Methodologies and applications. arXiv preprint arXiv:230414670. 2023.
Weng Y, Zhu M, He S, Liu K, Zhao J. Large language models are reasoners with self-verification. arXiv preprint arXiv:221209561. 2022.
White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:230211382. 2023.
Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.L., et al., 2023. A brief overview of ChatGPT: the history, status quo and potential future development. IEEE/CAA J. Autom. Sin. 10, 1122–1136.