



Large language models powered system safety assessment: applying STPA and FRAM

Gulsum Kubra Kaya^{a,*}, Dominique Bovell^a, Mark Sujan^b, Graham Braithwaite^a

^a Cranfield University, Faculty of Engineering and Applied Science, Safety and Accident Investigation Centre, Cranfield, UK

^b Centre for Assuring Autonomy, Department of Computer Science, University of York, UK

ARTICLE INFO

Keywords:

LLM
AI chatbot
ChatGPT
Gemini
FRAM
STPA

ABSTRACT

The advancement of large language models (LLMs) shows immense promise in many domains. However, their reliability is still questionable. This study aims to comparatively examine the performance of ChatGPT and Gemini in conducting a stand-alone systems-based risk assessment using System-Theoretic Process Analysis (STPA) and the Functional Resonance Analysis Method (FRAM). Our findings revealed that both LLMs demonstrated weaknesses in their analyses, with ChatGPT generally outperforming Gemini regarding response comprehensiveness and adhering to the prompted format. Specifically, LLMs failed to use systems thinking in their stand-alone applications and failed to follow up on previous prompt outputs. While LLMs can provide substantial amounts of information quickly, the effectiveness of LLMs in system safety assessment is contingent on addressing their limitations and implementing strategies to improve their capabilities.

1. Introduction

Large Language Models (LLMs), a subset of artificial intelligence that can process and generate text data comparable with a human-level performance (Naveed et al., 2023), have started to be utilised in the safety domain. For instance, LLMs have been trialled in conducting accident classification (Ziakkas and Pechlivanis, 2023), generating safety cases (Sivakumar et al., 2024), and conducting hazard and risk analysis (Charalampidou et al., 2024; Collier et al., 2024; Nouri et al., 2024; Sujan et al., 2025).

While LLMs offer valuable insights by reducing the time required to apply complex hazard analysis methods (Charalampidou et al., 2024) and generating a reasonable safety analysis that analysts can use as a starting point for further expansion and modification (Charalampidou et al., 2024; Collier et al., 2024; Nouri et al., 2024; Sujan et al., 2025), various concerns have been shared regarding the reliability of their findings. For instance, Charalampidou et al. (2024) found that ChatGPT-4 was unable to provide correct and satisfactory unsafe control actions (UCAs). Qi et al. (2025) claimed that ChatGPT may be inadequate in producing reliable results in the STPA application, although they still highlighted its potential to outperform human experts with careful designing and prompt engineering.

In addition, LLMs, in general, present some challenges and

limitations. These include the models hallucinating, leading to the generation of incorrect or misleading information, which can present a significant challenge, especially for safety-critical applications (Peng et al., 2024). Moreover, their responses have been shown to represent hate speech, discrimination, malicious or harmful contexts, and racial, gender and religious bias, leading to ethical concerns in content generation (Peng et al., 2024; Tamkin and Brundage, 2021; Weidinger et al., 2022). LLMs also face concerns over their security, privacy and ethics regarding the unauthorised use, processing and disclosure of sensitive information (Feretzakis and Verykios, 2024) and environmental impacts regarding the carbon footprint of LLM-powered chatbots (Jiang et al., 2024).

Although a handful of studies have explored LLM applications in the safety domain, significant research gaps remain. Most existing research has assessed LLM performance with active human intervention, focusing on refining or validating LLM-generated outputs rather than evaluating their ability to conduct independent hazard analyses. However, there is a rising trend toward using Generative Pre-trained Transformers (GPTs) to conduct safety assessments with a single prompt request, without any follow-up or iterative guidance. This pattern reflects two emerging scenarios of concern: first, real-world misuse, where users rely on one-shot LLM outputs without expert validation; and second, autonomous or semi-autonomous deployments, where LLMs may be embedded in

* Corresponding author at: Cranfield University, Faculty of Engineering and Applied Science, College Road, Wharley End, Martell House, MK43 0AL Cranfield, UK.
E-mail address: kubra.kaya@cranfield.ac.uk (G.K. Kaya).

decision-support systems with minimal human oversight. These raise critical questions: Can LLMs perform a stand-alone system-based safety assessment without human involvement? Can they assess risks in the early design stages to reveal unknown unknowns without contextual prompts or human correction? How might their use transform traditional risk assessment processes?

This research seeks to address these gaps by examining the potential of LLMs to conduct a stand-alone system-based risk assessment without human intervention. The stand-alone application was selected to assess the baseline capability of LLMs, allowing to see LLM capability before human intervention and to consider the worst-case scenarios, where users may rely on the LLM models without adequate supervision. Specifically, this study utilises OpenAI's ChatGPT-4o and Google's Gemini Advanced to apply Systems Theoretic Process Analysis (STPA) and Functional Resonance Analysis Method (FRAM) in analysing the liquid hydrogen (LH₂) aircraft refuelling process, which is not a well-known process, that presents unique challenges in hazard identification. STPA and FRAM are frequently used in assessing risks, analysing hazards and exploring system interactions and performance variabilities in complex systems (Patriarca et al., 2020, 2022). These methods offer a comprehensive approach to hazard identification and risk assessment, enabling safety practitioners to account for dynamic system behaviours (Sadeghi and Goerlandt, 2023; Salehi et al., 2020). This study uses FRAM and STPA due to their widespread use, ability to address the limitations of traditional methods and, thus, assess risks in complex systems. This dual-method design allowed us to test the performance of LLMs across a spectrum of analytical demands, reflecting real-world challenges in safety analysis.

The remainder of the paper is structured as follows: Section 2 outlines the study methodology, including a brief description of the LLM tools used, the case study, methods applied (FRAM and STPA) and the overall study design, with emphasis on prompt development. Section 3 presents the findings from the LLM-powered applications of FRAM and STPA, following a step-by-step structure. Section 4 discusses LLM performance in executing FRAM and STPA under a single prompt (zero-shot) approach, along with the limitations identified. Section 5 concludes the study, summarising key outcomes and suggesting directions for future research.

2. Methods

2.1. LLM tools

This study used ChatGPT-4o and Gemini Advanced to examine their ability to conduct stand-alone FRAM and STPA applications. The aim was to assess the baseline capabilities of LLMs in single-prompt (a zero-shot) setting. This approach reflects two emerging scenarios: (1) worst-case human use, where users rely on LLM outputs without iteration, expert validation, or contextual understanding; and (2) autonomous or semi-autonomous deployment, where LLMs may be embedded into systems or tools with minimal human oversight. In both cases, the absence of adequate supervision, defined here as expert review, iterative engagement and safety-specific contextual knowledge, raises concerns about the reliability and safety of such applications. Our study aims to surface those risks and establish a foundational understanding of model performance under constrained conditions.

ChatGPT is built on the GPT architecture developed by OpenAI (Wu et al., 2023). The GPT model is pre-trained on large amounts of text data from sources such as web pages and books and is fine-tuned on specific datasets by in-context learning and reinforcement learning from human feedback (Ray, 2023; Wu et al., 2023). GPT models generate texts based on user queries and have succeeded in natural language processing tasks such as text generation, question answering, language translation, and sentiment analysis (Casheekar et al., 2024; Ray, 2023). ChatGPT's key features include task adaptability, contextual understanding, zero-shot and few-shot prompting and fine-tuning. These qualities mean that

ChatGPT can take on and be adapted to various tasks with very little or no training, making it suitable for a wide range of applications (Ray, 2023).

Gemini is Google AI's latest multimodal AI language model that generates texts, images, code and more (Imran and Almusharraf, 2024). Gemini is built on the decoder-only transformer model architecture but has been modified to encompass enhancing technologies such as tensor processing unit accelerators and a retrieval-augmented generation framework. These technologies allow Gemini to complete tasks that involve large amounts of information faster and more efficiently, as well as to retrieve relevant external information and incorporate it into its responses, leading to more accurate and factually correct responses (Islam and Ahmed, 2024). Gemini's distinguishing feature compared to ChatGPT is that Google's search engine is integrated, providing the most up-to-date information. Additionally, Gemini has the Google search function embedded into its interface, allowing the user to fact-check its output with a cross-Google search (Carlà et al., 2024).

2.2. Case study: liquid hydrogen aircraft refuelling process

This study selected the LH₂ aircraft refuelling process as a case study for three reasons: (1) analysing a process at early design stage: LH₂ refuelling systems are currently under development, which helps us evaluate LLM's ability perform in the absence of complete documentation and information; (2) analysing a complex system: The LH₂ refuelling process involves intricate interactions between human operators, mechanical systems, cryogenic storage and fuelling protocols, conditions that are ideal for testing both STPA (structured control-based analysis) and FRAM (variability and interaction-focused analysis); (3) analysing a process is within the interests of the authors. LH₂-powered aircraft refuelling is a critical area for research and development. The unique properties of LH₂ differentiate it from the current aviation fuels. LH₂ has a high flammability range between 4–75 %, a greater propensity to ignite due to lower ignition energy, burns with a blueish, near-invisible flame, and its cryogenic nature requires specialised vacuum-insulated storage at −253 °C to maintain its liquid form (Pritchard and Rattigan, 2010). Its properties give rise to potential hazards, including cryogenic burns, solidifying contaminants, blockage of components, pressure/expansion hazards of cryogenic fluid, component embrittlement and flammable hazards (Benson et al., 2019).

The large-scale deployment of LH₂ depends on the safety and economic feasibility of the refuelling systems (Hoelzen et al., 2022; Mangold et al., 2022). The LH₂ refuelling process is suggested to include docking the ground vehicle that executes the refuelling, connecting and purging the fuel hoses, chill-down of the hose and reduced mass flow, transfer of LH₂, a second purging of the system and finally, disconnection of the fuel hoses (Mangold et al., 2022).

2.3. FRAM

In this study, the FRAM was applied in four application steps: (1) identify and describe the functions, (2) identify system variability, (3) aggregate the variability, and (4) manage the variability. The first step is to identify the functions in six aspects (i.e., input, output, precondition, control, time and resource). The second step is to identify potential and actual variability within the system, in which this study could only consider the potential variability as the authors assessing a process that is in the early design stage. Variability within function outputs may arise due to variability within the function (internal), the working environment (external), and/or the coupling with upstream functions. This variability is characterised by timing (too early, on time, too late or not at all) and precision (precise, acceptable or imprecise). The third step is to analyse the upstream–downstream couplings between functions. It considers how the variability of one function may increase or decrease the variability of another. Finally, the fourth step proposes ways to manage uncontrolled performance variability identified in the previous

steps. It is to control the variability to prevent unwanted or adverse outcomes, and sustain the variability that leads to positive or desirable outcomes (Hollnagel, 2012).

The FRAM goes beyond the traditional cause-and-effect relationship and focuses on the notion of Work-As-Done (WAD) versus Work-As-Imagined (WAI) (Hollnagel, 2012). Rooted in resilience engineering principles, FRAM aims to overcome the limitations of traditional safety assessment methods, which are typically based on a decremental approach to safety (i.e., reducing the number of harmful outcomes), also referred to as Safety-I by those more aligned with incremental safety and Safety-II thinking. While Safety-I concentrates on barriers and defences for preventing failures and malfunctions, Safety-II, which FRAM is built on, focuses on understanding how things go right and the factors that contribute to the success within the system and considers the variability of human performance as inevitable and typically useful (Hollnagel et al., 2015; Patterson and Deutsch, 2015).

FRAM has been used for various purposes, including analysing accident safety performance indicators, comprehending WAI and WAD, exploring systems, assessing variability in everyday performance and enhancing emergency response strategies, across various industries (Buikstra et al., 2020; Damoiseaux-Volman et al., 2021; Kaya et al., 2019; Patriarca et al., 2020; Zheng et al., 2024). The authors selected FRAM due to its widespread use, different approach to safety, and ability to assess complex systems.

2.4. STPA

In this study, the STPA was applied in four steps: (1) identifying the purpose of the analysis, (2) modelling the control structure, (3) identifying unsafe control actions (UCAs), and (4) identifying loss scenarios. The first step involves defining the system boundaries, identifying losses, system-level hazards and constraints, and refining the hazards if necessary. The second step involves the identification of controllers, control actions, controlled processes and feedback loops to build the control structure. Control actions can be refined from the responsibilities of each entity in the control structure. In the next step, UCAs are identified by control actions that are (a) not provided, (b) provided, (c) providing too early, too late or in the wrong order or (d) lasting too long or stopped too soon, which will lead to a hazard. Controller constraints are identified based on each UCA identified. Finally, the loss scenarios describe the “causal factors that can lead to unsafe control actions and hazards.” (Leveson and Thomas, 2018). The STPA findings are used to develop safety recommendations and mitigation strategies.

STPA is built on a system-theoretic accident model and processes. STPA treats safety as a control problem and focuses on the whole system rather than solely focusing on the individual components (Leveson and Thomas, 2018). Similarly to FRAM, STPA concerns emergent properties of systems, which develop due to the interactions between system components.

STPA has been widely used to analyse hazards in various industries, including aviation, healthcare, manufacturing and space (Patriarca et al., 2022; Castilho et al., 2018; Dghaym et al., 2021; Kaya, 2021). Several studies compared STPA with traditional safety analysis methods and concluded that STPA outperforms other methods in conducting a more comprehensive analysis (Chatzimichailidou et al., 2018; Sulaman et al., 2019; Zhang et al., 2019). The authors selected STPA due to its wide use and ability to analyse complex systems in the early design stage.

2.5. Study design

This study was reviewed and approved through the Research and Ethics System (Reference: CURES/23347/2024). The study was designed to develop prompts that align with the FRAM and STPA steps, iteratively refine them, and examine the prompt outputs regarding their ability to follow FRAM and STPA steps (see Fig. 1). This study

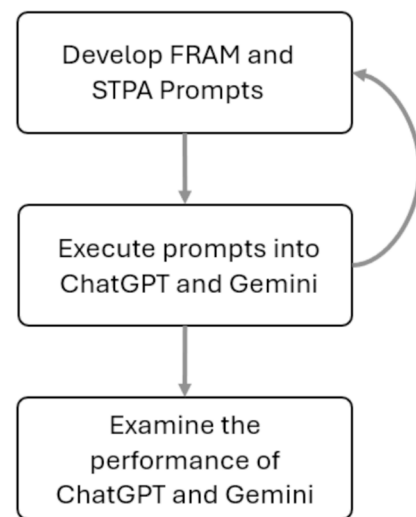


Fig. 1. Study design for conducting LLM-powered FRAM and STPA.

qualitatively examines the performance of two LLM models with their ability to follow FRAM and STPA steps without human intervention beyond the initiating prompts.

2.5.1. Developing the prompts

This study used various techniques to develop prompts for FRAM and STPA applications. A preparatory prompt was initiated to introduce the analyses and to prepare and provide the LLM with context to avoid ambiguity and to set precision, a technique also utilised in Sivakumar et al. (2024) and Sujan et al. (2025). In addition, persona prompting, a common prompt engineering technique (Chen et al., 2023), was utilised to prompt the LLMs to craft their responses from different perspectives: a safety engineer to conduct the FRAM and STPA. Finally, this study instructed the LLMs to follow a step-by-step process in the context of FRAM and STPA application and chain of thought prompting. This was also suggested by Chen et al. (2023) and Robertson et al. (2024), and it was found to provide clarity and instructions (Lo, 2023). To receive comprehensive outputs from each prompt, this study utilised controlling output verbosity by instructing the LLMs not to limit their responses, as suggested by others (Ekin, 2023; Lo, 2023).

In this study, after preparatory and persona prompting, the prompts followed the four basic structures: (1) ‘This is Prompt X’, (2) ‘This is the definition of...’, (3) ‘Here is an example...’ and (4) ‘This is the task...’. The FRAM prompts were developed using the FRAM book (Hollnagel, 2012), and the STPA prompts were developed using the STPA handbook (Leveson and Thomas, 2018). In this study, no follow-up prompts were provided to LLMs, as the study examines their ability to follow up on FRAM and STPA steps without human interference. After developing all prompts, the two authors set a meeting to execute and revise each prompt to ensure the clarity of the prompts and the comprehensiveness of their outputs.

In this study, the final FRAM prompts consist of seven prompts, and the STPA consists of ten prompts (see Fig. 2). The full prompt descriptions utilised for FRAM and STPA are in Appendix A in Tables A1 and A2, respectively.

2.5.2. Executing the prompts

ChatGPT-4o and Gemini Advanced LLM chatbots were utilised to run the prompts through the iteration process and the final prompt. After agreeing on the final prompts, the authors had single-prompt interactions with LLM chatbots, despite both methods requiring an iterative process. This was to explore the stand-alone performance of LLMs without human intervention, which is a research gap we aimed to address. The outputs from the finalised prompts were recorded on 3rd

FRAM Steps	FRAM Prompt ID	STPA Steps	STPA Prompt ID
Step 0: Recognise the purpose of the FRAM	Prompt 1- Explain concept	Step 1: Define the purpose of the analysis	Prompt 1: Explain concept
Step 1: Identify and describe the functions	Prompt 2: Identify functions		Prompt 2: Identify losses
	Prompt 3: Describe functions		Prompt 3: Identify system-level hazards
	Prompt 4- Identify couplings		Prompt 4: Identify system-level constraints
Step 2: Identify variability	Prompt 5: Identify variability	Step 2: Model control structure	Prompt 5: Identify controllers and control actions
Step 3: Aggregate variability	Prompt 6: Aggregate variability		Prompt 6: Identify feedback and controlled process
			Prompt 7: Identify unsafe control actions
Step 4: Consequence analysis	Prompt 7: Consequence analysis	Step 3: Identify unsafe control actions	Prompt 8: Identify controller constraints
		Step 4: Identify loss scenarios	Prompt 9: Identify loss scenarios
			Prompt 10: Identify safety recommendations

Fig. 2. The final FRAM and STPA prompts.

November 2024. The complete model responses are provided in [Appendix B](#) in [Tables B1–B4](#).

2.5.3. Examining the performance

The authors examined the LLM chatbots' performance, considering the LLM chatbot's ability to apply the methods and, so, follow up on the method steps. Two authors reviewed the full model outputs individually, using the qualitative examination metrics in [Table 1](#), which were developed based on FRAM and STPA application steps. The authors provided their review comments for each prompt output. After completing the independent reviews, the two authors held two full-day meetings to discuss their comments and the performance of the LLMs. Other authors reviewed the findings and provided further comments. All evaluators have research experience in system safety and have previously applied both FRAM and STPA in academic and applied contexts.

3. Results

3.1. FRAM findings

3.1.1. Step 1: identify and describe the functions (prompts 2, 3 and 4)

ChatGPT prompts 2 and 3 outputs identified 14 functions, listing 146 items in all functions' aspect descriptions. Gemini identified only 6 functions, listing 89 items in all functions' aspects and identified 3 additional supporting functions. The functions identified by ChatGPT were more detailed; however, Gemini was able to move beyond the task and identified background functions of <to communicate and coordinate>, <to manage emergencies> and <to train and maintain competency>. Gemini described the aspects of the functions in more detail. While ChatGPT had an average of 10.4 aspect items for each function, Gemini described an average of 14.8 items. [Fig. 3](#) provides two similar functions that ChatGPT (a) and Gemini (b) identified and described.

The LLMs did not provide a thorough and comprehensive coupling of aspects in the Prompt 4 outputs, where couplings and background functions were identified. In this prompt, the chatbots were tasked to create background functions to complete the couplings. ChatGPT

identified three background functions, while Gemini identified 11. However, both failed to link them with foreground functions. The reason is that although there is a link between the functions, the LLMs failed to write the aspects identically. With the FRAM Visualiser Model (FMV), the couplings can only be created if the upstream and downstream aspects are written identically with the exact wording.

ChatGPT revised the function aspects to have the exact wording. However, it led to some aspects being removed and was still unsuccessful in creating the couplings. ChatGPT only produced output-input couplings. For example, ChatGPT identified a coupling between the output <fuel transfer initiated> of the function <To initiate fuel transfer> acting as an input of the function <To monitor the refuelling process>.

Gemini provided an incrementally better performance using the exact wording, so couplings could be made in the FMV and produced a more nonlinear model. For example, the output <liquid hydrogen transferred> of the function <To transfer liquid hydrogen> acted as an input of the function <To complete refuelling>, and another output <monitored fuel quantity, temperature, pressure, and flow rate data> acted as an input to <To monitor and control the refuelling process>.

For both LLMs, Prompt 4 outputs required manual inputs from analysts to couple the functions and complete the FRAM modelling. [Fig. 4](#) shows the FRAM models using the outputs from ChatGPT (a) and Gemini (b).

3.1.2. Step 2: identify variability (prompt 5)

ChatGPT and Gemini had different focuses when identifying variability. ChatGPT identified 91 variabilities from its 14 identified functions, giving potential sources of variability in timing (too early, on time, too late, not at all) and precision (precise, acceptable, imprecise). For example, 7 variability cases were identified for the output <fuel transfer initiated> of the function <To initiate fuel transfer>. These were for timing: "The transfer is initiated too early, possibly before all checks or preparations are complete." (too early), "The transfer starts immediately after the hose is connected." (on time), "The transfer is delayed, leading to an extended refuelling time." (too late), and "The transfer is not

Table 1
LLM FRAM and STPA application performance examination metrics.

Prompts	Reviewer Considerations with LLM Chatbots Performance
FRAM	
Prompt 1*	<ul style="list-style-type: none"> to understand the system safety engineer persona to understand the case study context to understand the overall task
Prompt 2	<ul style="list-style-type: none"> to understand the given first and last functions to generate a comprehensive list of functions between the given functions
Prompt 3	<ul style="list-style-type: none"> to examine if functions were identified in six aspects to examine if there were functions with multiple entities identified in an aspect
Prompt 4	<ul style="list-style-type: none"> to examine if an explanation is given if an aspect is not identified. to examine if couplings between functions are identified. If identified, to examine if the couplings linked different aspects like output to input, output to precondition and output to control. to examine all background functions are added.
Prompt 5	<ul style="list-style-type: none"> to examine if variability sources are identified: endogenous, exogenous and upstream function. to examine if variabilities in terms of all timing and precision conditions are considered. to examine if variabilities are identified for all outputs identified in prompt 3.
Prompt 6	<ul style="list-style-type: none"> to examine if the impact of the variability of all outputs from all functions are considered to examine if no impact, dampening effect and amplifying effect cases are identified to examine if the simultaneous variabilities in the outputs from different functions are identified
Prompt 7	<ul style="list-style-type: none"> to examine if recommendations provided to prevent/mitigate uncontrolled performance variability leading to undesirable outcomes, to examine if recommendations are provided to sustain performance leading to desirable outcomes
STPA	
Prompt 2	<ul style="list-style-type: none"> to examine if all key stakeholders are considered to examine if losses identified considering each stakeholder
Prompt 3	<ul style="list-style-type: none"> to examine if all system-level hazards are identified following the hazard description formula to examine if all hazards are linked to losses
Prompt 4	<ul style="list-style-type: none"> to examine if system-level constraints are identified following the formula to examine if the system-level constraints are identified for each hazard to examine if the system-level constraints are indeed preventative/mitigative to hazards
Prompt 5	<ul style="list-style-type: none"> to examine if the high-level control structure is identified by considering high-level controllers to examine if the sub-level controllers are identified for a detailed control structure to examine if responsibilities identified for each controller, and these responsibilities are used to identify control actions to examine if controller's control algorithm and process models are identified.
Prompt 6	<ul style="list-style-type: none"> to examine if the feedback loops are defined to examine if responsibilities and control actions are used to refine feedback to examine the controlled processes are identified
Prompt 7	<ul style="list-style-type: none"> to examine if unsafe control actions are identified following the formula to examine if unsafe control actions are identified under four different conditions to examine if multiple unsafe control actions identified for each control action under the same condition to examine if explanations are given in case of not using the different conditions to examine if unsafe control actions are linked to hazards
Prompt 8	<ul style="list-style-type: none"> to examine at least a controller constraint is identified for an unsafe control action. to examine if the controller constraint identified can prevent the unsafe control action.
Prompt 9	<ul style="list-style-type: none"> to examine if multiple scenarios are identified for an unsafe control action to examine if two types of scenarios are considered

Table 1 (continued)

Prompts	Reviewer Considerations with LLM Chatbots Performance
	<ul style="list-style-type: none"> to examine if the scenario is linked to unsafe control action and so the hazard. to examine if scenario explains the causality of the unsafe control action and its consequences.
Prompt 10	<ul style="list-style-type: none"> to examine specific recommendations are identified focusing on the causalities on the loss scenarios. to examine if system-level constraints and controller constraints are refined to provide actionable suggestions

*STPA Prompt 1 is same as FRAM Prompt 1.

initiated, preventing the refuelling process.” (not at all), and for precision: “The fuel transfer is initiated smoothly with all parameters within the required limits.” (precise), “The transfer is mostly correct, though some minor adjustments might be necessary.” (acceptable), and “The transfer is poorly managed, leading to potential safety issues or inefficiencies.” (imprecise).

Gemini produced 57 potential variabilities by considering variability in terms of too early, too late and imprecise. In other words, Gemini did not consider the all-time and precision-related variabilities. However, it focused on all three variability sources: endogenous, exogenous and upstream variability. For example, five variability cases were identified for the output <liquid hydrogen transferred> of the function <To transfer liquid hydrogen>. These were endogenous variability in terms of time: “Transfer may take longer than planned due to equipment issues (pump malfunctions, flow restrictions), variations in fuel temperature or pressure, or unexpected complications during the transfer process” and precision: “Inaccurate fuel quantity transferred due to metering errors, leaks, or other technical issues”, exogenous variability in terms of time: “External factors like extreme weather conditions or sudden changes in ambient temperature can affect the transfer process and potentially cause delays” and precision: “Contamination of fuel lines or equipment can impact the quality and precision of the transfer”, and upstream variability in terms of time: “Delays in preparation or equipment issues can lead to a delayed start of the transfer process”. In addition, Gemini identified variability by considering the contextual and causality factors. However, both LLMs failed to provide variability cases for all outputs identified in the previous prompt.

3.1.3. Step 3: aggregate variability (prompt 6)

ChatGPT identified 42 aggregated variability cases and 3 simultaneous variability cases. For example, the variabilities on the output of <fuel transfer initiated> could result in “The transfer may start before all safety conditions are verified, introducing variability. This could potentially extend the refuelling process beyond the scheduled turnaround time or, if the fuel transfer is not properly initiated, lead to significant operational and safety issues, further amplifying variability in the monitoring process.”. For the simultaneous variability of delays in the functions <To confirm refuelling request> and <To prepare refuelling equipment>, the aggregated variability was identified as “If both functions experience delays, this could lead to significant amplification of variability in all subsequent functions. The delay in confirmation cascades through the equipment preparation, positioning, securing, and eventually leads to a much longer overall turnaround time.”

Gemini identified 8 aggregated variability cases and 3 simultaneous variability effects. For example, amplified variability in the function <to transfer liquid hydrogen> amplifies variability in the function <to monitor and control the refuelling process> by explaining that “delayed transfer puts pressure on monitoring and control to catch up and ensure safety.” and “imprecise transfer or inaccurate data can lead to misinterpretations, incorrect adjustments, and potential safety issues.”, resulting in “further delays, increased risk of overfill, leaks, or other incidents, potential for emergency situations”. Gemini did not provide specific simultaneous variability scenarios; however, it stated, “Simultaneous delays in early functions can create a cascading effect,

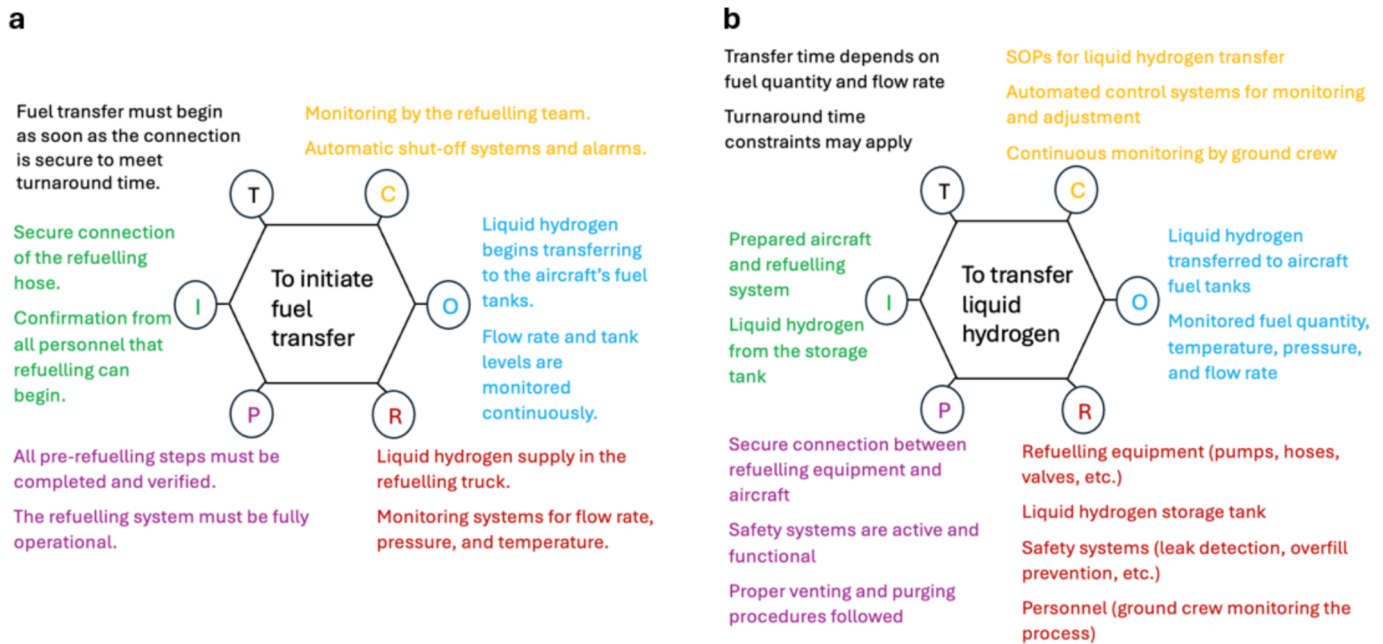


Fig. 3. An example of function identification and description by ChatGPT (a) and Gemini (b).

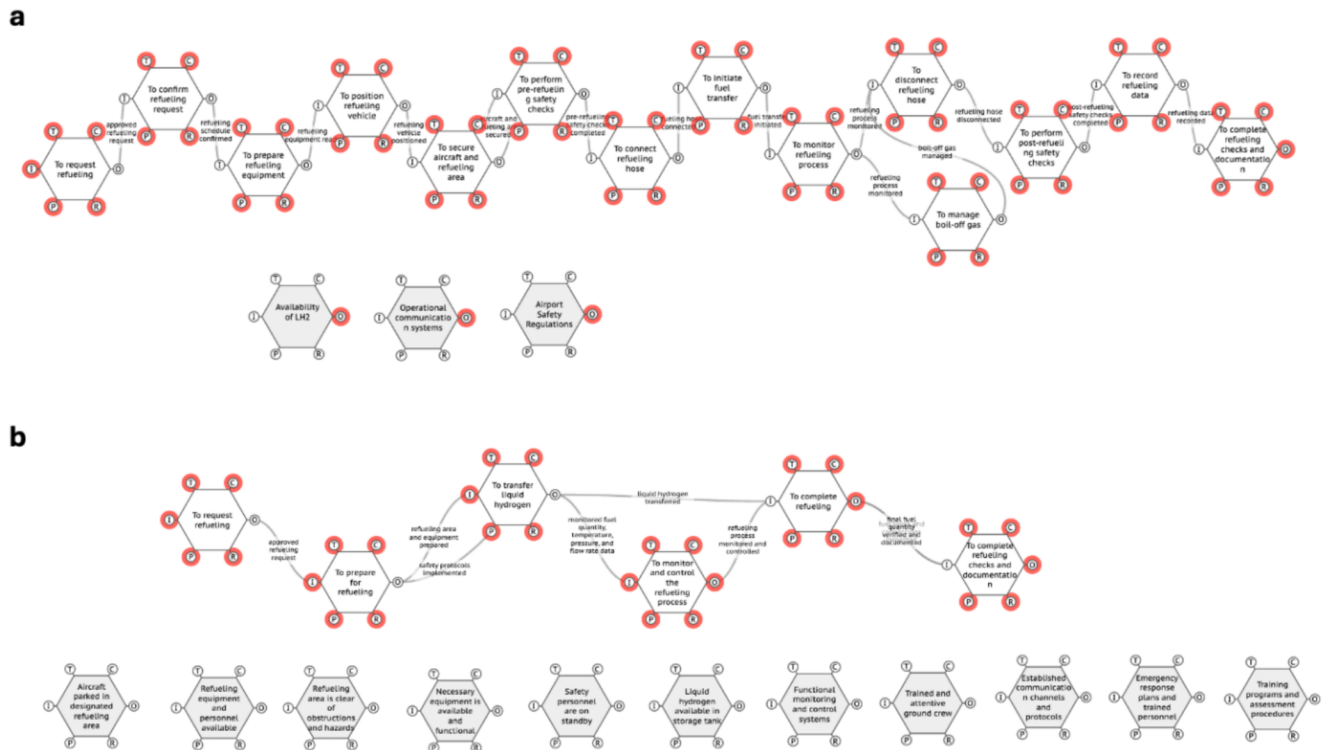


Fig. 4. The initial FRAM models generated by ChatGPT (a), and Gemini (b) outputs.

significantly impacting turnaround time and increasing pressure on subsequent functions, potentially leading to rushed procedures and increased safety risks."

ChatGPT provided examples for amplified and dampening impacts, whereas Gemini only provided examples for amplifying variability impacts. Both models identified multiple effects for each coupling and assessed the overall impact on other functions. Additionally, both identified simultaneous variability or combined effects, although only amplifying impacts were observed, and Gemini provided generic

statements. Neither model could continue the findings from prompt 3, leading to inconsistencies in the number of variability conditions considered.

3.1.4. Step 4: consequence analysis (prompt 7)

ChatGPT identified 42 recommendations for managing undesired variability, 14 recommendations for enhancing positive outcomes, and 4 additional considerations for managing aggregated variability. Despite ChatGPT's identified variabilities in terms of "on time" and acceptable,

the follow-up prompts did not build on that. That is why the recommendations for enhancing positive outcomes were not necessarily recommendations to sustain desirable variability; they were generic recommendations to ensure the successful operation of the given function, which was the same with Gemini. Gemini identified 21 recommendations for mitigating undesired variability, 11 for sustaining desirable variability, and 4 generic recommendations. Table 2 provides selective outputs of ChatGPT and Gemini for managing variabilities. Despite useful recommendations, both LLMs provided recommendations for the function rather than focusing on the specific aggregated variability scenarios. LLMs could have generated some of those recommendations if a single prompt had been executed to create safety recommendations without applying any methods.

3.2. STPA findings

3.2.1. Step 1: define the purpose of the analysis (prompts 2, 3 and 4)

In Prompt 2 outputs, ChatGPT and Gemini identified losses related to hydrogen storage and handling. ChatGPT provided 8 high-level and 26 sub-level losses, whilst Gemini provided 4 high-level and 15 sub-level losses. Both LLMs identified loss of human life, environmental damage/losses, and economic losses as high-level losses. Both identified unique losses, such as ChatGPT identifying loss of operational continuity, regulatory and compliance losses, and social and psychological losses, and Gemini identifying loss of sensitive information and delays in the adoption and development of hydrogen aviation.

In Prompt 3 outputs, ChatGPT identified 10 hazards, while Gemini identified 5 high-level and 15 sub-level hazards. Both provided a comprehensive list of hazards, including a range of hazards, such as physical risks, procedural inefficiencies, and environmental concerns; however, Gemini presented hazards in a more narrative, sentence-like format, whereas ChatGPT used a more structured breakdown (e.g., separating “System”, “Unsafe Condition” and “Link to Losses”). Also, Gemini identified sub-hazards and also hazards that are related to external factors and uncertainties, such as “Adverse weather conditions (e.g., strong winds, lightning) impacting refuelling operations”, “Unauthorized access or sabotage of the refuelling process or

Table 2

LLM output examples for recommendations.

ChatGPT	Gemini
<p>Recommendations for managing variabilities on the outputs from the function <To initiate fuel transfer></p> <p>Mitigation of Adverse Outcomes:</p> <p>Enhancing Positive Outcomes:</p> <ul style="list-style-type: none"> Automated Fuel Transfer Systems: Use automated systems that initiate fuel transfer based on predefined safety conditions being met. Redundant Systems: Implement redundant checks to ensure that all conditions are ideal before initiating fuel transfer. Emergency Shut-Off: Ensure an emergency shut-off system is in place and tested regularly to quickly stop fuel transfer if any anomalies are detected. Smooth Transfer Protocols: Develop protocols that allow for smooth initiation of fuel transfer, minimizing delays and ensuring optimal fuel flow. 	<p>Recommendations for managing variabilities on the outputs from the function <To transfer liquid hydrogen></p> <p>Mitigate unwanted variability:</p> <p>Sustain desirable variability:</p> <ul style="list-style-type: none"> Employ advanced sensor technology and real-time monitoring systems to detect leaks, pressure fluctuations or other anomalies early Implement automated shut-off and safety systems to prevent overflow, excessive pressure, or other hazardous conditions. Conduct rigorous pre-transfer checks of equipment and connections to ensure integrity and minimize the risk of leaks or malfunctions. Provide specialized training for ground crew on liquid hydrogen handling and emergency procedures. Explore the use of adaptive control systems that can adjust flow rates and pressures based on real-time conditions to optimize transfer efficiency and safety Encourage continuous monitoring and data analysis to identify opportunities for process improvement and further risk reduction

infrastructure”, and “Unexpected air traffic movements or ground vehicle intrusions near the refuelling area”. However, both LLMs included causal factors and recursive wording in hazard identification.

In Prompt 4 outputs, each LLM provided one system-level constraint linked to its corresponding hazard. ChatGPT was more effective in structuring system constraints as it followed the prompted format. Table 3 provides examples of losses, hazards and system-level constraints identified by ChatGPT and Gemini.

3.2.2. Step 2: model control structure (prompts 5 and 6)

ChatGPT identified 13 responsibilities for 5 high-level controllers and 12 for five sub-system-level controllers, refining 25 Control Actions (CAs). Gemini identified 15 responsibilities for 5 high-level controllers

Table 3

LLM output examples for identifying losses, hazards and system-level constraints.

ChatGPT Findings	Gemini Findings
<p>Losses (L)</p> <p>L-1 Loss of Life or Human Injury</p> <ul style="list-style-type: none"> L1.2: Fatalities or injuries to ground crew or refuelling personnel due to exposure to liquid hydrogen or its hazards (e.g., extreme cold, fire, explosion). <p>L7. Regulatory and Compliance Losses</p> <ul style="list-style-type: none"> L7.2: Delays in the certification or approval of new hydrogen-powered aircraft or refuelling infrastructure due to safety concerns. <p>Hazards (H)</p> <p>H3: Failure of Temperature Control Leading to Hydrogen Gasification</p> <ul style="list-style-type: none"> System: Liquid hydrogen storage and transfer system Unsafe Condition: Failure to maintain cryogenic temperatures Link to Losses: (L1, L2, L3, L4, L5, L6) <p>If the cryogenic temperature is not maintained, liquid hydrogen may gasify rapidly, leading to a pressure build-up and potential explosion. This can cause loss of life, injuries, environmental contamination, damage to infrastructure, and significant operational and economic impacts.</p> <p>System-level Constraints (SC)</p> <p>SC-3: Temperature Control Systems Must Maintain Cryogenic Temperatures During Hydrogen Storage and Transfer</p> <ul style="list-style-type: none"> System: Liquid hydrogen storage and transfer system Condition to Enforce: Ensure continuous monitoring and control of temperatures to keep hydrogen in its liquid state Link to Hazards: [H3] <p>This constraint mandates that the temperature of liquid hydrogen is continuously monitored and maintained to prevent gasification, which could lead to pressure build-up and potentially hazardous situations.</p>	<p>Losses (L)</p> <p>L1. Loss of Human Life or Human Injury</p> <ul style="list-style-type: none"> L1.1 Fatalities or injuries to ground crew during refuelling operations due to hydrogen leaks, fires, or explosions. <p>L4. Other Losses</p> <ul style="list-style-type: none"> L4.3 Delays in the adoption and development of hydrogen aviation technology due to safety concerns or incidents. <p>Hazards (H)</p> <p>H3: Cryogenic Hazards</p> <ul style="list-style-type: none"> H3.1 System: Refuelling Personnel & Unsafe Condition: Inadequate personal protective equipment (PPE) or training for handling cryogenic hydrogen & Link to Losses: (L1.1) H3.2 System: Refuelling Equipment & Unsafe Condition: Material embrittlement or failure of components due to exposure to cryogenic temperatures & Link to Losses: (L1.1, L3.2, L3.3) H3.3 System: Aircraft & Refuelling Infrastructure & Unsafe Condition: Condensation and ice formation leading to blockages, malfunctions, or structural damage & Link to Losses: (L1.1, L1.2, L3.1, L3.2, L3.3) <p>System-level Constraints (SC)</p> <p>SC3: Management of Cryogenic Hazards</p> <ul style="list-style-type: none"> SC3.1: Refuelling Personnel & Mandatory use of appropriate PPE and comprehensive training programs for handling cryogenic hydrogen. (H3.1) SC3.2: Refuelling Equipment & Selection of materials and components compatible with cryogenic temperatures, and implementation of thermal insulation and protection measures. (H3.2) SC3.3: Aircraft & Refuelling Infrastructure & Design and operational procedures to prevent and manage condensation and ice formation, including active heating or insulation. (H3.3)

and 8 for three sub-system-level controllers, refining 14 CAs. ChatGPT and Gemini identified 4 common high-level controllers: Civil Aviation Authority (CAA), Airport Authority, Airline, and Ground Handling Crew. ChatGPT identified responsibilities and control actions (CAs) for each controller listed, whilst Gemini provided responsibilities for all controllers but control actions for only the sub-level controllers. Both LLMs provided inconsistencies in the identification of CAs; they might be related to the responsibility, but not entirely or explicitly linked to the responsibilities, resulting in incomplete control structures.

In identifying feedback loops (Prompt 6), both LLMs identified feedback control loops. Table 4 provides examples from Prompt 5 and 6 outputs. ChatGPT identified feedback considering the process model needs and controllers. Gemini identified feedback in a less structured way. Building the control structure with Gemini outputs required more user interpretations.

The authors used the Prompt 6 findings from ChatGPT and Gemini to model the control structures (see Figs. 5 and 6, respectively). Despite ChatGPT offering insights from the process models, they were incomplete, and Gemini did not provide relevant information. Thus, the control structures did not include process models to ensure consistency and clarity of the Figures.

3.2.3. Step 3: identify unsafe control actions (prompt 7)

ChatGPT identified 36 UCAs from 5 controllers based on 9 CAs, while Gemini identified 22 UCAs for 14 CAs. However, both failed to effectively carry out this task as they struggled to generate and link UCAs to all hazards, and not all CAs were considered when generating UCAs.

Gemini was inconsistent in labelling the UCAs, which could lead to further confusion in the analysis. Gemini failed to plausibly link UCAs to hazards. For example, UCA1.1: Supervisor fails to monitor critical parameters (e.g., pressure, temperature) during refuelling is linked with hazards: “H1.1. Aircraft fuel system and refuelling infrastructure leak or rupture in hydrogen lines, tanks, or couplings during transfer or

Table 4

LLM outputs examples for controllers, responsibilities, control actions and feedback loops.

ChatGPT Findings	Gemini Findings
Controller: Refuelling Staff Responsibilities: <ul style="list-style-type: none"> R-1: Operate the refuelling equipment safely and efficiently [Linked to SC-1, SC-3]. Control Actions: <ul style="list-style-type: none"> CA-1: Connect and disconnect refuelling lines according to established procedures [Linked to R-1]. 	<ul style="list-style-type: none"> Controller: Refuelling Staff Responsibilities: <ul style="list-style-type: none"> Operate refuelling equipment and connect/disconnect hoses and couplings. (SC1, SC3) Control Actions: <ul style="list-style-type: none"> CA1: Connect and disconnect hoses and couplings according to procedures.
Feedback Control Loops Refuelling Staff <ul style="list-style-type: none"> Process Model Needs: <ul style="list-style-type: none"> Current status of refuelling equipment, hydrogen levels, and environmental safety. Feedback and Information Flows: <ul style="list-style-type: none"> From Equipment Sensors: Real-time data on hydrogen flow rates, pressure, temperature, and leak detection. From Supervisor: Operational directives and safety instructions based on the overall refuelling status. To Supervisor: Continuous feedback on refuelling progress and any equipment issues encountered. 	Feedback Between Controllers and Controlled Processes <ul style="list-style-type: none"> Controllers ↔ Refuelling Equipment & Aircraft Fuel System: <ul style="list-style-type: none"> Visual inspections, gauge readings, sensor data, and alarms provide real-time feedback on system status, enabling controllers to monitor progress, identify potential issues, and take corrective action.

storage”, “H1.3. Refuelling Infrastructure & Unsafe Condition: Malfunction or failure of safety relief valves or pressure control systems”, and “H2.2. Aircraft fuel system overfilling or over-pressurisation of hydrogen tanks leading to rupture and release”. However, it is not readily apparent how the omission of critical parameter monitoring would cause malfunction or failure of safety relief valves (H1.3).

In identifying controller constraints (CCs), ChatGPT identified 18 CCs with multiple UCAs linked to each CC by omitting to link several other UCAs. In general, there was no perfect one-to-one correspondence between UCAs and CCs. For instance, ChatGPT could not identify a CC for UCA2.1: Refuelling staff do not start hydrogen flow when the aircraft is ready, leading to delays that could affect the turnaround time or the safety of the operation [Linked to H10].

Meanwhile, Gemini identified 15 CCs with multiple UCAs linked to each CC identified in the previous prompt. Gemini, however, continued the analysis by focusing on 3 controllers (supervisor, refuelling staff, and other ground crew), which was also the case when identifying UCAs. Table 5 provides example outputs from ChatGPT and Gemini.

3.2.4. Step 4: identify loss scenarios (prompts 9 and 10)

ChatGPT identified 18 loss scenarios for 9 UCAs by describing the scenario and identifying causal factors and sequence events. For example, ChatGPT identified a scenario as follows:

Scenario 1.2.2: The refuelling staff receive incorrect feedback indicating that the aircraft is ready when it is not, due to a failure in the communication system [Linked to UCA-1.2, H1, H6].

- Causal Factors: Inadequate feedback; communication system failure.
- Sequence of Events: Due to a communication error between the aircraft and ground systems, refuelling staff believe the aircraft is ready and proceed with the connection, potentially causing a hazardous situation.

Although the scenarios were well-defined, there was an opportunity for ChatGPT to identify scenarios for the remaining 27 UCAs. Moreover, additional scenarios could have been generated for each UCA. Both identified two scenarios for a single UCA, as it was prompted to identify scenarios considering unsafe controller behaviour and improper or no execution of control action.

Gemini identified 40 loss scenarios for 20 UCAs (among 22 UCAs). Gemini performed better in generating scenarios based on the number of UCAs identified. For example, Scenario 1 for UCA3.2 is “The refuelling staff notices a slight pressure fluctuation but decides not to report it to the supervisor, assuming it is a minor issue. The fluctuation later escalates into a significant leak (H1.1).”, which is considered unsafe controller behaviour. Scenario 2 is “The refuelling staff is hesitant to interrupt the supervisor, who is busy communicating with the flight crew and delays reporting a temperature anomaly, leading to a potential safety hazard (H3.2).”, which is considered no execution of the control action of “Monitor pressure and temperature gauges and report any anomalies.”

Prompt 10 required the generation of safety recommendations based on loss scenarios and considering SCs and CCs. ChatGPT provided 12 recommendations based on loss scenarios (out of 18), 10 for SCs (one per each SC) and 11 for 17 CCs (among 18 CCs). In total, ChatGPT provided 33 recommendations. Table 6 lists some of the loss scenarios and recommendations provided by ChatGPT.

Gemini performed comparatively poorly. When tasked to generate recommendations based on loss scenarios, SCs and CCs, Gemini primarily focused on the sub-level controllers and linked them to their UCAs. While the recommendations were generally valid and informative, they lacked a direct link to individual loss scenarios. Recommendations were almost not specific to STPA findings. Gemini produced 15 recommendations for 5 of the 15 SCs and 10 recommendations for CCs, totalling 37 recommendations. Table 7 lists some of the recommendations provided by Gemini.

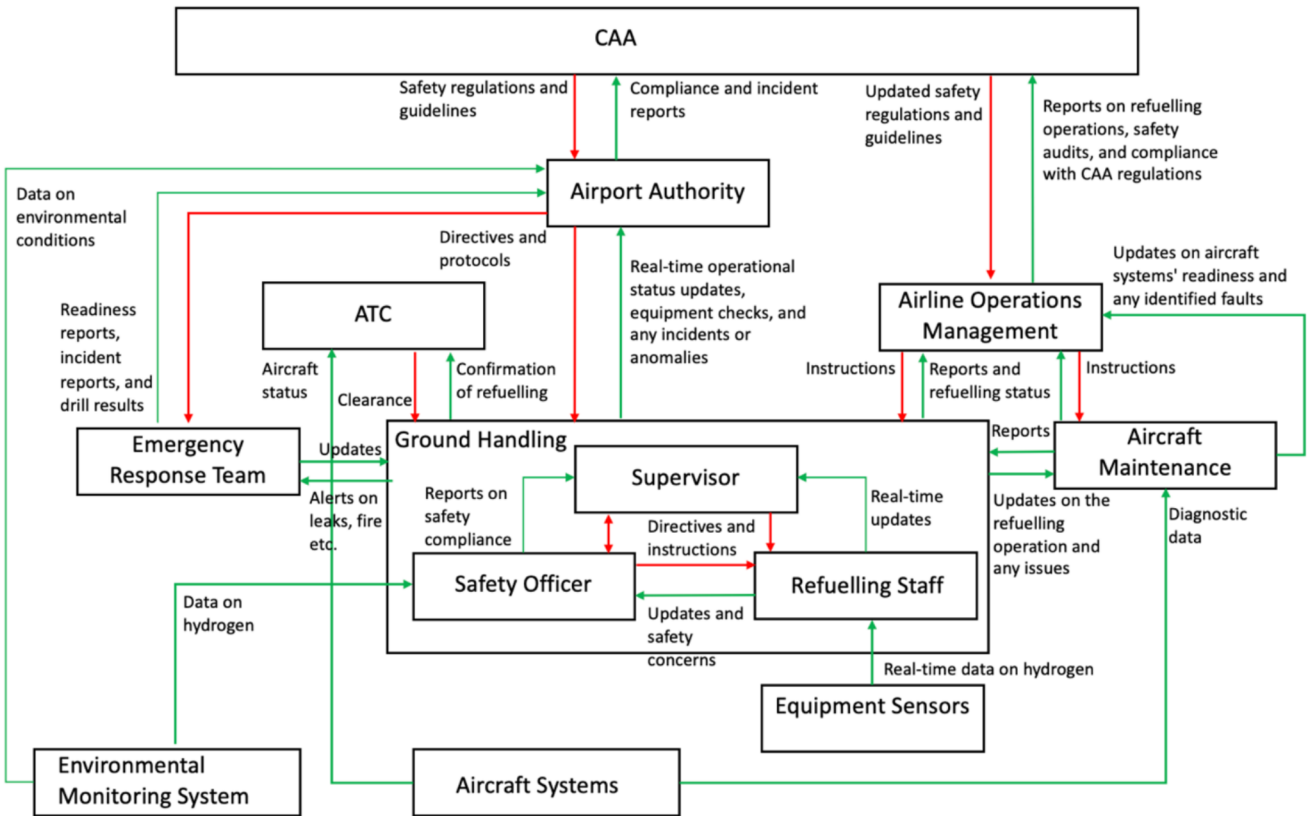


Fig. 5. Control structure modelled using ChatGPT outputs.

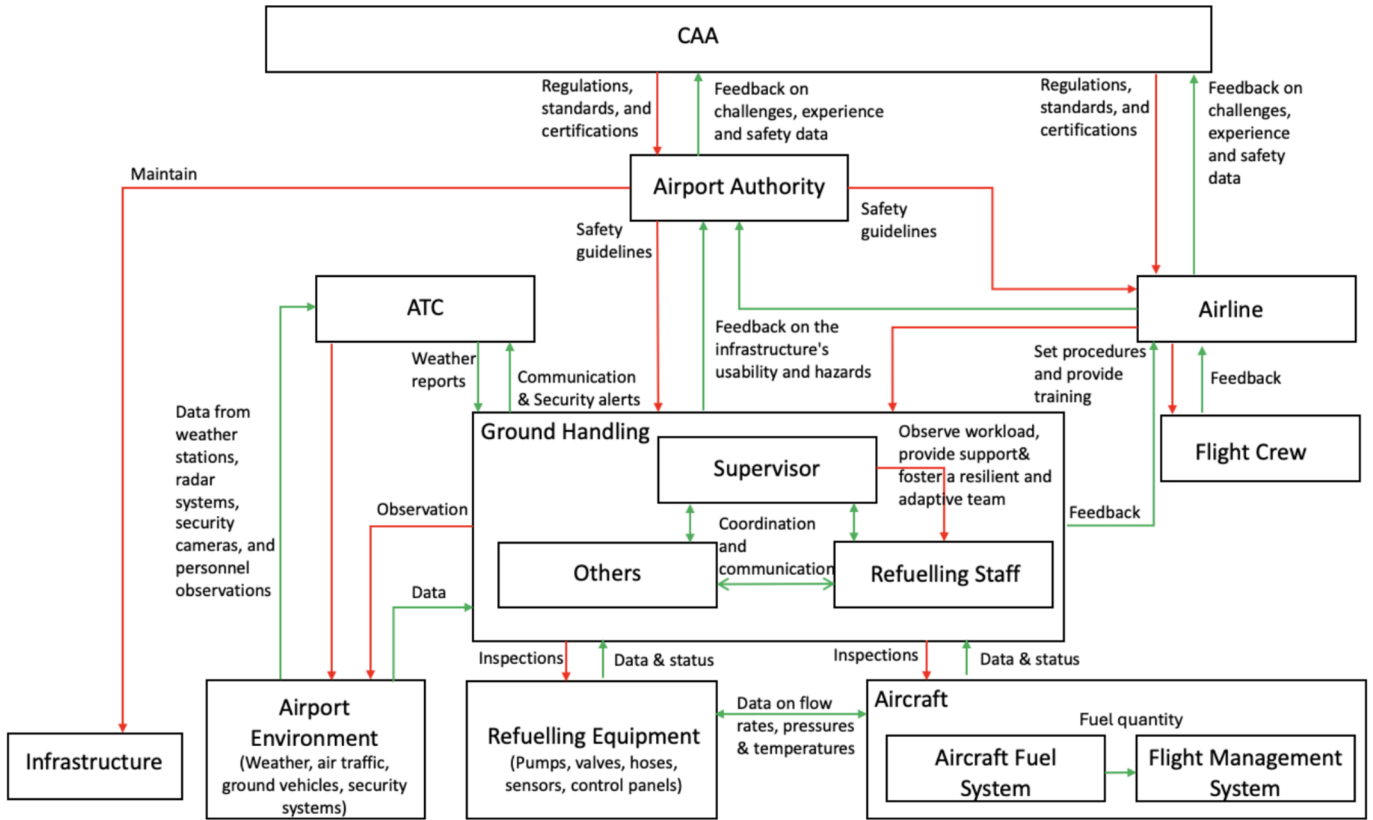


Fig. 6. Control structure modelled using Gemini output.

Table 5
LLM output examples for UCAs and CCs.

ChatGPT Findings	Gemini Findings
<p>1. Refuelling Staff Control Actions</p> <p>Control Action: Connect Refuelling Equipment to Aircraft</p> <ul style="list-style-type: none">UCA-1.1: Refuelling staff do not connect refuelling equipment to the aircraft when the aircraft is ready for refuelling and hydrogen levels are within safe limits [Linked to H1, H6].UCA-1.2: Refuelling staff connect refuelling equipment to the aircraft when the aircraft's systems are not ready, or there are detected anomalies in the hydrogen system [Linked to H1, H6].UCA-1.3: Refuelling staff connect the equipment before verifying the readiness of the aircraft or environmental conditions, leading to premature hydrogen flow [Linked to H1, H6].UCA-1.4: Refuelling staff disconnect the refuelling equipment too soon, before the process is completed, causing a potential leak or hydrogen spill [Linked to H1, H3]. <p>Refuelling Staff Controller Constraints</p> <ul style="list-style-type: none">CC-1: Refuelling staff must connect refuelling equipment to the aircraft only when the aircraft is confirmed to be ready for refuelling and hydrogen levels are within safe limits [Linked to UCA-1.1, UCA-1.2].	<p>Refuelling Staff</p> <p>CA1: Connect and disconnect hoses and couplings according to procedures</p> <ul style="list-style-type: none">UCA1.1: Refuelling staff fails to follow proper connection/disconnection procedures [H1.1]UCA1.2: Refuelling staff damages hoses or couplings during handling [H1.1] <p>Refuelling Staff Controller Constraints</p> <ul style="list-style-type: none">CC1.1: Refuelling staff must strictly adhere to established procedures for connecting and disconnecting hoses and couplings. [UCA1.1, UCA1.2]

4. Discussion

In this research, ChatGPT and Gemini were prompted to conduct FRAM and STPA applications for the given case study, and their performance was comparatively analysed. Overall, this study showed that ChatGPT performed better than Gemini regarding following FRAM and

STPA steps. Gemini’s architecture focuses on short, direct and general responses due to its emphasis on factual precision and brevity. In contrast, ChatGPT tends to be more verbose and provides greater detail, with a better capability to provide tailored responses based on past interactions within the same conversation due to its adeptness at capturing subtle nuances (Rane et al., 2024). Collier et al. (2024) claimed that ChatGPT performs well in divergent thinking tasks (e.g., brainstorming failure modes); however, it also claimed the outputs were too generic despite being a good starting point. Our findings revealed that while Gemini provided greater details on some FRAM analysis tasks, ChatGPT provided a more comprehensive and detailed analysis of most STPA outputs. This could be due to the STPA method being more structured, and so prompts being more structured. However, it is difficult to reach a single conclusion.

In general, Gemini provided more limited outputs than ChatGPT, as there was scope to produce more comprehensive outputs. ChatGPT outperformed Gemini in most FRAM (Prompts 2, 3, 5, 6 and 7) and STPA (Prompts 2, 4, 5, 7 and 10) prompts. However, Gemini produced better outputs in a few prompts or tasks in prompts. Considering the STPA outputs, Gemini was able to identify hazards related to external factors and uncertainties (Prompt 3), link controller constraints to all UCAs (Prompt 8) and generate more loss scenarios (Prompt 9). Considering FRAM outputs, Gemini was able to identify more background functions and couplings (Prompt 4) and identify variability sources (Prompt 5). Gomez-Cabello et al. (2024) found that Gemini was shown to have greater conciseness than ChatGPT, whereas several studies have shown that ChatGPT outperforms Gemini in providing more specific answers and higher accuracy (Carlà et al., 2024; Gomez-Cabello et al., 2024). In the FRAM analysis, Gemini identified more detailed function aspects. However, both failed in identifying all couplings between functions, as shown in Fig. 4. These missing interdependencies illustrate a core limitation in the models’ ability to adopt a systems-thinking approach, which requires understanding how functions interact dynamically rather than operating in isolation. To demonstrate this, the authors presented the completed FRAM models in Figs. 7 and 8. In both FRAM and STPA, the outputs lacked in revealing functional couplings and feedback-control loops. This limitation aligns well with previous findings by Ziakkas and Pechlivanis (2023) where ChatGPT performed

Table 6
Example ChatGPT outputs for loss scenarios and recommendations.

ChatGPT Findings
<p>Refuelling Staff Loss Scenarios</p> <p>UCA-1.1: Refuelling staff do not connect refuelling equipment to the aircraft when the aircraft is ready for refuelling and hydrogen levels are within safe limits.</p> <ul style="list-style-type: none">Scenario 1.1.1: The refuelling staff’s process model incorrectly indicates that the aircraft is not ready for refuelling due to faulty sensors or communication errors, leading to the decision not to connect the equipment [Linked to UCA-1.1, H1, H6].<ul style="list-style-type: none">Causal Factors: Inadequate process model; incorrect feedback from aircraft systems.Sequence of Events: Sensors on the aircraft erroneously report an issue with the hydrogen system, leading the staff to delay refuelling. The delay could result in operational inefficiencies or missed turnaround time. <p>Safety Recommendations Based on Loss Scenarios</p> <p>Loss Scenario 1.1.1: Refuelling staff’s process model incorrectly indicates that the aircraft is not ready for refuelling due to faulty sensors or communication errors.</p> <ul style="list-style-type: none">Recommendation 1: Implement redundancy in sensor systems and ensure regular calibration and maintenance of sensors to provide accurate feedback to refuelling staff. [Linked to UCA-1.1, SC-1, CC-1] <p>Loss Scenario 1.1.2: Refuelling staff are distracted or lack adequate training.</p> <ul style="list-style-type: none">Recommendation 2: Establish regular training and certification programs for refuelling staff to maintain high levels of awareness and procedural adherence, even under operational pressure. [Linked to UCA-1.1, CC-1, CC-8] <p>Safety Recommendations Based on System-level Constraints (SCs)SC-3: Temperature control systems must maintain cryogenic temperatures during hydrogen storage and transfer.</p> <ul style="list-style-type: none">Recommendation 15: Introduce redundant temperature control systems and regular maintenance checks to prevent failures that could lead to hydrogen gasification. [Linked to SC-3, UCA-1.4, UCA-2.4] <p>Safety Recommendations Based on Controller Constraints (CCs)</p> <p>CC-1: Refuelling staff must connect refuelling equipment only when the aircraft is ready.</p> <ul style="list-style-type: none">Recommendation 23: Introduce mandatory pre-connection checklists and interlocks that require confirmation from multiple systems before allowing connection. [Linked to CC-1, UCA-1.1]

Table 7
Example Gemini outputs for loss scenarios and recommendations.

Gemini Findings
Refuelling Staff Loss Scenarios UCA1.1: Refuelling staff fails to follow proper connection/disconnection procedures Scenario 1: The refuelling staff is rushing to complete the task and neglects to properly tighten a coupling, leading to a hydrogen leak during the transfer (H1.1). Scenario 2: The refuelling staff is unfamiliar with the specific aircraft model and connects the hoses incorrectly, causing damage to the fuel system and a potential leak (H1.1).
Safety Recommendations Based on Loss Scenarios Refuelling Staff: Provide comprehensive hands-on training for refuelling staff, covering proper connection/disconnection procedures, valve operation, and gauge monitoring. [Linked to Loss Scenarios related to UCA1.1, UCA1.2, UCA2.1, UCA2.2] <ul style="list-style-type: none">• Implement a buddy system or peer-checking mechanism to reduce the risk of errors or omissions during critical tasks. [Linked to Loss Scenarios related to UCA3.1, UCA3.2]• Use clear and standardized visual aids and alarms to enhance situational awareness and prompt responses to anomalies. [Linked to Loss Scenarios related to UCA3.1, UCA3.2]
Safety Recommendations Based on System-level Constraints (SCs) SC3: Management of Cryogenic Hazards: <ul style="list-style-type: none">• Select materials and components for the refuelling system that are resistant to embrittlement and maintain their structural integrity at cryogenic temperatures.• Apply advanced insulation and thermal management techniques to minimise heat transfer and prevent ice formation on critical components.• Provide continuous monitoring of cryogenic systems and implement automatic shutdown or alarm systems in case of abnormal temperature or pressure conditions.
Safety Recommendations Based on Controller Constraints (CCs) Refuelling Staff <ul style="list-style-type: none">• Provide refresher training on connection/disconnection procedures and emphasize the importance of following them meticulously. [Linked to CC1.1]• Use color-coded or keyed couplings and hoses to prevent incorrect connections and minimize the risk of damage. [Linked to CC1.1]• Install automated valve control systems with interlocks to prevent rapid or improper valve operation. [Linked to CC2.1]

better with linear models like bow-tie compared to HFACS. This suggests that LLM chatbots might perform better with traditional tools, as the existing literature and practice are predominantly based on traditional safety assessment tools and techniques. Consequently, LLM will prioritise the established traditional thinking and application.

However, [Sujan et al. \(2025\)](#) found that the model outputs could consider the interactions between functions and consider systems thinking when prompted accordingly. Again, this shows the inconsistent findings in the literature, possibly due to the peculiarity of the assessed processes and the tools used. While using an LH₂ refuelling case study provided the findings as described in this paper, a different case could have led to different conclusions, as in [Sujan et al. \(2025\)](#), who analysed an established clinical process, where we can assume that the LLM would have had ready access to relevant prior data and the method has been widely used in the healthcare domain. This suggests that LLM chatbots might provide better findings with known systems.

With FRAM applications, it is crucial to analyse how things could go right and identify potential trade-offs made by the operators to ensure safe operations. In this study, both models failed to reveal those trade-

offs. However, it should be acknowledged that the case study was challenging as there is no everyday practice on refuelling of LH₂-powered aircraft. Yet, ChatGPT could still generate variability examples of being on time, acceptable, and precise, which leads to considerations of how things could go right.

Considering the LLM’s performance on safety assessment, the STPA outputs were more comprehensive in both models. This could be due to the nature of the method and the fact that STPA is a more systematic method and closer to traditional thinking in comparison to FRAM. While the FRAM analysis produced comprehensive variability scenarios, the STPA offered a more holistic perspective on the analysis by incorporating a wider range of stakeholders, including regulatory and airport authorities. This oversight is significant as these stakeholders influence functions and activities, particularly in WAD scenarios. However, it should be noted that the FRAM analysis was limited by giving the first and final functions to be analysed. STPA also provided more relevant recommendations. The STPA recommendations focused on human–human and human–machine interactions and mostly on preventing humans from making unsafe control actions, in addition to equipment

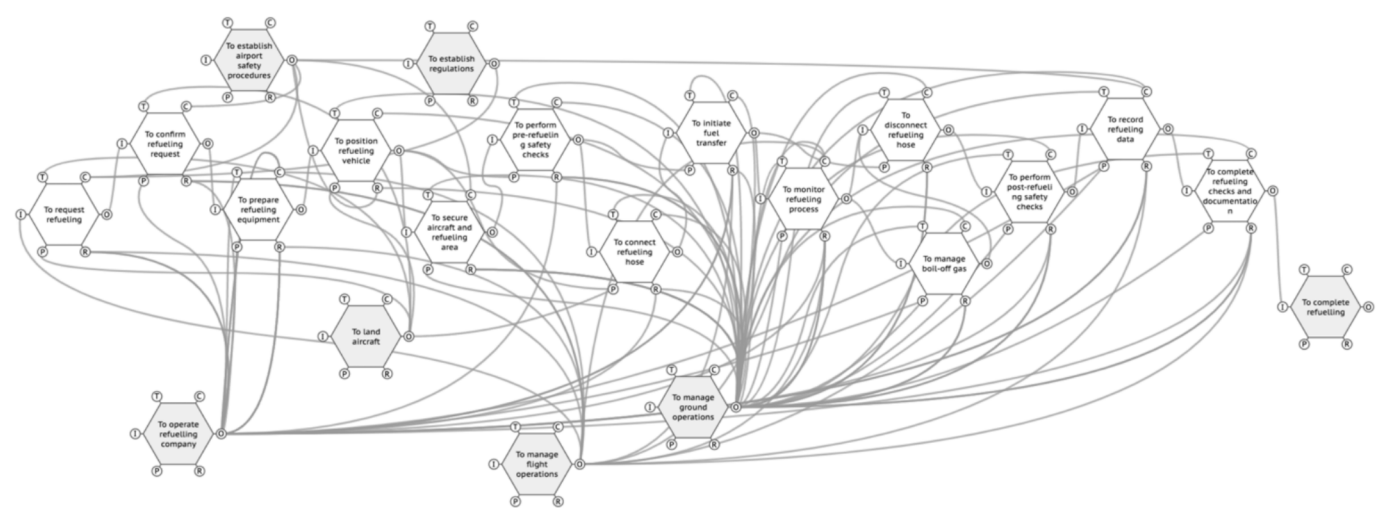


Fig. 7. Completed FRAM model using ChatGPT outputs.

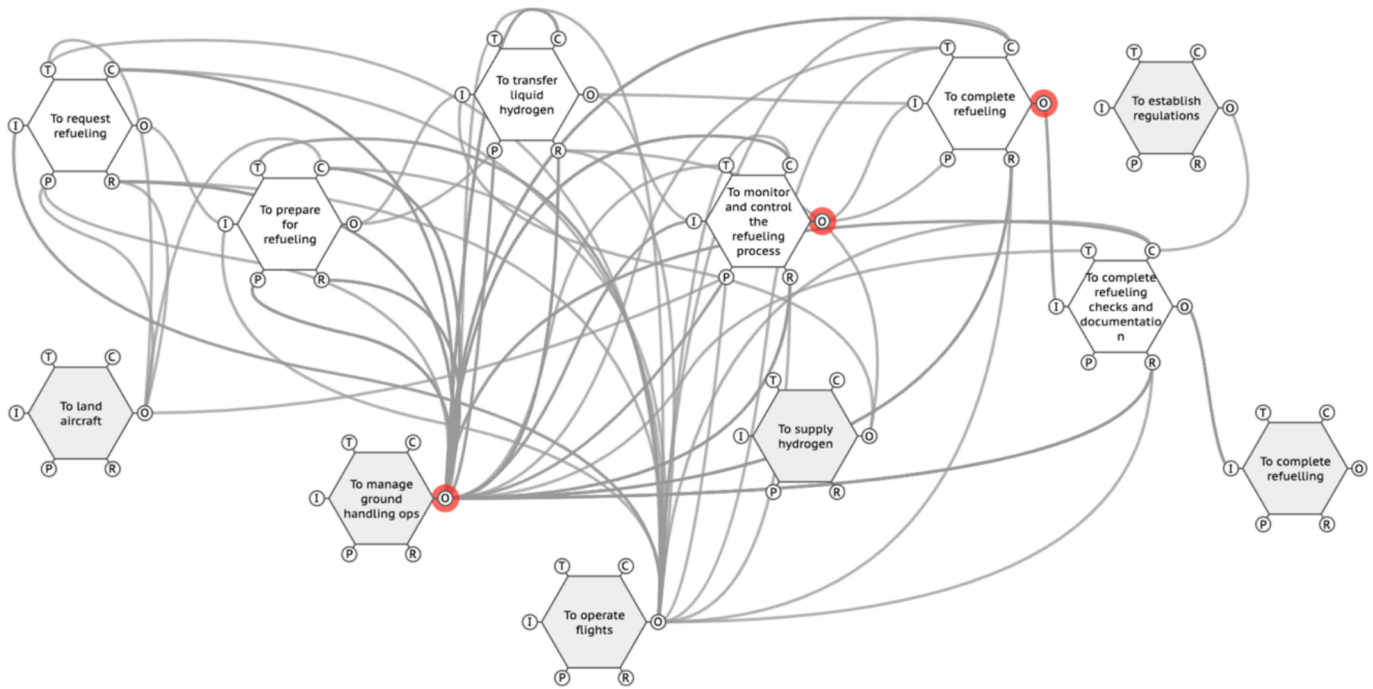


Fig. 8. Completed FRAM model using Gemini outputs.

and individual component-related recommendations. Meanwhile, the FRAM recommendations focused mainly on the equipment, ensuring that the equipment is ready, automation is implemented to streamline the process, and protocols and procedures are in place in order, which reflected its identification of variability in time and precision. FRAM's recommendations utilised more generic system descriptions.

In this study, ChatGPT and Gemini offered several advantages to support system-based safety assessments. One of them is that they can generate large amounts of useful information, which is acknowledged by Charalampidou et al. (2024) and Qi et al. (2025). Others can be listed as the user interface of LLMs, which is easy and straightforward. LLMs generate outputs quickly and use different personas. LLMs can enhance human analysis; however, LLM findings can be unreliable when there is no human intervention, which is also highlighted by Qi et al. (2025). The need for human analyst involvement is highlighted in various studies (Collier et al., 2024; Sujan et al., 2025; Ziakkas and Pechlivanis, 2023). Our findings revealed that we are still not at the point of leaving the whole analysis to LLM.

Despite several benefits of using LLM in system safety, they exhibit several limitations. A key disadvantage of the LLMs is the time-consuming process of developing effective prompts for thorough analysis. This requires a moderate understanding of LLM performance and prompt engineering techniques. This research provides a structured prompt engineering approach for hazard analyses, specifically STPA and FRAM. However, the lack of standardised prompts or processes can make it challenging for novice analysts to develop effective prompts without a comprehensive understanding of LLM performance and prompt engineering. Another limitation of LLMs is their non-deterministic nature. The same prompt can produce varying results with each execution, affecting the number of elements (e.g., functions, losses and hazards) and the output content. This can lead to inconsistencies in the comprehensiveness of the generated information. Also, LLMs failed to consistently follow up on previous prompt outputs and did not revise prompt outputs based on given self-criticisms. Additionally, their outputs varied when prompted at different times or on different devices simultaneously, raising concerns about the reliability of LLM chatbots in system safety assessments.

The key takeaways from our experience can be listed as follows:

- Prompt engineering and quality assessment are essential: Prompt engineering techniques such as preparatory prompts, persona prompting, and step-by-step breakdowns are crucial for improving output quality. Standardised assessment processes should be followed to evaluate LLM's ability to follow the method's steps and the output quality.
- LLM chatbots' non-deterministic nature: LLM chatbots can produce different outputs for the same prompts when prompted at different times.
- Performance variability and challenges with systems thinking: ChatGPT performed better in structured, rule-based analyses like STPA, whereas Gemini demonstrated a closer-to-human approach in system-wide assessments such as FRAM. However, both struggled with systems thinking, favouring traditional, linear risk assessment approaches.
- LLMs' dependency on existing knowledge and dominated information: LLM chatbots might provide better findings with known systems and traditional tools.
- Human analyst involvement: Despite generating valuable insights, both LLMs failed to follow up on previous prompt outputs, requiring human oversight to ensure each step was followed, outputs were verified, and analytical integrity was maintained.
- Potential for future enhancement: While current LLMs exhibit limitations, improvements through fine-tuning models, integrating domain-specific datasets, and developing structured safety assessment frameworks could enhance their utility in system safety assessments.

4.1. Limitations

This study also has several limitations. A key limitation of this study is that it examined the performance of LLMs using a single-prompt, unaided approach. In practice, both FRAM and STPA are typically conducted iteratively and benefit from human judgment, contextual insight and the refinement of models over time. Applying LLMs in such a context would ideally involve multi-turn interaction, feedback, and co-construction with human experts. Our study does not replicate this

more reflective process. However, the single-shot design was intentional. With the increasing deployment of LLMs for generating first-draft outputs, templates or rapid analyses, sometimes with minimal oversight, it is important to understand how these models perform in such settings. While this approach does not reflect best practice, it offers insights into the kinds of responses and limitations that may arise when models are used in a “zero-shot” or automated capacity. We hope this work can serve as a foundation for future studies that examine human-in-the-loop interaction and iterative refinement.

Additionally, this study focused on the performance of LLMs on following up prompts to apply FRAM and STPA rather than focusing on the quality of the outputs. However, it was due to the focus of this study to analyse LLMs’ capability in stand-alone FRAM and STPA applications. Future studies can develop and use performance evaluation metrics that focus on the reliability of the outputs, completeness and corrections of the outputs, logical consistency and justification ability, and a specific focus on the recommendations regarding their impact and feasibility. Moreover, the exploration of LH₂ as an alternative aviation fuel is still in its early design stages, with limited technical data on its impact on operational safety, particularly within the refuelling process for aircraft. Consequently, the limited data availability might impact the performance of LLMs. The use of well-documented, less complex or complex but well-known systems could have led to different conclusions. Also, the study focused solely on the refuelling process, making it challenging for the hazard analyses and LLMs to capture the integration with broader airline and airport operations.

However, despite these limitations, this study contributes to both safety science and practice by providing empirical evidence on the comparative application of the two widely used LLM models within the context of real-world safety methods (STPA and FRAM), using a structured evaluation framework. The study directly addresses a growing trend in the development of GPT-based tools that aim to support and, in some cases, replace human analysts in system safety tasks by testing whether state-of-the-art LLMs can perform such analyses autonomously. The study also offers valuable insights on the limitations of LLMs and provides practical lessons and future research directions, guiding the integration of AI-driven methodologies in safety-critical domains.

5. Conclusion

This study examined the capabilities of ChatGPT-4o and Gemini Advanced in performing system-based hazard analyses using FRAM and STPA, which were applied to the LH₂ aircraft refuelling case study. The findings indicate that both LLMs can generate valuable safety-related insights but face significant limitations in applying systems thinking, ensuring consistency across analytical steps, and conducting a stand-alone comprehensive safety assessment.

ChatGPT demonstrated stronger performance in structured tasks, particularly in STPA, where predefined steps and logical structures facilitated its outputs. Conversely, Gemini considered broader system factors FRAM application. However, substantial manual intervention is still required to ensure completeness and coherence in the FRAM model. Despite these advantages, neither model proved reliable enough for stand-alone application in safety-critical domains.

Moving forward, further research is needed to develop standard prompting frameworks, explore fine-tuning approaches for LLMs in safety applications, and investigate hybrid models where human expertise and AI-driven analysis complement one another. In addition, future research can focus on making a comparative analysis of the performance of LLM depending on the selection of the tools and system to be assessed, including the use of linear methods with complex systems and complex systems with linear or well-known complex systems. The potential of LLMs in supporting safety assessments is promising, but their integration into safety-critical decision-making should be cautiously approached and supplemented by expert oversight.

CRedit authorship contribution statement

Gulsum Kubra Kaya: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Dominique Bovell:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mark Sujun:** Writing – review & editing, Validation, Supervision, Methodology. **Graham Braithwaite:** Writing – review & editing, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ssci.2025.106960>.

Data availability

Data supporting this study are openly available in [Appendices A1, A2, B1 and B2](#).

References

- Benson, C.M., Ingram, J.M., Battersby, P.N., Mba, D., Sethi, V., Rolt, A.M., 2019. An analysis of civil aviation industry safety needs for the introduction of liquid hydrogen propulsion technology. *Proc. ASME Turbo Expo.* 3, 1–11. <https://doi.org/10.1115/GT2019-90453>.
- Buikstra, E., Strivens, E., Clay-Williams, R., 2020. Understanding variability in discharge planning processes for the older person. *Safety Sci.* 121 (December 2018), 137–146. <https://doi.org/10.1016/j.ssci.2019.08.026>.
- Carlà, M.M., Gambini, G., Baldascino, A., Boselli, F., Giannuzzi, F., Margollicci, F., Rizzo, S., 2024. Large language models as assistance for glaucoma surgical cases: a ChatGPT vs. Google Gemini comparison. *Graefes Arch. Clin. Experiment. Ophthalmol.* 1–15. <https://doi.org/10.1007/S00417-024-06470-5/FIGURES/3>.
- Casheekar, A., Lahiri, A., Rath, K., Prabhakar, K.S., Srinivasan, K., 2024. A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: applications, open challenges and future research directions. *Comput. Sci. Rev.* 52 (July 2023), 100632. <https://doi.org/10.1016/j.cosrev.2024.100632>.
- Castilho, D.S., Urbina, L.M.S., de Andrade, D., 2018. STPA for continuous controls: a flight testing study of aircraft crosswind takeoffs. *Saf. Sci.* 108, 129–139. <https://doi.org/10.1016/j.ssci.2018.04.013>.
- Charalampidou, S., Zeleskidis, A., Dokas, I.M., 2024. Hazard analysis in the era of AI: assessing the usefulness of ChatGPT4 in STPA hazard analysis. *Saf. Sci.* 178 (July), 106608. <https://doi.org/10.1016/j.ssci.2024.106608>.
- Chatzimichaïlidou, M.M., Ward, J., Horberry, T., Clarkson, P.J., 2018. A Comparison of the Bow-Tie and STAMP approaches to reduce the risk of surgical instrument retention. *Risk Anal.* 38 (5), 978–990. <https://doi.org/10.1111/risa.12897>.
- Chen, B., Zhang, Z., Langrené, N., Zhu, S., 2023. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. <https://arxiv.org/abs/2310.14735v4>.
- Collier, Z.A., Gruss, R.J., Abrahams, A.S., 2024. How good are large language models at product risk assessment? *Risk Anal.* 1–24. <https://doi.org/10.1111/risa.14351>.
- Damoiseaux-Volman, B.A., Medlock, S., van der Eijk, M.D., Romijn, J.A., Abu-Hanna, A., van der Velde, N., 2021. Falls and delirium in older inpatients: work-as-imagined, work-as-done and preferences for clinical decision support systems. *Safety Sci.* 142 (June 2020), 105355. <https://doi.org/10.1016/j.ssci.2021.105355>.
- Dghaym, D., Hoang, T.S., Turnock, S.R., Butler, M., Downes, J., Pritchard, B., 2021. An STPA-based formal composition framework for trustworthy autonomous maritime systems. *Saf. Sci.* 136 (January 2020). <https://doi.org/10.1016/j.ssci.2020.105139>.
- Ekin, S., 2023. Prompt engineering for CHATGPT: a quick guide to techniques, tips and best practices. *TechRxiv*. <https://doi.org/10.36227/techrxiv.22683919.v2>.
- Feretakis, G., Vervikios, V.S., 2024. Trustworthy AI: Securing Sensitive Data in Large Language Models. <https://arxiv.org/abs/2409.18222v1>.
- Gomez-Cabello, C.A., Borna, S., Pressman, S.M., Haider, S.A., Sehgal, A., Leibovich, B.C., Forte, A.J., 2024. Artificial intelligence in postoperative care: assessing large language models for patient recommendations in plastic surgery. *Healthcare (Switzerland)* 12 (11). <https://doi.org/10.3390/healthcare12111083>.
- Hoelzen, J., Flohr, M., Silberhorn, D., Mangold, J., Bensmann, A., Hanke-Rauschenbach, R., 2022. H2-powered aviation at airports – design and economics of LH2 refueling systems. *Energy Convers. Manag.* X 14 (November 2021), 100206. <https://doi.org/10.1016/j.ecmx.2022.100206>.

- Hollnagel, E., 2012. *FRAM: the Functional Resonance Analysis Method Modelling Complex Socio-Technical Systems*. Ashgate.
- Hollnagel, E., Wears, R., Braithwaite, J., 2015. From safety-I to safety-II: a white paper. *The Resilient Health Care Net*.
- Imran, M., Almusharraf, N., 2024. Google Gemini as a next generation AI educational tool: a review of emerging educational technology. *Smart Learn. Environ.* 11 (1). <https://doi.org/10.1186/s40561-024-00310-z>.
- Islam, R., Ahmed, I., 2024. Gemini-the most powerful LLM: Myth or Truth. 2024 5th Information Communication Technologies Conference, ICTC 2024, 303–308. doi: [10.1109/ICTC61510.2024.10602253](https://doi.org/10.1109/ICTC61510.2024.10602253).
- Jiang, P., Sonne, C., Li, W., You, F., You, S., 2024. Preventing the immense increase in the life-cycle energy and carbon footprints of LLM-powered intelligent chatbots. *Engineering* 40, 202–210. <https://doi.org/10.1016/j.eng.2024.04.002>.
- Kaya, G.K., 2021. A system safety approach to assessing risks in the sepsis treatment process. *Appl. Ergon.* 94 (October 2020), 103408. <https://doi.org/10.1016/j.apergo.2021.103408>.
- 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. <https://doi.org/10.1016/j.ssci.2019.06.020>.
- Leveson, N.G., Thomas, J.P., 2018. STPA handbook. http://psas.scripts.mit.edu/home/get_file.php?name=STPA_handbook.pdf.
- Lo, L.S., 2023. The CLEAR path: a framework for enhancing information literacy through prompt engineering. *J. Acad. Librariansh.* 49 (4), 102720. <https://doi.org/10.1016/j.jacalib.2023.102720>.
- Mangold, J., Silberhorn, D., Moebis, N., Dzikus, N., Hoelzen, J., Zill, T., Strohmayr, A., 2022. *Procedures and Aircraft Design Implication*. 1–39.
- Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A., 2023. A comprehensive overview of large language models. *ArXiv*. <https://arxiv.org/abs/2307.06435v9>.
- Nouri, A., Cabrero-Daniel, B., Törner, F., Sivencrona, H., Berger, C., 2024. Welcome Your New AI Teammate: On Safety Analysis by Leashing Large Language Models. *Proceedings - 2024 IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI, CAIN 2024*, 1, 172–177. doi: [10.1145/364481](https://doi.org/10.1145/364481).
- Patriarca, R., Chatzimichailidou, M., Karanikas, N., Di Gravio, G., 2022. The past and present of system-theoretic accident model and processes (STAMP) and its associated techniques: a scoping review. *Safety Sci.* 146 (December 2020), 105566. <https://doi.org/10.1016/j.ssci.2021.105566>.
- Patriarca, R., Di Gravio, G., Woltjer, R., Costantino, F., Praetorius, G., Ferreira, P., Hollnagel, E., 2020. Framing the FRAM: a literature review on the functional resonance analysis method. *Saf. Sci.* 129, 104827. <https://doi.org/10.1016/j.ssci.2020.104827>.
- Patterson, M., Deutsch, E.S., 2015. Safety-I, safety-II and resilience engineering. *Curr. Probl. Pediatr. Adolesc. Health Care* 45 (12), 382–389. <https://doi.org/10.1016/j.cppeds.2015.10.001>.
- Peng, B., Chen, K., Li, M., Feng, P., Bi, Z., Liu, J., Niu, Q., 2024. Securing Large Language Models: Addressing Bias, Misinformation, and Prompt Attacks. <https://arxiv.org/abs/2409.08087v1>.
- Pritchard, D.K., Rattigan, W.M., 2010. Hazards of liquid hydrogen Position paper. <https://www.hse.gov.uk/research/rrpdf/rr769.pdf>.
- Qi, Y., Zhao, X., Khastgir, S., Huang, X., 2025. Safety analysis in the era of large language models: a case study of STPA using ChatGPT. *Mach. Learn. Appl.*, 100622.
- Rane, N., Choudhary, S., Rane, J., 2024. Gemini versus ChatGPT: applications, performance, architecture, capabilities, and implementation. *J. Appl. Artif. Intell.* 5 (1), 69–93. <https://doi.org/10.48185/jaai.v5i1.1052>.
- Ray, P.P., 2023. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys. Syst.* 3 (March), 121–154. <https://doi.org/10.1016/j.iotcps.2023.04.003>.
- Robertson, J., Ferreira, C., Botha, E., Oosthuizen, K., 2024. Game changers: a generative AI prompt protocol to enhance human-AI knowledge co-construction. *Bus. Horiz.* <https://doi.org/10.1016/J.BUSHOR.2024.04.008>.
- Sadeghi, R., Goerlandt, F., 2023. A proposed validation framework for the system theoretic process analysis (STPA) technique. *Saf. Sci.* 162 (February), 106080. <https://doi.org/10.1016/j.ssci.2023.106080>.
- Salehi, V., Veitch, B., Smith, D., 2020. Modeling complex socio-technical systems using the FRAM: a literature review. *Hum. Factors Ergon. Manuf. (October)*, 1–25. <https://doi.org/10.1002/hfm.20874>.
- Sivakumar, M., Belle, A.B., Shan, J., Khakzad Shahandashti, K., 2024. Prompting GPT-4 to support automatic safety case generation. *Expert Syst. Appl.* 255, 124653. <https://doi.org/10.1016/J.ESWA.2024.124653>.
- Sujan, M., Slater, D., Crumpton, E., 2025. How can large language models do a FRAM analysis? *Saf. Sci.* 181, 106695. <https://doi.org/10.1016/j.ssci.2024.106695>.
- Sulaman, S.M., Beer, A., Felderer, M., Höst, M., 2019. Comparison of the FMEA and STPA safety analysis methods—a case study. *Softw. Qual. J.* 27 (1), 349–387. <https://doi.org/10.1007/s11219-017-9396-0>.
- Tamkin, A., Brundage, M., Ganguli, D., 2021. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. <https://arxiv.org/abs/2102.02503v1>.
- Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L.A., Rimell, L., Isaac, W., Gabriel, I., 2022. Taxonomy of risks posed by language models. *ACM Int. Conf. Proc. Series* 22, 214–229. <https://doi.org/10.1145/3531146.3533088>.
- Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.L., Tang, Y., 2023. A brief overview of ChatGPT: the history, status quo and potential future development. *IEEE/CAA J. Autom. Sin.* 10 (5), 1122–1136. <https://doi.org/10.1109/JAS.2023.123618>.
- Zhang, J., Zhang, W., Xu, P., Chen, N., 2019. Applicability of accident analysis methods to Chinese construction accidents. *J. Saf. Res.* 68, 187–196. <https://doi.org/10.1016/j.jsr.2018.11.006>.
- Zheng, Q., Liu, X., Yang, M., Wang, W., Adriaenssens, A., 2024. Enhancing emergency response planning for natech accidents in process operations using functional resonance analysis method (FRAM): a case of fuel storage tank farm. *Process Saf. Environ. Prot.* 188 (January), 514–527. <https://doi.org/10.1016/j.psep.2024.05.132>.
- Ziakkas, D., Pechlivanis, K., 2023. Artificial intelligence applications in aviation accident classification: a preliminary exploratory study. *Decis. Anal. J.* 9 (April), 100358. <https://doi.org/10.1016/j.dajour.2023.100358>.



DSPACE

<https://dspace.org/>

Large language models powered system safety assessment: applying STPA and FRAM

Kaya, Gulsum Kubra; Bovell, Dominique; Sujan, Mark; Braithwaite, Graham R.

2025-11

Elsevier

<https://doi.org/10.1016/j.ssci.2025.106960>

Kaya GK, Bovell D, Sujan M, Braithwaite G. (2025) Large language models powered system safety assessment: applying STPA and FRAM. Safety Science, Volume 191, November 2025, Article number 106960