

Generative AI Chatbots' Performance Across Different Domains of Industrial Safety Engineering

Nazurudin Bin Syed Ahmed, Mukesh Kumar Likhmawat Prashanth, Arockianathan Sam Michael, Wang Xinyu, Fan Keli and Loh Tzu Yang

Department of Chemical & Biomolecular Engineering, National University of Singapore (NUS)
21 Lower Kent Ridge Road, Singapore

e1499316@u.nus.edu, e0878475@u.nus.edu, e1499209@u.nus.edu, e1521235@u.nus.edu,
e1553529@u.nus.edu, tzuyang@nus.edu.sg

Abstract

This study evaluates the performance of four widely used generative AI chatbots (ChatGPT, Google Gemini, Grok, and Perplexity) in industrial safety engineering using a controlled comparative protocol. Nine standardized prompts (three safety domains × three prompt categories) were administered across Electrical, Mechanical, and Fire Safety, covering Fundamental Knowledge, Applied Reasoning, and Image-Based Analysis. Responses were assessed against pre-defined gold standards using a 1–5 rubric across three key performance indicators (accuracy, completeness, and relevance) by three independent evaluators, with scores averaged. Overall, the chatbots performed best on text-based prompts (Fundamental Knowledge: 88.67%; Applied Reasoning: 79.33%) and lowest on Image-Based Analysis (68.67%). By domain, performance was highest in Fire Safety (87.33%) and lowest in Electrical Safety (70.67%). Across KPIs, relevance and accuracy were consistently stronger than completeness, indicating that responses were generally aligned and correct but frequently lacked the depth expected for professional judgement. The findings suggest that AI chatbots can support HSE work in drafting, summarization, and preliminary decision support, but human oversight remains essential, especially for applied reasoning in safety-critical and visual interpretation tasks.

Keywords

AI chatbots; Industrial Safety Engineering; Fundamental Knowledge; Applied Reasoning, Image-Based Analysis.

1. Introduction

Industry 5.0 emphasizes human-centric technologies that enhance sustainability, resilience, and productivity in industrial systems (European Commission, 2021). Within this landscape, generative AI, particularly large language model (LLM) chatbots, offers potential value for occupational safety and health by accelerating information synthesis, drafting, and preliminary decision support. However, their use in safety-critical contexts remains under-examined, especially where errors, omissions, or misinterpretations can translate into real-world harm.

This study is motivated by the persistent global challenge of industrial accidents across electrical, mechanical, and fire safety domains, where significant operational risks remain despite technological, regulatory and safety advancements. Evidence from Singapore's Workplace Safety and Health (WSH) statistics show an increase from 2023 to 2024, with workplace fatalities rising from 36 in 2023 to 43 in 2024 at a rate of 1.2 per 100,000 workers (MOM, 2024). In 2024, one fatality was attributed to a Fire-related incident, which poses a serious workplace hazard and warrant targeted risk assessment and control efforts. Similarly, machinery-related hazards remain a significant contributor to injury risk in the manufacturing sector, with 52 fatal and major injury cases reported in machinery-related incidents (MOM, 2024).

Despite their growing application in Health, Safety, and Environment (HSE) profession, the reliability of AI chatbots in safety-critical applications remains uncertain, with risks related to accuracy, bias, and data integrity, particularly where misinformation may lead to serious injury or fatalities. This highlights the need for empirical evaluation of chatbot performance against various facets of industrial safety domains prior to widespread adoption. In addition, prior evaluations of LLMs in safety contexts often focus on single domains, single models, or purely text-based tasks. This study contributes a cross-domain, cross-model comparison using standardized prompts, locally grounded gold standards (Singapore regulatory context), and a structured KPI-based rubric that includes multimodal (image-based) safety interpretation.

1.1 Objectives

The objective of this research is to systematically evaluate and compare the performance of four widely used AI chatbots (ChatGPT, Google Gemini, Grok, and Perplexity) across three key domains of industrial safety engineering: Electrical Safety, Mechanical Safety, and Fire Safety. The research follows a structured framework (refer to Table 1) to ensure balanced evaluation and consistent assessment of the AI chatbots.

Table 1. Summary of Research Framework

Safety Domains considered:	Electrical Safety, Mechanical Safety, and Fire Safety
AI chatbots evaluated:	ChatGPT, Google Gemini, Grok, and Perplexity
Prompt Categories:	• P1: Fundamental Knowledge • P2: Applied Reasoning • P3: Image-Based Analysis
Key Performance Indicators (KPIs):	Accuracy, Completeness, and Relevance

2. Literature Review

2.1. Evolution and Capabilities of AI chatbots

Adamopoulou and Moussiades (2020) define a chatbot as a computer program designed to simulate human conversation using artificial intelligence (AI) and natural language processing (Adamopoulou & Moussiades, 2020). Powered by Large Language Models (LLMs), these systems are trained to recognize patterns to generate context-specific responses, mimicking human interaction. Ndlela (2024) notes that chatbots can bridge communication gaps between organizations and knowledge sources by providing employees with rapid access to information that previously required subject-matter experts.

Sherr et al. (2025) state that the landscape of generative AI has diversified and transformed significantly due to their growing popularity and competitiveness among a wide range of users (Scherr, Cao, & et al., 2025). According to Arash Heidari (2024) OpenAI's ChatGPT is widely recognized for its ability to automate workflows, generate reports, and summarize complex texts (Arash Heidari, 2024). Beyond ChatGPT, other chatbots such as Google Gemini use multimodal model to analyze imagery and videos without any specific training. According to Setzen (2025) et al. Google Gemini has the potential to improve any diagnostic accuracy and minimize human errors by identifying any sort of abnormalities and pinpoint narration (Setzen, 2025). According to Manju Lata (2025), GROK, powered by xAI advanced large language model, can understand and generate human like reports. Similarly, Perplexity is an AI based search engine that utilizes large language model technology that is used to answer various queries, and they have wide applications in scientific research to compile scientific articles efficiently (Manju Lata, 2025).

2.2. AI in Industrial Safety Engineering

Traditional risk assessment relies heavily on historical data, past incident records, and established analytical methods including Risk Matrices, Hazard and Operability Studies (HAZOP), Fault Tree Analysis (FTA), and Bow-Tie diagrams for evaluating the likelihood and severity of foreseeable risks. They are usually performed by designated teams comprising of risk assessment members who maintain, document, and update these risk registers. In contrast, the integration of AI chatbots in risk assessment represents a significant operational advancement by streamlining data

synthesis, optimizing speed and productivity of the risk assessment process, and reducing the manual labor associated with traditional risk assessment methods.

Bhatti et al. (2024) suggest that by training AI models with relevant safety standards, regulations, and historical data encompassing frequently encountered high-risk activities, organizations can strengthen proactive hazard control and risk management system (Bhatti, 2024). Zhu X. L (2022) has mentioned that in operational settings, AI-enhanced cameras are already used to detect Personal Protective Equipment (PPE) non-compliance in real-time (Zhu X. L., 2022). Specific to chatbots, Zhu & et al. (2022) indicates their utility in hazard identification by analyzing historical incidents and operational logs to develop preventive measures. Additionally, Khurram et al. (2025) highlight the role of chatbots like ChatGPT and Grok in incident reporting, documentation, and scenario-based training (Khurram, 2025). According to Hostetter (2024) Multimodal platforms like Gemini are further utilized for rapid risk assessments, emergency response coordination and situational awareness (Hostetter, 2024).

2.3. Limitations of AI chatbots

Despite the technological advancements of AI chatbots, significant limitations persist, particularly when applied imprudently in dynamic and high-risk industries. Iyengar et al. warn that AI chatbots often lack deep conceptual understanding and regulatory compliance. Furthermore, Anjum et al. (2025) has identified the critical risk of “hallucinations” – the generation of false or fabricated evidence due to factors such as provision of partial learning data, intrinsic model ambiguity and the probabilistic nature of token prediction (Anjum, 2025). Arash Heidari (2024) has mentioned that these technical shortcomings are compounded by issues of data credibility where models like ChatGPT often rely on potentially outdated or biased datasets and often lack the capability to verify the accuracy of sourced information, making manual verification essential for confidential or time-sensitive data (Arash Heidari, 2024).

According to Buckley (2025) beyond technical deficiencies, the integration of AI introduces significant human-centric risks and operational challenges. Buckley et al. warn that overdependency on these tools may erode critical thinking, diminish human oversight, and disrupt workforce dynamics (Buckley, 2025). To address these vulnerabilities, Mohamed El-Helaly (2024) emphasizes that the drawbacks of AI can only be minimized through structured governance, advocating for the establishment of robust workplace regulations and comprehensive training in AI for both employers and employees (El-Helaly, 2024). Therefore, successful implementation requires a balanced framework that prioritizes human verification and policy-driven oversight to mitigate the risks of misinformation and cognitive complacency.

3. Methodology

This research adopts a controlled comparative analysis to evaluate the competency of AI chatbots in industrial safety engineering. It employs a dual-method approach, utilizing quantitative scoring based on Key Performance Indicators (KPIs) and qualitative analysis of response generated by AI chatbots. The methodological framework is detailed below:

3.1. Prompt Development

To assess varying dimensions of AI chatbots cognitive capability, a standardized set of three prompts was developed for each safety domain (Electrical, Mechanical, and Fire Safety). These prompts were categorized into three distinct levels of complexity:

1. **P1: Fundamental Knowledge:** Benchmarked against the Multitask Language Understanding (MMLU). According to Hendrycks et al., these prompts measure factual recall and technical definitions across safety disciplines (Hendrycks, 2021).
2. **P2: Applied Reasoning:** Designed to evaluate the chatbot’s capacity for applied reasoning and contextual problem-solving. According to Guo et al. this dimension aligns with established metrics in AI explainability and decision-support systems, where logical consistency, analytical coherence, and situational reasoning serve as key performance indicators (KPIs) for assessing reasoning quality and interpretability (Guo & et al., 2023).

3. **P3: Image-Based Analysis:** Based on recent multimodal AI studies, these prompts test the chatbot's ability in image recognition with contextual safety analysis which is a significant frontier for current LLMs.

Each prompt consisted of a single, well-defined question (no follow-up or sub-questions asked) designed to elicit a specific type of response (conceptual, applied, or analytical) from the chatbots based on the prompt. The same set of prompts were administered across all chatbots to ensure standardization and comparability. This approach ensured that all chatbots were evaluated under identical input conditions, eliminating prompt variability as a confounding factor.

3.2. Establishment of 'Gold Standards'

Prior to testing, a 'gold standard' representing the ideal or expected answer was prepared for each set of prompts. These gold standards were constructed based on Singapore's WSH Acts and Regulations, Approved Code of Practice, Singapore Standards, and verified literature. This ensured that the evaluation process remained objective and served as a benchmark for the evaluators when assessing the AI-generated responses and ensured objectivity during performance scoring.

3.3. Prompt Testing Protocol

To maintain objectivity and fairness across all AI chatbot systems, a strictly standardized prompt testing protocol was enforced:

- **System Consistency:** All tests were conducted under controlled conditions using two identical laptops operating on Windows 11.
- **Platform Access:** The default or currently available free versions of ChatGPT, Perplexity, Google Gemini, and Grok were used to simulate the typical user experience.
- **Context control:** To eliminate "conversational memory bias," each prompt was entered as the initial input in a new, independent chat session.
- **Documentation:** Responses were captured via unedited text for analysis and screenshots for visual verification.

3.4. Evaluation Framework

The responses generated by the AI chatbots were quantified using a rubric-based scoring system derived from existing literature (Subasinghe et al., 2025; Yazdi, Zarei, Adumene, & Beheshti, 2024; Schulze Balhorn, Weber, Buijsman, et al., 2024). The framework assesses three core KPIs which are described below:

- **Accuracy:** It is the most fundamental metric for assessing the credibility and reliability of AI-generated responses. Studies by (Chen & et al., 2023) and (Bang & et al., 2023) emphasize that factual precision is essential when chatbots are deployed in safety-critical or situational awareness needed to support human judgements, where any misinformation presented could lead to real-world harm.
- **Completeness:** It is vital in evaluating the depth and breadth of a model's response, ensuring that generated content encapsulates all relevant dimensions of a query (Shen & et al., 2023). In safety-related contexts, incomplete information may lead to misinterpretation or unsafe actions carried out by the end users. Research by (Shen & et al, 2023) and (Zhang & et al, 2024) highlight that completeness, along with factual accuracy, constitutes the two pillars of reliable AI evaluation, particularly in technical and knowledge-intensive domains.
- **Relevance:** It focuses on how well the chatbot's answers match the user's intent and context. It ensures that responses are not only correct but also meaningful and appropriate to the situation. As noted by (Mehri & Eskenazi, 2020), relevance is a key factor in determining the overall quality and usefulness of AI-generated dialogue.

The evaluation scores were converted to percentages using 20-point increments (1 = 20%, 2 = 40%, 3 = 60%, 4 = 80%, 5 = 100%) to standardize comparison across chatbots (Table 2).

Table 2. Evaluation Rubrics of Chatbots' Response

Score	Accuracy	Completeness	Relevance
5	Excellent	Fully Complete	Highly Relevant & Actionable
4	Very Good/Minor Flaws	Mostly Complete	Very Useful & Specific
3	Acceptable/Oversimplified	Moderately Complete	Adequate Relevance
2	Poor/Misleading	Significantly Incomplete	Tangentially Relevant
1	Hallucination	Fails to Address Prompt	Irrelevant

The evaluation team was composed of evaluators from diverse backgrounds to ensure a balanced mix of industry experience and academic insights, thereby reducing individual bias and strengthening the credibility of the evaluation outcomes.

1. **Internal Evaluator:** Mr. Nazurudin Syed Ahmed, a Ministry of Manpower (MOM) registered Safety Officer (ID: 07828) in Singapore with a rich work experience in oil and gas sector for over 15 years.
2. **Internal Evaluator:** Mr. Mukesh Kumar Likhmawat Prashanth, an M.Sc. in Safety, Health, and Environmental Technology graduate from the National University of Singapore (NUS).
3. **External Evaluator:** Mr. Nawfal Shahib, a Ministry of Manpower (MOM) registered Safety Officer (ID: 05672) in Singapore was engaged to provide external independent evaluation. Mr. Shahib possesses over 15 years of operational experience in HSE in the construction and marine industries.

3.5. Data Analysis

Upon completion of the evaluation, the individual scores from the three evaluators were aggregated to compute the arithmetic mean for each prompt. These final values were subsequently utilized to generate graphical visualizations illustrating performance trends.

4. Data Collection

By synthesizing these quantitative metrics with qualitative observations, the study systematically identified operational patterns to evaluate the relative competence and domain-specific effectiveness of the AI models (Table 3).

Table 3. Compilation of AI chatbots' Responses Across Different Safety Domains

Domain A: Electrical Safety	
Prompt 1: Fundamental Knowledge	
Question:	What is the purpose of grounding (earthing) in an electrical system, and how does it protect individuals from electric shock?
Gold Standard:	Earthing serves three purposes. They are: 1. To provide a safe path for currents to flow to the ground. 2. Proper equipment grounding helps discharge fault currents safely, preventing electrical hazards and reducing the risk of equipment damage during normal operation and fault conditions. (Singh, 2024). 3. To protect a human from getting electric shock.
Prompt 2: Applied Reasoning	
Question:	A factory has experienced frequent tripping of circuit breakers and reports of minor electrical shocks from equipment. As a safety officer, how would you systematically identify the root cause of these issues and propose corrective actions?
Gold Standard:	Electric shocks can occur when there are faults or inadequacies in electrical installations, such as incorrect wiring, damaged insulation, or poorly maintained components (Kulor, Apprey, Agbevannu, & Gasper, 2024). As a safety officer, preventing the recurrence of electrical incidents begins with identifying the key contributors to electrical hazards, which is essential for effective hazard elimination. Once these contributors are identified, preventive measures should be implemented through regular maintenance, periodic equipment checks, and consistent use of checklists to document inspection findings.

Prompt 3: Image-based Analysis	
Question:	What is the purpose of this device? 
Gold Standard:	The above image is a dissipation array system (DAS) which is a key component within a larger lightning protection system that is designed to prevent lightning strikes, rather than intercept them. It uses multiple points to emit a low-level corona discharge, which neutralizes the buildup of charge and interrupts the formation of upward streamers that could connect with a downward lightning leader (Bazelyan & Drabkin, 2003).
Domain B: Mechanical Safety	
Prompt 1: Fundamental Knowledge	
Question:	What are the key differences between inherent safety design and safeguarding measures in machinery safety?
Gold Standard:	Inherently safe design measures are at the top of the Hierarchy of protective measures stated by the ISO 12100:2010 standard [3]. (Chinniah, Nix, & Jocelyn, 2019). This means that in machinery safety, inherently safe design emphasizes the elimination or fundamental reduction of hazards through design modifications such as geometry optimization, energy limitation, material selection, or layout reconfiguration, ensuring that the hazard itself no longer exists or is significantly minimized. The second layer in the hierarchy is called “Safeguarding and complementary protective measures.” This layer includes physical means to enforce safe distance or prevent access to hazards, and active controls like interlocks and safeguarding devices that control hazardous energy. (Chinniah, Nix, & Jocelyn, 2019). Safeguarding measures are introduced only when risks cannot be sufficiently reduced by design.
Prompt 2: Applied Reasoning	
Question:	When designing a conveyor system in a factory, what specific safety measures should be implemented to prevent entanglement and crushing injuries?
Gold Standard:	Open belt conveyors, such as the one used in grain handling, pose serious entanglement and crushing hazards due to exposed inward-running pinch points at pulleys, idlers, and belts. Workers are especially at-risk during maintenance or cleaning when limbs, loose clothing, or tools can be pulled into the equipment. Best practices to prevent injuries include guarding all pinch points, providing emergency stops and audible alarms, implementing lockout/tagout procedures, maintaining clear walkways, and training employees on safe practices (Casner, 2022). One study as quote “<i>property of the belt, according to ISO 703 or DIN 22107</i>”, can be set as benchmark example to refer to ISO or any other international standard to explain the design of a safe system. (Barburski, Marcin, 2014).
Prompt 3: Image-based Analysis	
Question:	What are the mechanical hazards in this image that can lead to an injury? 
Gold Standard:	Hazard 1: The operator wore loose-fitting clothing which could get caught in the rotating spindle. Hazard 2: There was no chip shield or transparent guard, so metal swarf or fragments could fly off the cutter and strike the operator’s face or eyes. Hazard 3: The operator placed his hand close to the moving table and saddle of the milling machine, risking a pinch/crush injury when the table moves.”
Domain C: Fire Safety	
Prompt 1: Fundamental Knowledge	
Question:	What are the two core objectives of fire safety design? What main risks does each objective aim to address?
Gold Standard:	According to the Fire from First Principles and Fire code 2023, it is clearly stated that the two core goals of fire safety are life safety and property protection. Life safety is mainly aimed at protecting people from the harm of

	smoke, while property protection is to keep buildings away from the damage caused by high temperatures (Stollard & Abrahams, 1999).
Prompt 2: Applied Reasoning	
Question:	How can people integrate fire compartment and smoke control systems in the design of a large shopping mall to enhance life safety?
Gold Standard:	The goal of fire compartments is to limit the spread of fire and smoke, while smoke exhaust systems are designed to separate smoke from people and keep evacuation routes unobstructed (Stollard & Abrahams, 1999). For large buildings with atriums, smoke curtains can be deployed to create a “smoke containment zone” while simultaneously activating the smoke exhaust system to vent smoke, thereby ensuring safe evacuation of occupants in the lower areas. If fire compartment requirements need to be relaxed, an automatic sprinkler system and mechanical smoke exhaust system must be installed. The entire system must maintain a certain level of survivability.
Prompt 3: Image-based Analysis	
Question:	Identify any violations or hazards shown, and explain the specific risks they pose during an emergency?
	
Gold Standard:	The function of a corridor in the SCDF Fire Code 2023 regulations is defined as "providing a passage from a room or space to an exit", which fundamentally rules out the possibility of it serving as a storage space. According to the key points in Fire from First Principles: Corridors "The areas designated as evacuation passages shall not be converted for other uses. Blocked corridors will impede the evacuation of people and be regarded as violations. Accessibility should always be maintained. Secondly, flammable materials should not be piled up. A large number of flammable materials provide sufficient "fuel" for the fire, which will accelerate the spread of the fire and smoke (Stollard & Abrahams, 1999).

5. Results and Discussion

5.1 Scoring of Chatbot's Responses

Based on Table 2. Evaluation Rubrics of Chatbots' Response, the three evaluators provided their independent scoring on the chatbot responses vis-à-vis the gold standard answers as shown in Table 4 below.

Table 4. Evaluation Scores of AI chatbots Across Different Safety Domains

AI Chatbots	Safety Domains	Prompts	Evaluators								
			Safety Officer (Internal)			Safety Officer (External)			M.Sc. SHE Graduate		
			KPI Score			KPI Score			KPI Score		
			Accuracy	Completeness	Relevance	Accuracy	Completeness	Relevance	Accuracy	Completeness	Relevance
ChatGPT	Electrical Safety	P1	5	3	4	5	4	4	5	5	5
		P2	5	5	5	5	4	4	5	5	5
		P3	5	5	5	5	4	4	5	5	5
	Mechanical Safety	P1	4	4	4	4	5	4	5	5	5
		P2	3	2	3	3	3	3	4	5	5
		P3	4	4	4	4	4	4	5	4	5
	Fire Safety	P1	5	5	5	5	5	5	4	5	5
		P2	5	4	4	5	4	4	2	3	3
		P3	5	4	5	5	4	5	5	5	5
Perplexity	Electrical Safety	P1	5	5	4	5	5	5	4	4	5
		P2	4	4	4	5	4	4	5	4	5
		P3	1	1	1	1	1	1	1	1	1
	Mechanical Safety	P1	4	4	4	4	4	4	5	5	5
		P2	4	4	4	4	3	4	5	4	5
		P3	4	4	5	5	5	5	3	3	4
	Fire Safety	P1	5	5	5	5	5	5	5	5	5
		P2	5	4	4	5	4	4	3	3	3
		P3	4	3	4	3	4	4	3	3	4
Google Gemini	Electrical Safety	P1	4	3	4	3	3	3	5	5	5
		P2	5	4	4	4	4	4	5	5	5
		P3	1	1	1	1	1	1	1	1	1
	Mechanical Safety	P1	3	3	3	4	3	3	5	4	5
		P2	2	2	2	3	2	2	3	3	5
		P3	4	4	4	4	4	4	3	2	3
	Fire Safety	P1	5	4	4	4	4	4	5	5	5
		P2	4	4	4	4	4	4	2	3	3
		P3	4	4	4	5	4	4	5	5	5
Grok	Electrical Safety	P1	4	4	4	3	4	4	5	5	5
		P2	4	4	3	3	3	3	5	5	5
		P3	1	1	1	1	1	1	1	1	1
	Mechanical Safety	P1	4	4	4	4	4	4	5	3	5
		P2	4	4	4	4	4	5	4	5	5
		P3	4	5	4	4	4	4	3	4	5
	Fire Safety	P1	5	5	5	5	5	5	5	5	5
		P2	4	5	5	5	5	5	3	4	5
		P3	5	5	5	5	5	5	4	4	4

and illustrated (Refer to Figure 1, Figure 2, and Figure 3) to systematically discern performance patterns and consistency among the four AI chatbots.

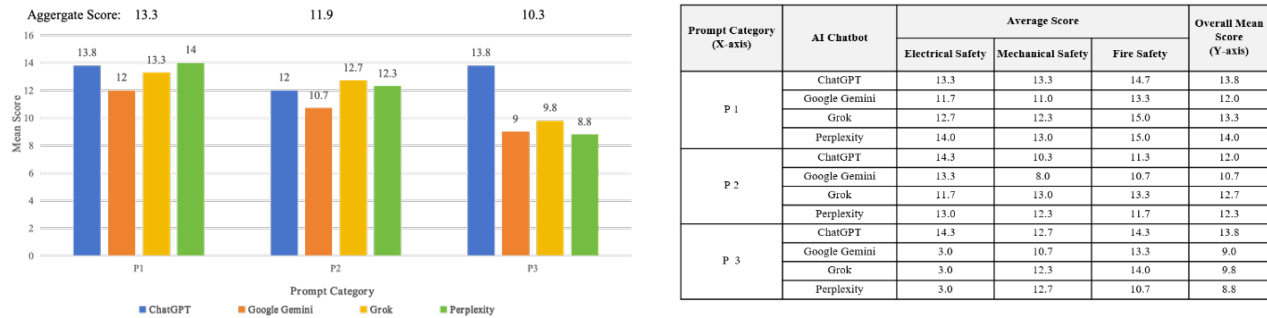


Figure 1. Performance of AI chatbots across different prompt categories

Across the prompt categories (P1–P3), the four AI chatbots demonstrated broadly consistent performance. Aggregate scores were highest for Fundamental Knowledge (P1), followed closely by Applied Reasoning (P2), while Image-Based Analysis (P3) recorded the lowest overall scores. Strong performance in P1 across the four chatbots indicates a sound baseline understanding of safety concepts, whereas the marginal decline in P2 reflects the higher cognitive demands of reasoning-based tasks (Figure 1).

In contrast, P3 exhibited the greatest performance disparity and weakest results, confirming that visual interpretation remains a key limitation for current AI chatbots. For example, in fire safety domain, the chatbots - ChatGPT, Grok, and Google Gemini correctly identified the Housing and Development Board (HDB) flat that is common to Singapore in the image and gave responses based on that. Surprisingly, for electrical safety domain, only ChatGPT provided the most accurate and contextually aligned response to the gold standard. The remaining chatbots – Google Gemini, Grok, and Perplexity – produced misinterpreted or hallucinated responses (like some comparing the dissipation array system in the image to a wind turbine), highlighting the potential limitation of these AI chatbots in visual reasoning and contextual analysis.

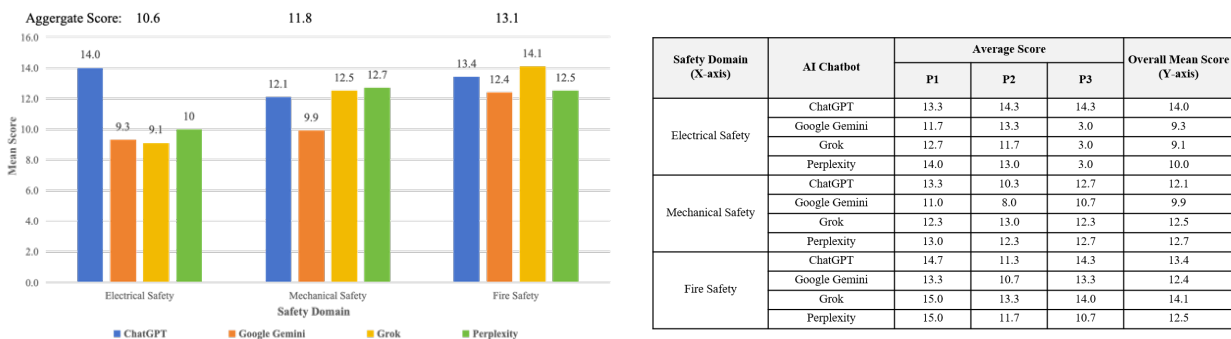


Figure 2. Performance of AI chatbots across different safety domains

Comparative analysis across the three safety domains reveals clear variations in domain-specific competency among the four AI chatbots. The strongest overall performance was observed in the fire safety domain (Figure 2), where all chatbots demonstrated strong alignment with established gold standards and exhibited consistent levels of fundamental knowledge and applied reasoning. Minimal score deviation suggests a shared familiarity with core fire safety principles. Minor limitations were observed in the applied reasoning responses, particularly the inconsistent explanation of technical acronyms such as CFD (Computational Fluid Dynamics), which were not clearly defined in certain responses. This ambiguity necessitated an additional search to interpret the acronym accurately, thereby indicating minor limitations in content clarity and completeness. Notably, the chatbots appropriately referenced Singapore Civil Defence Force (SCDF) guidelines, indicating strong awareness of local regulatory frameworks.

Although the study's gold standards were derived from peer-reviewed literature, alignment with SCDF requirements underscores the contextual accuracy and regulatory relevance of the responses within the Singapore fire safety context. Mechanical safety performance was moderate across all AI chatbots and generally aligned with the established gold standards. Most chatbots referenced relevant international standards, including ISO 12100 and ISO 13857, supporting the technical validity of their recommendations. In contrast, Google Gemini did not cite any ISO standards, which adversely affected its performance score, as reference to recognized standards is a key indicator of technical accuracy and defensible safety reasoning.

The electrical safety domain recorded the weakest overall performance, reflecting the higher technical complexity associated with regulatory application, specialized terminology, and hazard control measures such as grounding, isolation, and circuit protection. The wider score dispersion (9.1–14.0) indicates variability in model depth and precision in this domain. Overall, these findings suggest that while AI chatbots are effective in addressing general and procedural safety topics, domains requiring precise regulatory and engineering-based reasoning remain a critical area for further model refinement (Figure 3).

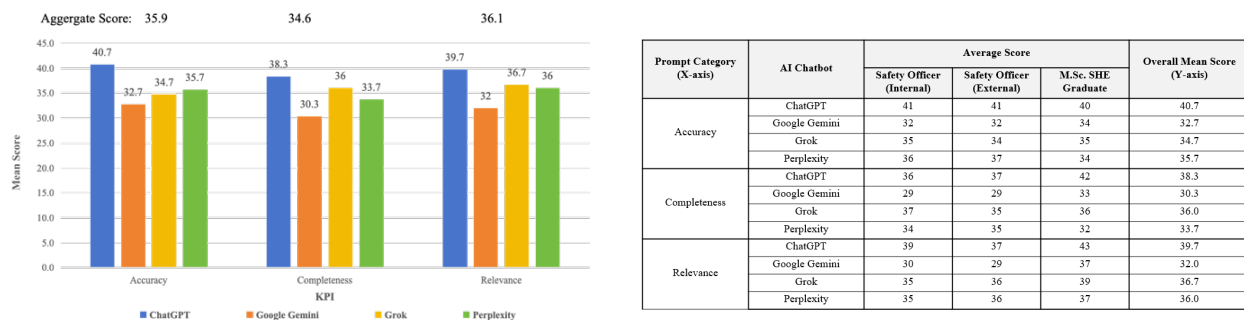


Figure 3. Performance of AI chatbots based on KPIs

5.3 Validation

Relevance emerged as the strongest overall KPI, with a combined average score of 36.1, indicating strong contextual alignment with safety-related objectives. ChatGPT achieved the highest relevance score (39.7), followed by Grok (36.7) and Perplexity (36.0), while Google Gemini scored lower (32.0), reflecting weaker alignment with prompt intent. Overall, ChatGPT, Grok, and Perplexity demonstrated greater consistency in contextual appropriateness across varied prompt structures.

Accuracy ranked closely behind relevance, with a combined average score of 35.9, suggesting that the chatbots were generally capable of producing factually correct responses. ChatGPT again performed best (40.7), followed by Perplexity (35.7), Grok (34.7), and Google Gemini (32.7). While ChatGPT showed stronger inter-evaluator consistency, the remaining models exhibited comparable mid-range performance, indicating moderate but reliable factual precision.

Completeness recorded the lowest aggregate score (34.6), reflecting uneven coverage of prompt requirements. ChatGPT achieved the highest completeness score (38.3), indicating a stronger ability to deliver holistic responses, followed by Grok (36.0) and Perplexity (33.7). Google Gemini scored lowest (30.3), suggesting comparatively limited depth and elaboration in its outputs.

Overall interpretation of results

The comparative analysis of the four AI chatbots reveals distinct performance patterns across prompt categories, safety domains, and KPIs. While chatbots exhibit a robust foundation in general safety principles, they display slight underperformance in reasoning, contextual interpretation, and visual analysis, with average score deviations ranging between ± 2 to 4 points.

Firstly, performance was strongest for text-based prompts involving fundamental knowledge and applied reasoning, highlighting the effectiveness of current large language models (LLMs) in structured information retrieval and

interpretation. In contrast, image-based analysis consistently underperformed, underscoring persistent limitations in multimodal reasoning and contextual visual identification and interpretation.

In terms of safety domains, fire safety emerged as the most consistently well-handled area by the AI chatbots, with responses aligning closely with Singapore Civil Defence Force (SCDF) guidelines and established terminologies, suggesting strong representation of fire safety content in training datasets. Electrical safety recorded the weakest performance, reflecting the higher technical and regulatory complexity associated with engineering controls and standards-based reasoning.

Lastly, in terms of KPIs, accuracy and relevance were identified as collective strengths, whereas completeness remained a recurring limitation for the AI chatbots. Although responses were generally correct and contextually aligned, many lacked the depth and comprehensiveness expected in professional safety assessments.

6. Conclusion

This study demonstrates that AI chatbots possess substantial potential as assistive tools for HSE professionals. While these technologies cannot replace HSE practitioners, engineers, or managers; they can effectively support routine tasks such as information verification, calculations, training material development, etc. In this respect, AI chatbots occupy a role comparable to established digital knowledge platforms (e.g., search engines and online reference tools), with the added advantage of improved accessibility and usability across diverse user groups. Nevertheless, their output must be utilized cautiously in workplaces, as biases, inaccuracies, and contextual limitations may still occur.

HSE professionals are therefore encouraged to adopt AI chatbots as supportive frameworks rather than sole decision-makers. Applications such as drafting templates, improving language clarity, and assisting with non-critical analyses are appropriate, whereas safety-critical processes - particularly risk assessment - remain firmly grounded in human judgement, professional experience, and peer review.

Consistent with existing literature and the empirical findings of this study, AI chatbots exhibit broad functional capability but limited depth in specialized domains, reinforcing the characterization of AI as a “generalist” rather than a “domain specialist”. This underscores the continued necessity of human oversight when deploying AI tools in high-risk, regulation-driven safety environments. Lastly, further research should focus on designing effective prompt strategies such as development of a structured prompting toolkit tailored for HSE professionals which may enhance the accuracy, reliability, and consistency of AI-assisted outputs.

References

- Adamopoulou, E., and Moussiades, L. “An overview of chatbot technology,” 2020.
- Nayyar, A. “Deep insights for large language models,” in *Mastering Prompt Engineering*, Morgan Kaufmann, Chicago, p.250, 2025.
- Anjum, F. *From hallucination to harm: Unintended ethical risks in large language models*, ISRA, Mangalore, 2025.
- Heidari, A., and J. N. “Everything you want to know about ChatGPT: Components, capabilities, applications, and opportunities,” Wiley, 2024.
- Bang, Y., et al. “Multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity,” 2023.
- Barburski, M. “Analysis of the mechanical properties of conveyor belts on the three main stages of production,” 2014.
- Bérestégui, P. “Artificial intelligence in Industry 4.0: Implications for occupational safety and health,” *European Agency for Safety and Health at Work*, Discussion Paper, 2024.
- Bhatti, U.J. “Transforming petrochemical safety using a multimodal AI visual analyzer,” *Engineering Proceedings*, p.78, 2024.
- Buckley, C. “Revisiting the consequences of relying too much on AI,” *Buckley Planet*, 2025.
- Carvalho, I., and Ivanov, S. “ChatGPT for tourism: Applications, benefits and risks,” *Tourism Review of AIEST*, pp.290–303, 2024.
- Casner, L. “Working around open belt conveyors,” 2022.
- Chen, S., et al. “Benchmarking factuality evaluation of large language models,” 2023.
- Chinniah, Y., Nix, S.D., and Jocelyn, S. “Safety of machinery: Significant differences in two widely used international standards for the design of safety-related control systems,” 2019.
- Electrical Safety Foundation International (ESFI). “Electrical safety,” 2025.

- El-Helaly, M. "Artificial intelligence and occupational health and safety: Benefits and drawbacks," 2024.
- European Commission. "Industry 5.0: Towards a sustainable, human-centric and resilient European industry," 2021.
- Guo, Z., et al. "Evaluating large language models: A comprehensive survey," 2023.
- Hassanain, M.A. "Fire safety risk assessment of workplace facilities: A case study," *Frontiers in Built Environment*, 2022.
- Hendrycks, D., et al. "Measuring massive multitask language understanding," 2021.
- Hostetter, H.N. "The role of large language models (AI chatbots) in fire engineering: An examination of technical questions against domain knowledge," *Engineering Applications of AI*, 2024.
- Iyengar, P. "Clever Hans in the loop? A critical examination of ChatGPT in a human-in-the-loop framework for machinery functional safety risk analysis," *Mathematics and Engineering*, p.31, 2025.
- Khurram, M. "Artificial intelligence in manufacturing industry worker safety," *Processes*, p.1312, 2025.
- Lightning Eliminators. "Lightning protection strategies for oil and gas facilities," 2025.
- Manju Lata, and K. V. "Exploring the future of conversational technology in education," in *Chatbots and Beyond*, Wiley, p.304, 2025.
- Mehri, S., and Eskenazi, M. "USR: An unsupervised and reference-free evaluation metric for dialog generation," 2020.
- Ministry of Manpower, Singapore. *WSH National Statistic Report 2023*, 2024.
- Ministry of Manpower, Singapore. *WSH National Statistic Report 2024*, 2025.
- Ndlela, M.N. *Organizational communication in the digital era*, Springer, Switzerland, 2024.
- Owidi, S., Lyanda, J.N., and Wangila, E. "Comparative study on accuracy of responses by select AI tools: ChatGPT and Perplexity AI versus human responses," 2024.
- Paltrinieri, N., and Khan, F. "Dynamic risk analysis in the chemical and petroleum industry," 2016.
- "Real-time personal protective equipment non-compliance recognition on AI edge cameras," *Electronics*, p.2290, 2024.
- Scherr, S., Cao, B., et al. "Explaining the use of AI chatbots as context alignment: Motivations behind the use of AI chatbots across contexts and culture," 2025.
- Setzen, S.A. "AI-powered laryngoscopy: Exploring the future with Google Gemini," Wiley, 2025.
- Shen, et al. "Evaluating content completeness and contextual depth in AI-generated responses," *Computers in Human Behavior*, 2023.
- Szwajca, D. "The importance of reputation of a country in the process of building its competitive advantage on the global market," 2017.
- Subasinghe, S.M.S., Gersib, S.G., Frueh, T.M., and Mankad, N.P. "Can large language models (LLMs) act as virtual safety officers?" *ACS Chemical Health & Safety*, Vol.32, No.1, pp.39–47, 2025.
- Schulze Balhorn, L., Weber, J.M., Buijsman, S., et al. "Empirical assessment of ChatGPT's answering capabilities in natural science and engineering," *Scientific Reports*, Vol.14, p.4998, 2024.
- Trends, G.A. "Google's Gemini and xAI's Grok rapidly close gap on ChatGPT in consumer AI race," 2025.
- Zhang, T., et al. "Multi-dimensional evaluation framework for conversational AI systems," *Expert Systems with Applications*, 2024.
- Zhu, X.L. "Can a chatbot enhance hazard awareness in the construction industry?" *Frontiers in Public Health*, 2022.
- Yazdi, M., Zarei, E., Adumene, S., and Beheshti, A. "Navigating the power of artificial intelligence in risk management: A comparative analysis," *Safety*, Vol.10, No.2, p.42, 2024.

Biographies

Nazurudin Syed Ahmed is currently pursuing Master of Science in Safety, Health, and Environmental Technology in the Department of Chemical and Biomolecular Engineering at the National University of Singapore (NUS). He holds a Bachelor of Science (Hons) in Safety, Health, and Environmental Management from Leeds Beckett University. With over 16 years of industrial experience, Mr. Ahmed is a seasoned Health, Safety, and Environment (HSE) specialist in the Oil & Gas, Petrochemical, and Aerospace sectors. He currently serves as the Senior HSE Engineer at Safran Landing Systems, where he leads strategic safety compliance and operational risk management. He is a Ministry of Manpower (MOM) registered Workplace Safety and Health Officer (WSHO) and Fire Safety Manager (FSM) in Singapore. His professional expertise encompasses Safety Management Systems (ISO 45001/SS651), emergency response planning, and behavioral safety leadership.

Prashanth Mukesh Kumar Likhmawat is a recent graduate of Master of Science in Safety, Health, and Environmental Technology with a specialization in General Safety and Sustainability from the Department of Chemical and Biomolecular Engineering at the National University of Singapore (NUS). Mr. Prashanth also holds a

Bachelor of Science in Microbiology with Distinction from Madras Christian College, University of Madras. His primary research interests focus on the intersection of advanced technology and occupational safety, Risk Management, and the application of Generative AI in complex and dynamic workplaces. His current work evaluates the operational competency of Large Language Models (LLMs) in high-risk engineering domains. He is actively seeking to leverage data-driven strategies to enhance organizational safety culture and regulatory compliance in Singapore.

Wang Xinyu is currently pursuing Master of Science in Safety, Health, and Environmental Technology with a specialization in General Safety and Sustainability from the Department of Chemical and Biomolecular Engineering at the National University of Singapore (NUS). Mrs. Wang Xinyu also holds a Bachelor of Architecture in Landscape Architecture with Sichuan Agricultural University. Her current research interests focus on sustainable safety and energy environments, risk management, and the application of artificial intelligence for industrial safety in various complex settings, while evaluating the feasibility and effectiveness of AI deployment throughout this process. She is currently actively exploring how to better apply digitalization to safety, health, and environmental technologies. Then enhance regulatory reliability and improve work efficiency while reducing risks in the workplace.

Sam Michael Arockianathan is currently pursuing Master of Science in Safety, Health, and Environmental Technology with a specialization in Safety and Sustainability from the Department of Chemical and Biomolecular Engineering at the National University of Singapore (NUS). Mr. Michael has completed his Environmental Engineering degree at Griffith University Australia and has done his thesis on Design and Development of Carbon Nanodots to Sense Phenols in Wastewater. Mr. Michael is an Environmental Consultant/Engineer with six years of experience in environmental site assessments, site remediation, and a few HSE projects. He has led projects involving ISCO treatment, offshore and nearshore oil spill response, landfill gas and odor assessments, air, water, and noise monitoring including source and stack emissions testing. His current research interest involves the implementation of various technologies to improve energy production in power plants.

Fan Keli is a Master of Science student in Safety, Health, and Environmental Technology at the Department of Chemical and Biomolecular Engineering, National University of Singapore (NUS). She holds a bachelor's degree in landscape architecture, with technical training in environmental systems, spatial planning, and infrastructure-related design. Building on this interdisciplinary foundation, her current research interests focus on industrial safety risk assessment, sustainable safety and energy environments, and the application of artificial intelligence and digital tools to support hazard identification, risk evaluation, and safety decision-making. She is particularly interested in improving the reliability and efficiency of safety management systems through data-driven and AI-enabled approaches in complex engineering and workplace settings.

Loh Tzu Yang is an Associate Professor and Co-Director of the M.Sc. in Safety, Health, and Environmental Technology program at the National University of Singapore. Passionate about advancing occupational safety and health, he integrates academic rigor with practical expertise to train future leaders in workplace safety, risk analysis, and occupational hygiene.