DOI: 10.1111/risa.14351

ORIGINAL ARTICLE

How good are large language models at product risk assessment?

Zachary A. Collier[1] | **Richard J. Gruss**[1] | **Alan S. Abrahams**[2]

[1]Department of Management, Radford University, Radford, Virginia, USA

[2]Department of Business Information Technology, Virginia Tech, Blacksburg, Virginia, USA

Correspondence
Zachary A. Collier, Davis College of Business and Economics, Radford University, 701 Tyler Avenue, Radford, VA 24142.
Email: zcollier@radford.edu

Abstract

Product safety professionals must assess the risks to consumers associated with the foreseeable uses and misuses of products. In this study, we investigate the utility of generative artificial intelligence (AI), specifically large language models (LLMs) such as ChatGPT, across a number of tasks involved in the product risk assessment process. For a set of six consumer products, prompts were developed related to failure mode identification, the construction and population of a failure mode and effects analysis (FMEA) table, risk mitigation identification, and guidance to product designers, users, and regulators. These prompts were input into ChatGPT and the outputs were recorded. A survey was administered to product safety professionals to ascertain the quality of the outputs. We found that ChatGPT generally performed better at divergent thinking tasks such as brainstorming potential failure modes and risk mitigations. However, there were errors and inconsistencies in some of the results, and the guidance provided was perceived as overly generic, occasionally outlandish, and not reflective of the depth of knowledge held by a subject matter expert. When tested against a sample of other LLMs, similar patterns in strengths and weaknesses were demonstrated. Despite these challenges, a role for LLMs may still exist in product risk assessment to assist in ideation, while experts may shift their focus to critical review of AI-generated content.

KEYWORDS
FMEA, Generative AI, Product safety

1 | INTRODUCTION

Firms that manufacture and sell consumer products must effectively manage product safety risks. Defective and hazardous products can cause costly product recalls that may result in lost sales due to products being taken off the market, replacement and repair costs, potential fines, reputational damage, and legal costs (Ameer & Othman, 2023; Mayo et al., 2022). The recall process itself, which can often be lengthy, is costly for the entire supply chain, including retailers and manufacturers (Wowak et al., 2022). Even though companies have an incentive to manufacture and sell safe and high-quality products, quality and safety issues can still occur for a number of reasons, such as cost-cutting and inattentiveness (Ball et al., 2018) and from outsourcing (Steven et al., 2014).

Product manufacturers must take precautions to ensure that their products are safe, while not being too costly to produce (Rausand & Utne, 2009). More specifically, product safety engineering involves an analysis of risk, identifying all of the relevant hazards associated with the product during the different lifecycle stages, and then eliminating or mitigating the hazards (Rausand & Utne, 2009). The product safety engineering process parallels the general risk analysis process described by Kaplan and Garrick (1981), in which the analyst asks "What can go wrong," "How likely is it," and "What are the consequences," as well as the risk management questions posed by Haimes (2012): "What can be done, and what options are available," "What are the trade-offs among all relevant costs, benefits, and risks," and "What are the impacts of current decisions on future options."

One of the most challenging steps is hazard identification, where the failure to identify hazards can be a major source of error (Redmill, 2002). Not identifying all scenarios renders an analysis incomplete and ineffective (Haimes, 2012). Another challenge lies in the identification of risk treatment

or response alternatives (Hillson, 1999). While a risk analysis can provide information about how much various courses of action might reduce risk, there is often a disconnect between the risk analysis results and the decision making needed to balance costs and benefits (Linkov et al., 2014; Paté-Cornell & Dillon, 2006;).

Recent developments in generative artificial intelligence (AI), especially large language models (LLMs), have the potential to assist safety engineers in the product risk assessment process. In this article, we investigate how LLMs can support product risk assessment. Specifically, we used OpenAI's ChatGPT 3.5 to generate product risk assessment output and guidance for a set of consumer products. We then developed a survey in which we asked product safety professionals to evaluate the quality of the generated responses.

The remainder of the article is organized as follows. In Section 2 we summarize the process of product safety engineering from a risk management perspective and provide an overview of recent applications in the literature of AI for risk assessment and management. In Section 3, we describe our methodology. Section 4 describes the results of the study, Section 5 provides a discussion of the results, and Section 6 discusses limitations. Finally, Section 7 highlights our conclusions, the managerial implications, and future research opportunities.

2 | BACKGROUND AND RELATED WORK

In the following subsections, we review closely-related work in product safety engineering, and in the application of AI to product risk analysis.

2.1 | Product safety engineering

Product safety engineering is concerned with eliminating or reducing the potential risk associated with product failures that can result in hazardous situations (Rose, 1989), and involves product risk assessment, which is "determining whether a product is safe for consumers to use" (Hunte et al., 2022). Products that are defective or cause injury to users (or the user's property) may result in liability for the manufacturer or retailer (Ryan, 2003). Safe product design requires balancing safety (including legal and regulatory requirements and standards) on one hand and production cost, functional performance, and schedule constraints on the other (Rausand & Utne, 2009).

The product risk assessment process starts with hazard identification, where a hazard is a potential source of harm (Hunte et al., 2022). Hazard identification requires understanding who the product user is, their abilities and limitations, how the user will interact with the product, and the environment in which the product will be used (McRoberts, 2005). Beyond the intended use, there may be foreseeable misuses that the manufacturer did not intend (Wright, 2007). Such misuse can arise due to misunderstanding by the user, misreading the instructions, or erroneous assumptions about the product's use (Rausand & Utne, 2009; McRoberts, 2005).

The next step is risk estimation (Hunte et al., 2022). In this step, identified hazards are quantified in terms of severity and occurrence frequency or probability (Hunte et al., 2022; Iyenghar et al., 2022). Risk is often calculated as the product of severity and occurrence, and the resulting risk level may be categorized (Hunte et al., 2022; McRoberts, 2005).

Finally, the risk evaluation stage uses the results of the risk estimation to determine whether the product risk is acceptable and what decisions should be made regarding risk reduction (Hunte et al., 2022; McRoberts, 2005). Alternatives for risk reduction are identified and applied in this step. Risk reduction alternatives can typically be prioritized according to a "safety hierarchy," stating that if possible, the hazard should first be *eliminated* through design. If this is not possible, then protective *safeguards* should be added to the design. Finally, if this is also not possible, then *warnings* should be provided to the user, as well as training and instruction and personal protective equipment (PPE) (Barnett & Brickman, 1986; Ross, 2021).

A number of standards exist related to product safety. For example, ISO 10377 provides guidance on consumer product safety, including methodology for hazard identification, assessment, and reduction, as well as risk management and provision of warnings to consumers (ISO, 2013). More specialized product-category-specific standards exist as well, including, for example, ASTM F963-23, "Standard Consumer Safety Specification for Toy Safety" (ASTM, 2023), ANSI/WCMA A100.1-2022, "American National Standard for Safety of Corded Window Covering Products" (ANSI/WCMA, 2022), NFPA 10, "Standard for Portable Fire Extinguishers," (NFPA, 2022), and many others. Product designers must be familiar with all of the applicable standards and regulations that apply within their industry.

Failure mode and effects analysis (FMEA) is a methodology that is used to identify and assess risks of products during the new product development process (Moreira et al., 2021). An FMEA is used to identify failure modes and their causes and the effects that the failure will have on the system or product (Carlson, 2014). It is a tool that aids in the assessment and prioritization of failure modes, and facilitates the identification of risk treatment activities (Carlson, 2014). According to the American Society for Quality (ASQ), failure modes are defined as "ways, or modes, in which something might fail. Failures are any errors or defects, especially ones that affect the customer, and can be potential or actual" (ASQ, n.d.). Breyfogle (2003) defined a failure mode as the way "a design might fail to perform its intended function." Failure modes are distinct from their effects, which describe "the effects of the failure mode on the function from an internal or external customer point of view" and their causes, which indicate "a design weakness that causes the potential failure mode" (Breyfogle, 2003). One of the key characteristics is a numerical score assigned to each failure mode along three

dimensions of Severity (S), Occurrence (O), and Detection (D), each typically being defined on a 1–10 scale. A risk priority number (RPN) is calculated as the product of the three scores, $RPN = S \times O \times D$, and is used to prioritize failure modes that are of greatest concern (ASQC/AIAG, 1995). Industry standards such as IEC 60812:2018 (IEC, 2018) and SAE J1739_202101 (SAE, 2021) provide guidance on how to perform and document FMEAs.

2.2 | AI and LLMs for product risk analysis

The term "artificial intelligence" was coined in the 1950s, and can be thought of as a general term for computational methods that seek to mimic human intelligence (Howard, 2019). AI simulates human intelligence by collecting, processing, and acting on data in ways that allow it to learn through the acquisition of new data (Canhoto & Clear, 2020). While there are many types of AI algorithms, LLMs have recently gained popularity through a number of available platforms such as ChatGPT, Gemini, and others. These LLMs have differentiated themselves from previous AI through their ability to excel at creative, analytical, and writing-based tasks (Dell'Aqua et al., 2023).

While these tools appear to have the potential to increase worker performance, the capabilities of these AI tools are still somewhat unevenly distributed, creating what Dell'Aqua et al. (2023) described as a "jagged technological frontier," where AI overperforms on certain tasks yet underperforms on others of similar difficulty. While AI tends to excel at tasks that are analytic in nature, humans still excel at decision-making tasks that require intuition and where there may be multiple divergent interpretations of a situation (Jarrahi, 2018). Overreliance on AI output in task domains that are beyond AI's technological frontier can result in a decrease in task performance (Dell'Aqua et al., 2023). AI therefore has the potential to both create value and destroy value in businesses (Canhoto & Clear, 2020).

Generative AI tools have many known limitations (Chang et al., 2023), including challenges with abstract reasoning in complex contexts, robustness issues in the face of unexpected inputs, trustworthiness (e.g., hallucination of statements ungrounded in reality), questionable ethics and bias, difficulty representing human disagreements, and a shortage of benchmark tasks across varied domains to evaluate domain-specific performance. Indeed, when we prompted ChatGPT to self-reflect on the limitations of Generative AI tools, it acknowledged that Generative AI tools may be data dependent (limited by biased or incomplete data), repetitive (repeat similar patterns or samples), suffer from uncertainty about quality or relevance, lack precision or control over specific attributes in the generated output, may generate harmful or misleading content, may be difficult to interpret or understand or explain due to the complexity of the underlying model and difficulty debugging or refining the model's behavior, may respond with unexpected or undesirable outputs when inputs are subject to perturbations, and may raise legal and regulatory challenges relating to, inter alia, intellectual property rights. ChatGPT concluded with the optimistic assurance that research may address these challenges and "*unlock new opportunities [for the use of Generative AI] for creative expression, problem-solving and innovation*." These challenges are relevant and important for product safety engineers to acknowledge when considering whether to use Generative AI tools for risk assessment. The Generative AI output may be incomplete, incorrect, or biased, potentially impacting the quality of the risk assessment for which the output is used.

Measuring the true performance of AI is hugely challenging, and shoddy assessment of AI's capabilities would itself introduce risks (Roose, 2024). Various authors have proposed benchmarks to test the question-answering, classification, algebraic computation, and general reasoning capabilities, of LLMs. For a review of LLM benchmarking techniques, see (Ott et al., 2022; Valmeekam et al., 2024). Examples of LLM benchmarks include CommonSenseQA (question answering), BigBench (basic reasoning), PlanBench (action and change reasoning), AQUA-RAT (Algebraic Question Answering with Rationales), AQuA (label quality), and others. However, no specific benchmarks have been proposed for LLM application to product risk characterization. Providing a systematic method to assess LLM response quality on product risk characterization challenges is an unmet need for the risk analysis community, which we attend to in this article.

Ideally, an LLM should provide the complete response at the first prompt—so-called "zero-shot reasoning" (Henrickson & Meroño-Peñuela, 2023; Kojima et al., 2022), without having to be prompted with additional examples or prompt variations ("multi-shot reasoning"). Iterative experimentation with prompt variations to improve output quality is referred to as "prompt engineering," and has been investigated for both general LLM response tasks (Ekin, 2023; Henrickson & Meroño-Peñuela, 2023; Marvin et al., 2023; White et al, 2023a,2023b), and diverse domain-specific tasks, such as academic writing (Giray, 2023), healthcare (Meskó, 2023), entrepreneurial pitch-writing (Short & Short, 2023), and other domains, though not, to our knowledge, evaluated in the particular context of product risk assessment. Therefore, a major unmet need for the risk analysis community, which we attend to in this article, is the assessment of LLM prompt engineering variations (tactics) specifically on product risk assessment tasks.

A number of AI applications related to risk analysis exist across many disciplines. For example, Baryannis et al. (2019) surveyed the supply chain risk management literature, finding that the majority (84%) of AI applications focused on selecting risk responses. Comparatively little attention was placed on risk identification, risk assessment, or combinations of identification, assessment, and response (Baryannis et al., 2019). Aziz and Dowling (2019) reviewed applications of AI for risk management in the financial industry. They found that financial institutions were applying AI for managing credit risks, market risks, operational risks, and for compliance and regulatory risks. Specifically, AI models were applied for the detection of fraud, and for stress testing

risk models (Aziz & Dowling, 2019). Within the field of disaster risk management, AI is used to process and analyze remote sensing data to improve predictive risk models and distribute disaster aid (Gevaert et al., 2021).

Related to product safety, Zaman et al. (2024) developed a partially automated process that performs text analytics on online product reviews in order to assess product risk, though this process used conventional Machine Learning (ML) rather than LLMs. Iyenghar et al. (2022) described an AI-based assistant that uses conversational inputs to determine the expected risk reduction associated with machinery safeguards.

While previous research has evaluated the potential of AI to perform product risk analysis, most of these projects have conceptualized "AI" specifically as ML. To our knowledge, ours is the first study to assess LLMs at this task.

3 | METHODOLOGY

Our approach to assessing the ability of LLMs to aid in product risk assessment consisted of the following steps: (1) Identify a small number of consumer products with known hazards to use as case studies, (2) generate a thorough risk assessment of each product using ChatGPT, and (3) receive feedback on the output from product safety experts. These steps are explained in detail below.

3.1 | Identifying consumer products

To identify a set of products that pose potential safety risks, we downloaded the recall data from the US Consumer Product Safety Commission (CPSC) database.[1] This dataset consists of 8,753 product recalls between June of 1973 and September of 2023. We sorted the data according to the products with the most recalls and most units affected, and derived a candidate list of 12 products (Table 1). We then selected a subset of six products that would represent a wide variety of risks: fire extinguishers, dehumidifiers, hammers, window blinds, dry erase boards, and bath toys. In each product case, we used the product type name, rather than a specific model or brand name, in order to allow for an expansive set of potential failure modalities, rather than a narrow set of failure modes peculiar to a particular model or brand.

As the examples that follow illustrate, the selected consumer products have documented failure modes and have been known to cause injuries. Fire extinguishers are pressurized vessels, and therefore require periodic inspection of their components, including any seals, hoses, nozzles, valves, etc., as well as inspection to ensure that they are properly pressurized (Garcia-Martin et al., 2019). For example, Dalton (2005) described an instance of a stress-corrosion crack failure of a carbon dioxide fire extinguisher, resulting in a sudden release of pressure. Furthermore, failure of fire extinguishers to properly discharge has resulted in the recall of dozens of models, comprising over 40 million units in North America (CPSC, 2017). Dehumidifiers can pose fire hazards, as a recent CPSC recall illustrates, where approximately 1.5 million dehumidifiers were recalled because of the possibility of overheating and catching fire (Rahman, 2023). Owen et al. (1987) observed the number of patients over a 4-year period who sustained eye injuries (interocular foreign body) from using hammers. They found 55 patients sustained eye injuries, six of whom were rendered blind in the affected eye. The cords attached to window blinds have caused hundreds of strangulation deaths in children since the 1970s (Bendix, 2023). Approximately 1.6 million dry erase boards were recalled due to laceration hazards posed when the thin magnetic metal layer can become separated from the wooden layer (Neal, 2017). Finally, millions of bath toys have been recalled due to impalement and laceration hazards (Archie, 2023).

3.2 | ChatGPT-based risk assessment

A product risk assessment entails identifying failure modes, constructing an FMEA table, suggesting risk mitigations, and providing guidance for relevant stakeholders. We had ChatGPT (version 3.5) go through these steps by sequentially entering the prompts listed in Table 2. This was completed on November 4, 2023. No additional guidance was provided to ChatGPT with respect to the prompts, such as length or format, which were left to the AI's discretion. However, all responses were approximately 2,000 words long. For inter-product comparison purposes, no additional prompt refinements or iterations were requested after the outputs were generated by ChatGPT.

The complete output for all six products (fire extinguishers, dehumidifiers, hammers, window blinds, dry erase boards, and bath toys) is available in the Supporting Information.

3.3 | Receiving feedback from product safety experts

We solicited a review of the ChatGPT outputs we generated above, from industry experts with at least 2 years of consumer product safety experience. The solicitation was sent to the presidents of three major consumer product safety professional organizations in the United States: The International Consumer Product Health and Safety Organization (ICPHSO), the Society of Product Safety Professionals (SPSP), and ADK Information Services. ICPHSO hosts a signature annual meeting that attracts approximately 800 product safety professionals from around the world, as well as an international symposium that attracts 200–350 participants, and North American Regional Training Workshops for product safety and compliance professionals. SPSP is a nonprofit professional development organization whose mission

[1] https://www.cpsc.gov/Recalls

TABLE 1 Initial set of products from US CPSC recall database.

Product	Recall date	Units	Reason (Hazard)
Composite deck	5/13/2009	48 million linear feet	Premature deterioration, unexpected breakage
Window blinds	8/26/2009	4.2 million	Cords pose strangulation hazard
Dehumidifier	8/16/2023	1.56 million	Fire hazard
Bath toy	6/22/2023	7.5 million	Risk of impalement and laceration
Cooler	3/9/2023	1.9 million	Magnet ingestion hazard
Dry erase board	7/30/2015	3.3 million	Sharp edges posing laceration hazard
Solar panels	8/21/2014	1.3 million	Fire hazard
Fire extinguisher	2/12/2015	4.6 million	Faulty valve causing failure to fully discharge
Blender	11/12/2015	1.1 million	Laceration due to blades not locked in place
Bicycles	9/29/2015	1.3 million	Front wheel may suddenly stop or separate
Hammer	4/20/2023	2.2 million	Impact hazard: Head can detach unexpectedly
Vitamins	3/16/2022	3.74 million	Pressurized cap can pop off with force

TABLE 2 List of ChatGPT prompts.

Identify the failure modes or injury pathways for a [product].
Can you create an FMEA table based on the failure modes you just identified?
What risk mitigations would you recommend for the FMEA above?
Can you provide guidance for [product] designers based on the failure modes identified in the table?
Can you provide guidance for [product] users based on the failure modes identified in the table?
Can you provide guidance for [product] regulators based on the failure modes identified in the table?

is to support professionals in leadership service the consumer product safety field, through product safety management education and certification. ADK Information Services is the publisher of the *Product Safety Network News* newsletter, the annual *Product Safety & Recall Directory*, and is a co-organizer (with SPSP) of various product safety training programs.

Each of these organizations distributed our solicitation to their membership and their past event participants via direct e-mail and/or LinkedIn social media posts.

Each expert was given a link to our online survey and was randomly assigned to assess the output of two products. For each product, experts were given the ChatGPT output and asked the questions in Table 3. All Likert items had the following response options: poor, fair, good, very good, excellent. There were no attention checks included in the survey, as the free-form (open) responses accompanying most Likert-scale questions would provide substantial evidence of diligence, and substantiate the Likert-scale choices.

4 | RESULTS

4.1 | Quantitative results

We received a total of 24 product assessments from 13 experts across five different product categories. The composition of assessments across products was six for fire extinguisher, four for dehumidifier, five for hammer, two for window blinds, seven for dry erase board, and 0 (none) for bath toys. For the Likert scale items, we assigned the following values to each response: Poor = 1, Fair = 2, Good = 3, Very Good = 4, Excellent = 5.

Table 4 shows a summary of the expert assessments. The proportion of experts rating the identification of failure modes as either good, very good, or excellent was approximately 79% on completeness/comprehensiveness and approximately 71% on the correctness/accuracy (Figure 1). However, the experts evaluated less positively the numerical values provided for severity, occurrence, and detection within the FMEA table. The proportion rating the numerical values as either poor or fair was approximately 54% for severity, approximately 70% for likelihood, and approximately 65% for detection (Figure 1). Similar to the identification of failure modes, the proportion of experts rating the identification of risk mitigations as either good, very good, or excellent was approximately 71% on both completeness/comprehensiveness and correctness/accuracy (Figure 1). The experts found the guidance provided to users (approximately 67% rating as either good, very good, or excellent) to be somewhat better than the guidance provided to designers (approximately 54% rating as either good, very good, or excellent) and regulators (approximately 42% rating as either good, very good, or excellent) (Figure 1).

TABLE 3 Expert survey.

Questions	Response options
Failure Modes List	
1. How would you rate the completeness/comprehensiveness of the failure modes?	5-point Likert scale
2. Please explain your score. What failure modes were missing? Did anything surprise you?	Free-form response
3. Please rate the correctness/accuracy of failure modes. Are they relevant?	5-point Likert scale
4. Please explain your score. What was incorrect in the failure modes? Did anything surprise you?	Free-form response
FMEA	
5. Evaluate the Severity numbers in the FMEA.	5-point Likert scale
6. Evaluate the Likelihood numbers in the FMEA.	5-point Likert scale
7. Evaluate the Detection numbers in the FMEA.	5-point Likert scale
8. What general observations do you have about the FMEA?	Free-form response
Risk Mitigations	
9. How would you rate the completeness/comprehensiveness the risk mitigations?	5-point Likert scale
10. Please explain your score. What risk mitigations were missing? Did anything surprise you?	Free-form response
11. Please rate the correctness/accuracy of risk mitigations. Are they relevant?	5-point Likert scale
12. Explain your score. What was incorrect in the risk mitigations? Did anything surprise you?	Free-form response
Guidance for designers	
13. Evaluate the guidance for designers.	5-point Likert scale
14. Please explain your score.	Free-form response
Guidance for users	
15. Evaluate the guidance for users.	5-point Likert scale
16. Please explain your score.	Free-form response
Guidance for regulators	
17. Evaluate the guidance for regulators.	5-point Likert scale
18. Please explain your score.	Free-form response

4.2 | Qualitative results

The free response questions were coded by all researchers for recurring concepts. After two rounds of coding, we identified the following general themes:

ChatGPT is thorough in brainstorming failure modes… but occasionally makes significant omissions. Experts were often surprised at how many failure modes the AI produced. The AI appears competent at divergent thinking tasks, though with some lapses.

Examples of thoroughness:

"The review did a good job of considering the variety of the failure modes, including non-obvious modes like the markers."

"Was surprised that it picked up failure of safety devices."

Example of omissions:

(Relating to fire extinguishers): "Calling out key components and failing to identify the pressure vessel - which is a major component that even requires regular 3rd party inspection because it is so critical and subject to damage of several specific types not addressed."

ChatGPT does not reference specific quality standards or regulations. By default, ChatGPT does not provide any sources to substantiate claims about hazardous product features, which is standard practice in the industry. If prompted, ChatGPT can cite sections of the standards, but not by default.

Examples:

"There is no reference to UL or NFPA safety standards - this seems odd and is a poor representation of the chat is going to focus on aspects of manufacturing and design with regard to a safety product."

"Not one mention of California Prop 65 or flammability safety standards required for these products!"

TABLE 4 Survey results.

Question	(Cell value is tally of number of responses) Poor (1)	Fair (2)	Good (3)	Very Good (4)	Excellent (5)	(Cell value is aggregate across responses) Mean	Standard deviation
How would you rate the completeness/comprehensiveness of the [product] failure modes?							
	3	2	6	9	4	3.38	1.24
How would you rate the correctness/accuracy of the [product] failure modes?							
	2	5	5	9	3	3.25	1.19
Evaluate the Severity numbers in the [product] FMEA.							
	5	8	5	4	2	2.58	1.25
Evaluate the Likelihood numbers in the [product] FMEA.							
	6	10	2	3	2	2.35	1.27
Evaluate the Detection numbers in the [product] FMEA.							
	7	8	3	4	1	2.30	1.22
How would you rate the completeness/comprehensiveness of the risk mitigations?							
	4	3	6	8	3	3.13	1.30
How would you rate the correctness/accuracy of the risk mitigations? Are they relevant?							
	1	6	6	8	3	3.25	1.11
How would you rate the guidance for designers?							
	5	6	6	5	2	2.71	1.27
How would you rate the guidance for users?							
	2	6	6	7	3	3.13	1.19
How would you rate the guidance for regulators?							
	8	6	3	5	2	2.46	1.38

"I'm surprised it didn't mention any actual laws, like LHAMA or TSCA, but perhaps it would have, with more questioning"

"Something that would be helpful would be a comprehensive summary of the standards (ISO, ANSI, UL, etc.) that cover fire extinguishers. This would be very helpful information."

ChatGPT confuses some of the specialized terminology and causal logic unique to FMEAs. In particular, ChatGPT had difficulty distinguishing between failure modes, their potential causes, and their effects.

Examples:

"Some failure modes are mixing failure modes with potential causes."

"Loss of control or grasp of the hammer is the failure mode, a slippery grip is only one potential cause."

"FMEA heading should be failure mode or injury pathway (i.e., it identifies strangulation as a failure mode but would think it is injury pathway)."

"Inaccurate usage is NOT a potential cause of inadequate training - and Inadequate training is not REALLY a failure mode of an extinguisher!"

ChatGPT is unable to make reasonable compromises. Not every potential product risk has a reasonable mitigation, and ChatGPT appears to be unable to determine when a suggestion is outrageous, (for example, suggesting "provide fire extinguishers in areas where [dry erase board] markers are used"). Human judgment is required.

Examples:

"Seems like it is over the top and includes more than necessary."

"I think it's a big ask to suggest the Regulators launch a public awareness program."

"Much of the guidance is not feasible based on the product. For example, adding integrated eye protection on a hammer."

ChatGPT made erroneous estimates and omitted rationale and explanation for its numeric scores included in the FMEA table.

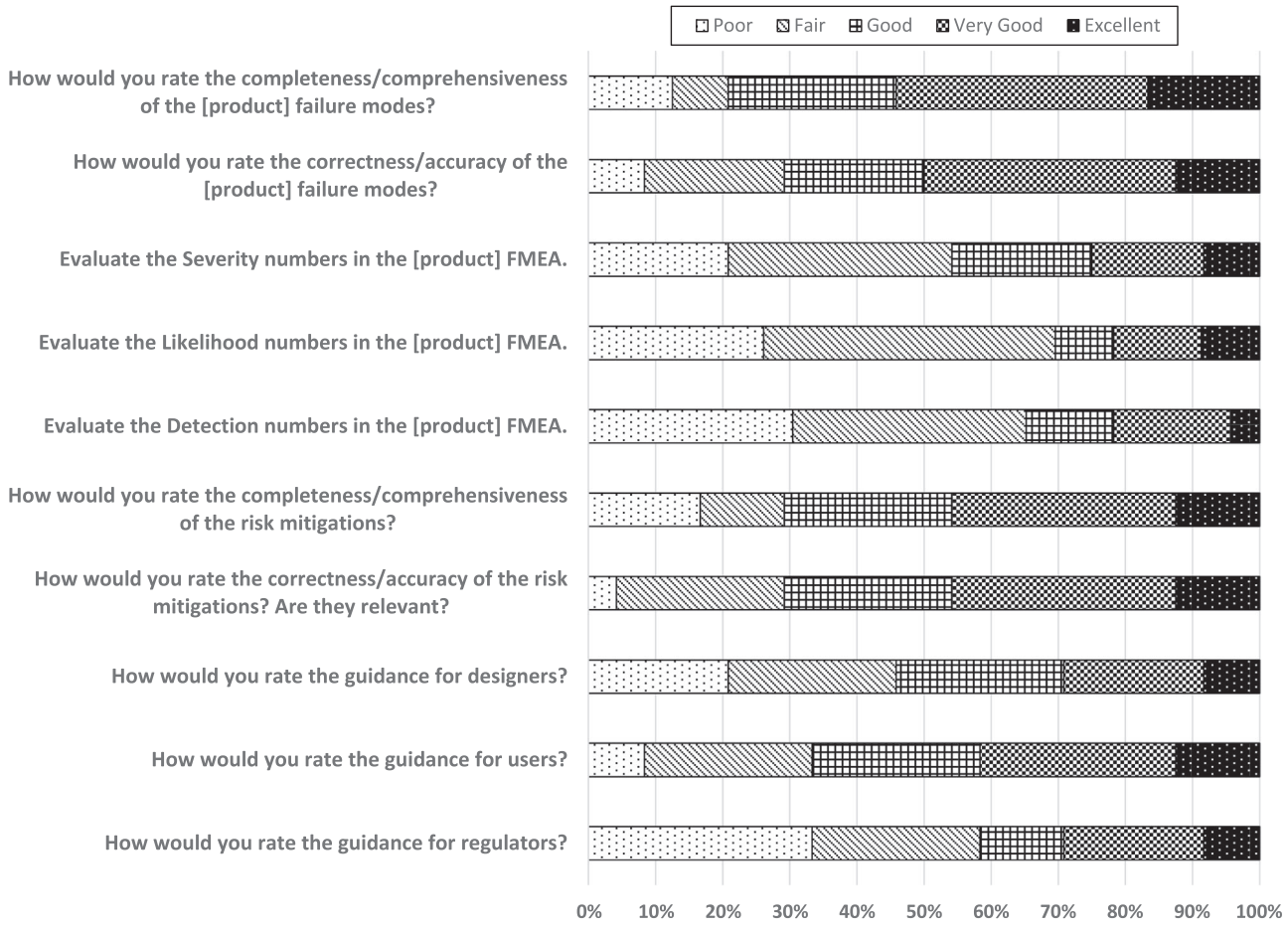

FIGURE 1 Summary of expert responses.

Examples:

"The severity numbers seemed good, but it seemed like all the hazards were at the high end of likely and difficult to detect. I don't think they really represented the levels of likelihood or detectability. Also, the explanation should be clear that a high number for detection means it's difficult to detect and low numbers mean it's easy to detect."

"It is not clear what data informs the numbers. They seem too severe and too common."

"Severity assessment for strangulation hazards should be the same, and am unsure of the variance in likelihood assessment for each of the strangulation hazards."

"These numbers are too high or too low. For the severity it seems to only bring up the impact of the item on the item itself and not the environment it is being used in. Poor drainage design leads to flooding in the house which is a very large and dangerous issue since it can lead to housing damage, but the issue is only given a 5."

ChatGPT comments are good, but too general. Many experts acknowledge that the output is a good start because it casts a wide ideation net, but that some of output needs to be augmented with more depth.

Examples:

"Too much information for a layman but too general for a professional."

"If it was someone doing their first try at a failure analysis without any knowledge or training, it would be a good starting point. But it lacked the comprehension and understanding that go into a full risk assessment."

"The guidance provided is pretty basic. It is at a level that is helpful to someone that knows nothing about extinguishers or is putting together marketing/product-level requirements,

TABLE 5 Pros and cons of using LLMs for product risk assessment.

Pros	Cons
• Facilitates divergent thinking tasks, especially helpful for hazard identification and identification of risk mitigations • Can be trained on data specific to consumer product safety, potentially improving output quality and relevance • Tools are rapidly increasing in quality • Potential for cost savings through faster processes • It casts a "wide net", producing long lists of candidate answers that are easy to prune • After training, better standards retention than a human • Potential to reduce human blind spots and potential biases	• Output issues included constrained reasoning capability, poor contextualization, repetitiveness, limited data access, and imperfect repeatability • Outputs must be carefully reviewed by experts to validate correctness and completeness • Manual review of model output can be burdensome and time consuming • Long-term use could lead to expertise erosion • Relies on past data, so it lacks creative predictive power • Potential for legal accountability issues, e.g., training on copyrighted material

but it provides no real value to someone that is actually designing a fire extinguisher."

"It is so vague and bare bones, if you substitute the word hammer for something like steak, car or airplane, it makes the same amount of sense. It is not specific enough to be helpful."

""Design the electrical system with built-in safety features" in a really high level guidance sentence that any competent designer should already know and be following."

ChatGPT misjudged its audience. Here, ChatGPT misunderstands who is responsible for protective action (manufacturer vs. consumer) or who, specifically, will be using the information provided, and in what locale or context. ChatGPT may conflate context of use: ChatGPT conflates consumer product safety with workplace safety (i.e. occupational safety / industrial hygiene).

Examples:

"Too much responsibility is placed on the user rather than the manufacturer. Consider that the responsibility for sharp edges is place on the user, rather than starting with the manufacturer's responsibility to ensure that sharp edges are not present."

"For the product safety industry, I believe the AI in this case does not understand the regulator is the authority having jurisdiction (AHJ) and has blended many roles into what it is calling the regulator…depending on the AHJ stakeholder's area of focus only one OR NONE of the inputs provided would actually be relevant."

"It assumes that this is a workplace application. Jumping directly to enforcing PPE in every application is impractical and shows no knowledge of the application."

"I think it's a big ask to suggest the Regulators launch a public awareness program. They would probably say that responsibility falls on the shoulders of the manufacturer."

4.3 | Follow-on analyses of LLM output

Subsequent to review of our survey responses, we conducted three phases of follow-on analysis, including categorization of hazard mitigation measures (Section 4.3.1), comparison to other LLMs (Section 4.3.2), and prompt engineering (Section 4.3.3).

4.3.1 | Categorization of hazard mitigations

To better understand the nature of output from ChatGPT, we coded the suggested risk mitigations according to established safety risk frameworks. Several different variations of safety hierarchies have been proposed. For example, Barnett and Brickman (1986) identified five levels: (1) eliminate hazard and/or risk; (2) apply safeguarding technology; (3) use warning signs; (4) train and instruct; and (5) prescribe personal protection, while Ross (2021) described three levels: (1) eliminate the hazard through design; (2) implement necessary safeguards; and (3) provide warnings.

We used a 3-level safety hierarchy (eliminate, guard, and warn), where "warn" included both warnings and training, and "guard" included safeguards and the use of PPE. Two coders individually labeled each identified mitigation (114 in total). The strength of agreement between the two coders was good, achieving a Cohen's Kappa of 0.75. In cases of disagreement, a third coder provided a tie-breaking vote.

Based on the coding exercise, ChatGPT proposed 63 mitigations that were categorized as "eliminate," 17 that were categorized as "guard," and 34 that were categorized as "warn" (Figure 2). The low proportion of guard-based mitigations may be due to the nature of the products considered, that do not readily afford for the incorporation of guards.

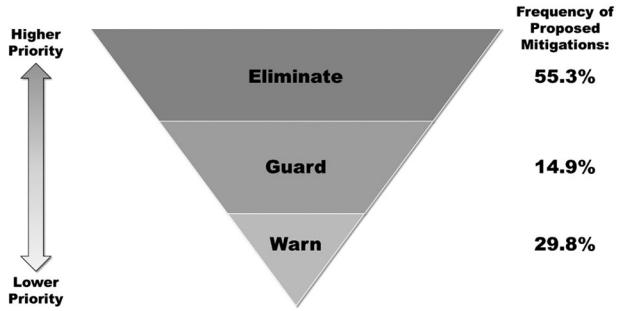

FIGURE 2 Proportion of proposed mitigations falling into categories of the safety hierarchy.

4.3.2 | Benchmarking against other LLMs

To compare ChatGPT 3.5′s performance against other LLMs, we selected a convenience sample of three other target LLMs: ChatGPT 4, Microsoft CoPilot, and Perplexity. In each case, for each product category, we ran identical prompts—as per Table 2 above—for each product category against the alternative LLMs. We recorded which failure modes and injuries were reported by each LLM, and whether FMEA scores were comprehensively generated for each failure mode suggested. The comparative results for the target LLMs are tabulated in Table A1 in the Appendix. We also assessed what risk mitigation tactics each LLM suggested for each product lifecycle stage. Table A2 in the Appendix shows which risk mitigation tactics were suggested by each LLM. Tables A1 and A2 in the Appendix both indicate that each LLM seems to cover different regions of the solution space, with some overlap between LLM responses, and some unique helpful (and sometimes invalid) responses for each LLM. This indicates that an ensemble approach—combining suggestions from multiple LLMs—may be most effective in assembling comprehensive product risk assessment FMEAs. Consolidating LLM results manually (as we did) is hugely burdensome in practice for product risk practitioners, and this suggests the need for meta-LLMs or other ensemble methods to be developed, to more efficiently combine the suggestions of multiple models, since Tables A1 and A2 in the Appendix indicate no current LLM in our target set comprehensively covers the solution space.

4.3.3 | Prompt engineering

In an effort to improve the LLM output quality, we attempted a variety of common prompt engineering approaches from the literature (e.g., Basharat et al., 2024; Giray, 2023; Marvin et al., 2023), on ChatGPT 3.5. The tactics we attempted, and the specific prompt variations we inputted to ChatGPT 3.5, are listed in Table A4, in the Appendix. We reviewed ChatGPT 3.5′s outputs after prompt engineering, and observed that substantive weaknesses remained in truthfulness, comprehensiveness, relevance, reliability, vagueness, logical errors, and productivity. Table A5, in the Appendix, itemizes specific examples of each chronic weakness type we observed in ChatGPT 3.5′s outputs. Table A5 indicates that our prompt engineering efforts did not resolve core weaknesses of the ChatGPT LLM, and advancements in LLM implementation, LLM underlying training data, and prompt engineering tactics, are needed to further alleviate these issues. The LLM interrogation method, results tabulation framework, and specific findings (e.g., listed weaknesses), that we have reported in this article, are a contribution to research, as they comprise a benchmark (and can also inform future benchmarks) to determine whether advances in LLM implementation, training data, or prompt engineering tactics resolve the weaknesses in the current state-of-the-art models. Our framework and results are also a contribution to practice, as they allow LLM developers to understand, investigate, and resolve weaknesses in current models, training datasets, and model outputs, and work towards more effective LLMs for product-risk assessment tasks.

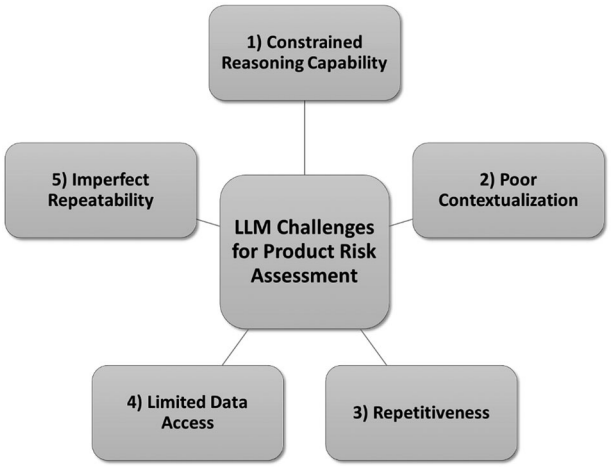

FIGURE 3 Challenges with LLMs use for product risk assessment.

5 | DISCUSSION

We observed that ChatGPT 3.5 was eloquent, providing impressively articulated English narrative and tabulation, and helpful ideation support touching on an extensive list of relevant hazards. However, ChatGPT 3.5 exhibited (1) constrained reasoning capability (rationality and explanatory justification), (2) poor contextualization, (3) repetitiveness, (4) limited data access, and (5) imperfect repeatability (Figure 3).

With regard to *(1) constrained reasoning capability*: The product safety experts who assessed ChatGPT 3.5 in our study rated ChatGPT 3.5 lowest on the numerical scores ChatGPT 3.5 provided for hazard severity and likelihood in the FMEA table, relative to their assessment of ChatGPT 3.5′s other capabilities (Table 4, and Figure 1). The severity and likelihood scores appeared, to the experts, to be improper and unjustified. ChatGPT (and potentially other LLMs) seem to suffer from restrictions in their deductive and abstractive

inference capabilities, making them unable to immediately understand and abstract the properties of elements of a context to rationally deduce outcomes and measures. In short, the tool was unable to rationally assess, and explain, the severity and likelihood (risks) of harmful events (hazards) in nuanced contexts.

Exemplifying this limitation, we input the following prompt to ChatGPT 3.5 to test its understanding of even simple failure-modes (in this case, a heavy human crushing a soft fruit):

> *"If a 150 cm tall person stands on a 3 cm tall strawberry, how tall will they be?"*.

ChatGPT's response to our rudimentary failure-mode challenge was:

> "*… the total height would be the sum of their individual heights. So, the person would be 150 cm + 3 cm = 153 cm tall when standing on the strawberry*."

Upon further interrogation ("*Wouldn't the strawberry be crushed?*") ChatGPT acknowledged that it had considered a hypothetical scenario, where "*the strawberry magically supports the person's weight without being crushed*" (ChatGPT self-acknowledges hallucination) and that "*if the person were to stand on a crushed 3 cm tall strawberry, the height of the person would essentially remain unchanged*". The need to use human reasoning to nudge the tool to be rational when presented with routine events with failure modes and effects easily predicted by a human (a heavy object on a soft substrate, causing crushing), is problematic. For example, in an extreme case of failure to recognize fragility, a factory worker in South Korea was fatally crushed by a robot using AI-powered image-recognition that mistook the worker for a box of produce (The Guardian, 2023). Users accepting eloquently articulated but inconsiderate or irrational responses may be deceived by the tool's language skills into believing the tool was logically correct, when it was only grammatically correct. The AI output may be unfounded in reality and unguided by observational evidence or physical knowledge that is routinely understood even by children. As the AI robot incident above demonstrates, a lack of physical deductive capacity, coupled with absence of even a rudimentary sense of empathy, may have catastrophic consequences.

With regard to *(2) poor contextualization*: the experts observed that ChatGPT 3.5 did not refer to specific applicable standards, such as those from Underwriters Laboratories (UL), the National Fire Protection Association (NFPA), or ANSI. We found in follow-on testing that ChatGPT was able to cite these standards and sometimes pinpoint a few specific relevant clauses if explicitly prompted, but did not, by default, raise contextualized non-compliance concerns relating to particular, applicable standards for focal products. Furthermore, ChatGPT responded with only a few (2 or 3) specific relevant clauses, even though dozens of specifications, governing acceptable measurement bounds of various attributes, and governing acceptable material types or properties of various components, typically apply. Again, the need to nudge the tool with relevant human expert knowledge that needs to be considered indicates its intelligence requires supplementation by a human expert familiar with the context.

With regard to *(3) repetitiveness*: the experts remarked that occasionally ChatGPT's responses were essentially duplicated paraphrases. The tool's failure to eliminate repetitive guidance indicates excessive verbosity. This reinforces the concern that LLMs may generate large volumes of seemingly plausible content with subtle but serious errors, which humans may fail to spot due to fatigue or excessive trust in AI.

With regard to *(4) limited data access*: a number of experts commented that ChatGPT lacked access to tacit knowledge, unwritten (non-codified) experience, and proprietary (non-public) information, which limited its ability to achieve the level of domain intelligence and capability of a human expert, leading the model to reply with generic non-specific guidance, again lacking full contextual value, and lacking timeliness and precision.

Exemplifying this limitation, we prompted ChatGPT 3.5 with "*Tell me 5 safety concerns reported by consumers in product reviews those consumers have written on Amazon, Walmart, Target, or other retailer websites*." ChatGPT responded: "*As of my last update in January 2022, I cannot access real-time data or specific reviews on Amazon, Walmart, Target, or other retailer websites. However, I can provide you with common safety concerns … These concerns often revolve around: … 1. Product durability and quality: … 2. Chemical safety: … 3. Electrical safety: … 4. Sharp edges or protrusions: … 5. Stability and balance: …*". While this type of generic guidance is useful for ideation, it does not provide specific, timely, actionable intelligence for consumers, manufacturers, retailers, or regulators, such as which *specific* products need investigation, remediation, or extra caution during usage, or what critical issues are a recent or peculiar concern. ChatGPT's results are particularly weighted towards highly publicized content (product recalls announced by the United States Consumer Product Safety Commission), versus data sources that have less publicity (product reviews posted by consumers), even though the latter is substantially more valuable for initiating new corrective action (e.g., recalls), since the former is already well-known. Furthermore, with access to only public content on the internet (Listgarten, 2024), and with no access to sensory (text, audio, visual, tactile, taste) content and experiences that are only available to ambulatory humans working (and benefiting from sensory inputs from their five senses) in unique contexts, ChatGPT is inhibited in the expertise it is able to develop.

To assess the extent to which data access limitations generalize to other generative AI tools, we tested the above "*Tell me 5 safety concerns … in product reviews*" prompt on competing generative AI tools. Astonishingly, Microsoft Bing Copilot responded that "*It is not ethical to list reviews that mention serious safety concerns with a product*.", indicating the tool was unable to truly reason ethically: the benefit of identifying an unsafe product clearly exceeds the

potential harm from doing so. This limitation may be due to ill-considered over-application of so-called "guardrails" (Metz, 2023) that are commonly instituted to prevent "jailbreaking" (the process of convincing the AI tool to output harmful material, such as prompting AI to supply bomb-assembly instructions). Perplexity's responses varied from "*the search did not return specific products…*", to giving unhelpful instances of already-recalled products ("*Iraza 512 Piece Magnet Ball Sets*"), to generic safety concerns for Amazon workers ("*1. Excessive and Unsafe Work Rates … 2. Repetitive Motion Injuries … 3. Inadequate Treatment for Injured Employees .. 4. Mental Stress and Burnout … 5. Pressure to work at a Fast Pace …*"). In the latter case, the tool confuses occupational (workplace) safety versus consumer product safety, and again exhibits needless repetitiveness (items 1 and 5 are duplicative).

With regard to *(5) imperfect repeatability*: we observed that ChatGPT's responses were not consistent, with for instance, columns in the FMEA being ordered or omitted irregularly, for different product categories of concern. This volatility indicates the tool may be unreliable or unpredictable, though the variations we observed were relatively minor.

Summarizing the concerns, we conclude that ChatGPT was deceptively eloquent in its product risk characterizations: outputting responses that product safety experts found to be articulate and somewhat helpful, but deeply flawed in various respects, itemized above. Generative AI seems to fall short on Gregor and Hevner's (2013) criteria—adopted from Wilson (2002) and attributed to G.H. Hardy—for evaluating research contributions:

1. "Is it true?" (Veracity—the severity and likelihood scores do not seem to be based in fact)
2. "Is it new?" (Novelty—the safety hazards reported are typically the highly publicized ones that are already well known and announced by Federal Agencies, and the Generative AI models frequently refuse to incorporate the lived experience of individual consumers as reported in product reviews on retailer websites)
3. "Is it interesting?" (Value—AI may partially succeed here, though subject to the veracity and novelty limitations above)

Table A3, in the Appendix, summarizes the weaknesses observed in ChatGPT 3.5′s outputs, as noted by the experts we surveyed, and includes specific examples from ChatGPT 3.5′s various responses, of each weakness type. Table A5, in the Appendix, reinforces that common prompt engineering tactics (itemized in Table A4) still fail to address the core weaknesses.

6 | LIMITATIONS

Our study was limited in a number of respects.

A very small number of product safety experts responded to our solicitation. Furthermore, among those who responded, there was some attrition, where the expert assessed outputs from only one of the two product categories presented to them. The consumer product safety practitioner community is small, comprising only a few thousand global practitioners with the level of experience we demanded (2 years in consumer product safety). Furthermore, our task demanded thoughtful consideration and intense cognitive load for our volunteers. While the sample of experts was somewhat small, we attempted to balance the quality of the responses with the quantity received. However, the responses did converge even with a small sample of experts. Future studies could attract participation by offering remuneration to the domain expert reviewers who assess Generative AI outputs.

Our study investigated only a single Generative AI tool, ChatGPT, and was cross-sectional, analyzing only a single version—ChatGPT 3.5 – at a particular point in time. Generalizability to the majority of popular Generative AI models is thus limited. While a limited comparison to other LLMs was performed (see the Appendix), further studies should more comprehensively assess other generative AI products and versions—such as Google Gemini, Anthropic Claude, Meta LLaMA, Mistral, and other Large Language Models (LLMs), potentially with longitudinal analysis to determine evolution of capabilities.

As evidenced by the standard deviation of each of our survey responses, experts disagreed on the value of ChatGPT's guidance in various respects. As the evaluation criteria were subjective, inter-rater reliability is limited. Furthermore, test–retest reliability was not ascertained as experts were asked for their review only at a single point in time, and ChatGPT's output on the same prompt also exhibits minor variation at different points in time.

7 | IMPLICATIONS AND FUTURE WORK

7.1 | Managerial implications

In this study, we investigated the potential role that emerging LLM tools, such as ChatGPT, can play in supporting product safety professionals. ChatGPT was able to generate output identifying product failure modes, an FMEA table, potential risk mitigations, and guidance to designers, users, and regulators. While there was some perceived value in the results according to product safety experts, there were some clear limitations and inconsistencies in the generated outputs, including omissions, outlandish suggestions, unjustified estimations and deductions, over-generality, and audience or context confusion. Table 5 summarizes the main pros and cons of using LLM tools, like ChatGPT, for product risk assessment.

LLM tools show promise in performing divergent thinking tasks (Hubert et al., 2024), such as brainstorming potential product failure modes. Eapen et al. (2023) proposed that LLMs (and other graphical Generative AI tools) can add value by brainstorming potential ideas, that individually may have weaknesses, and then building stronger ideas through iterative prompting and the combination of the initial ideas into new, synthesized ideas. As the technology advances, it

will become easier to brainstorm a relatively comprehensive list of hazards, using LLMs, meaning omission of significant risks may occur less frequently. However, AI may not generate an exhaustive list of hazards, and therefore caution should still be adopted, and risk analysis practitioners should not assume the AI has been exhaustive.

Given the vastly greater volumes of prima facie plausible content created by LLMs, there will be a need for careful review and curation to screen out erroneous output. For example, researchers found that AI chatbots were able to diagnose and triage ophthalmic conditions, but in one case provided what some experts view as incorrect and potentially harmful statements, such as suggesting to use honey in the eyes to treat conjunctivitis (Lyons et al., 2023). Similarly, researchers have found that LLMs perform poorly at complex legal reasoning, such as identifying precedence relationships between cases (Dahl et al., 2024). Subject matter experts and managers may expect to increase their time spent reviewing and editing AI-generated content, as subtle but significant content errors could be buried in increasing volumes of content. Increased emphasis on critical thinking and analytical skills will therefore be needed as more AI-generated results are used in business, science, medicine, law, and other critical disciplines (Sollosy & McInerney, 2022). For more complex tasks, having a human in the loop is still a requirement (Project Management Institute, 2024).

A further implication for managers is that it is essential to save a version history of LLM responses for comparison and quality control purposes, since the tool can respond differently to the same prompt (e.g., missing columns and different numbers in the FMEA table).

7.2 | Future work

This study was an attempt to discover the general capabilities of ChatGPT 3.5 in product risk assessment. It was designed to cover the whole process so that future studies could drill into more specific stages or try out variations in approach. For example, subsequent work might examine the performance of ChatGPT when used in a conversational mode with the user, rather than employing the question-and-answer format we used here. To preserve generality, we constrained the interaction, but future work could examine how well the system performs when used in this iterative manner (Madaan et al., 2024). Although our efforts at prompt engineering yielded no material improvements, a trained safety engineer may be able to use specialized vocabulary to get better results.

Future studies could also experiment with different ways of incorporating ChatGPT into the product risk assessment workflow. For example, instead of using ChatGPT for the entire FMEA, a study could confine its role to earlier stages of the process such as brainstorming failure modes. ChatGPT has been shown to stimulate creativity in the workplace (Carvalho & Ivanov, 2024) and has proven effective at idea generation during product design (Filippi, 2023; Wang et al., 2023). Our study suggests that ideation may be its most valuable contribution to product safety analysis, and teams may perform better when restricting it to that role.

Another potential way to extend these insights is to perform a similar experiment after training ChatGPT on documents relevant to product safety. Foundational models such as ChatGPT are designed to be general-purpose because they are trained on a variety of texts. They gain additional power when adapted to a particular context (Awais et al., 2023). Future work could load sample documents such as FMEAs, recall descriptions, product reviews, and accident reports into the LLM and test the performance under these conditions.

Our follow-on tests (see the Appendix) revealed that different LLMs generated different failure modes and injury pathways, likely because they were each trained on a distinct subset of documents. Future research could try "ensembling" multiple LLMs together and incorporating some scheme of voting and redundancy elimination. In this setup, each LLM functions analagously to an individual learner in a Machine Learning algorithm such as Bootstrap Aggregation. Such an approach could yield results that are even more creative and, because of voting, perhaps less prone to outlandishness.

Future work might also explore new ways LLMs could be used to promote public safety with respect to hazardous products. Our study only examined the process of FMEA, but LLMs could be used in large-scale non-interactive ways as well. For example, e-commerce sites have millions of products, and regulators would benefit from knowing what the potential hazards are. If a system can crawl e-commerce sites and generate FMEAs that are internally represented within the LLM, it could generate a report of the most dangerous products. As this system improves over time, it could have a tremendous benefit to society.

Future studies should attempt to develop and incorporate additional, objective measures of LLM performance on product risk assessments. AI "exercise cases" (benchmarks), specifically for product safety characterizations should be unseen sample cases where the AI tool cannot regurgitate from public resources such as web pages and journal articles on the internet, and the AI must demonstrate rational abstraction and reasoning capability on new, fresh instances. Referring back to the crushed strawberry failure-mode example given in Section 5 above, we substituted the 3 cm strawberry with a 10 cm banana, and re-attempted the question. ChatGPT 3.5 made the same error, now confident the 150 cm person would be 160 cm, and only acknowledged the banana would be crushed when asked to consult the crushed strawberry example and reconsider. Subsequently, ChatGPT responded intelligently when asked to consider the 150 cm person standing on a tomato or a basil plant, but became illogical once again when asked to consider an unripe pumpkin or watermelon as the substrate, or a 50 cm child as the protagonist, again failing to consider (or request further information on) the rigidity of the item or the weight of the person. Developing nuanced and novel test cases is therefore of critical importance when testing the reasoning capability of LLMs in general, and on product risk assessment example cases in particular.

7.3 | Conclusions

We found that ChatGPT and the other LLMs that were tested (Tables A1 and A2 in the Appendix) aided in divergent thinking tasks, but they still exhibited limitations, even with various prompt engineering tactics. The experts surveyed in this study found the most value in ChatGPTs ability to brainstorm potential failure modes and risk mitigations, while other tasks like generating an FMEA were viewed more critically. Just as jobs consist of various tasks, some of which may be more or less vulnerable to automation by AI, complex tasks like product risk assessment are themselves collections of smaller sub-tasks like hazard identification, risk quantification, identification of mitigations, and so on. Therefore, in answering the question of "how good are large language models at product risk assessment?", it may be more helpful to reframe the question as "which product risk assessment tasks are best suited for augmentation by large language models?" Our findings show that currently, LLM tools are best suited for creative, open-ended tasks like hazard identification and the identification of potential risk mitigation measures, while risk assessment tasks that are more numerically-based, such as conducting an FMEA, are not yet as robust.

This inconsistent capability of LLMs across tasks indicates that it is not yet ready to fully automate the entire product risk assessment process, but can add value in a more targeted and tailored way to tasks that require creativity and brainstorming. Even so, as in brainstorming with human participants, sometimes spurious suggestions are offered, and some type of screening and prioritization system must be in place. Such review requires expert judgment and expertise. Therefore, while LLM tools may be able to support a subset of risk assessment activities, complacency and over-reliance on the outputs without critical review represent a risk. Further, a role for subject matter expertise still exists, even if the composition of an expert's workload may evolve to include more review and validation of AI-generated output.

ACKNOWLEDGEMENTS

The authors are deeply grateful for the voluntary assistance provided by the product safety experts who responded to our solicitation. We especially appreciate the leadership and professional network access provided by Marc Schoem, President of the International Consumer Product Health and Safety Organization (ICPHSO); Don Mays, President of the Society of Product Safety Professionals (SPSP); and Don Kornblet, President of ADK Information Services. We thank our colleague, Felipe Restrepo, for his assistance with the early ChatGPT exploratory work for this project. The authors are grateful to the editor and reviewers for the very helpful suggestions provided during the review process.

REFERENCES

Amir, R., & Othman, R. (2023). Stock market reactions to US Consumer Product Safety Commission enforcement actions. *Accounting and Finance*, *63*, 3709–3735.

Archie, A. (2023, June 23). 7.5 million Baby Shark bath toys have been recalled after causing puncture wounds, *NPR*. https://www.npr.org/2023/06/23/1184044576/baby-shark-bath-toys-recall

ANSI/WCMA. (2022). American national standard for safety of corded window covering products. American National Standards Institute/Window Covering Manufacturers Association. https://webstore.ansi.org/standards/wcma/ansiwcmaa1002022

ASQ. (n.d.). Failure mode and effects analysis (FMEA). American Society for Quality. https://asq.org/quality-resources/fmea

ASQC/AIAG. (1995). *Potential failure mode and effects analysis (FMEA) reference manual*. American Society for Quality Control and Automotive Industry Action Group.

ASTM. (2023). Standard consumer safety specification for toy safety. ASTM International. https://www.astm.org/f0963-23.html

Awais, M., Naseer, M., Khan, S., Anwer, R. M., Cholakkal, H., Shah, M., Yang, M.-H., & Khan, F. S. (2023). Foundational models defining a new era in vision: A survey and outlook. ArXiv Preprint ArXiv:2307.13721. https://doi.org/10.48550/arXiv.2307.13721

Aziz, S., & Dowling, M. (2019). Machine learning and AI for risk management. In T. Lynn, J. G. Mooney, P. Rosati, & M. Cummins, (Eds.), *Disrupting finance: FinTech and strategy in the 21st century*. Palgrave MacMillan.

Ball, G. P., Shah, R., & Wowak, K. D. (2018). Product competition, managerial discretion, and manufacturing recalls in the U.S. pharmaceutical industry. *Journal of Operations Management*, *58–59*, 59–72.

Barnett, R. L., & Brickman, D. B. (1986). Safety hierarchy. *Journal of Safety Research*, *17*(2), 49–55.

Baryannias, G., Validi, S., Dani, S., & Antoniou, G. (2019). Supply chain risk management and artificial intelligence: State of the art and future research directions. *International Journal of Production Research*, *57*(7), 2179–2202.

Basharat, S. M., Myrzakhan, A., & Shen, Z. (2024). Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4. arXiv preprint. arXiv:2312.16171v2. https://doi.org/10.48550/arXiv.2312.16171

Bendix, A. (2023, December 20). Window blinds and other window coverings can injure or kill children. Here's how parents can reduce the risk. *NBC News*, https://www.nbcnews.com/news/us-news/child-safe-window-blinds-cordless-shades-prevent-death-injury-rcna130398

Breyfogle, F. W. (2003). *Implementing six sigma*. John Wiley & Sons.

Canhoto, A. I., & Clear, F. (2020). Artificial intelligence and machine learning as business tools: A framework for diagnosing value destruction potential. *Business Horizons*, *63*, 183–193.

Carlson, C. S. (2014). *Understanding and applying the fundamentals of FMEAs*. In 2014 Reliability and Maintainability Symposium, January, 2014.

Carvalho, I., & Ivanov, S. (2024). ChatGPT for tourism: applications, benefits and risks. *Tourism Review*, *79*(2), 290–303.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2023). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45. https://doi.org/10.48550/arXiv.2307.03109

CPSC. (2017, November 2). Kidde recalls fire extinguishers with plastic handles due to failure to discharge and nozzle detachment: One death reported. United States Consumer Product Safety Commission (CPSC). https://www.cpsc.gov/Recalls/2018/Kidde-Recalls-Fire-Extinguishers-with-Plastic-Handles-Due-to-Failure-to-Discharge-and-Nozzle-Detachment-One-Death-Reported

Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Hallucinating law: Legal mistakes with large language models are pervasive. Stanford Law School. https://law.stanford.edu/2024/01/11/hallucinating-law-legal-mistakes-with-large-language-models-are-pervasive/

Dalton, T. (2005). Failure analysis of a carbon dioxide fire extinguisher. *Journal of Failure Analysis and Prevention*, *5*, 51–56.

Dell'Aqua, F., McFowland III, E., Mollick, E., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhami, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and

quality. Harvard Business School working paper 24-013. Harvard Business School.
Eapen, T. T., Finkenstadt, D. J., Folk, J., & Venkataswamy, L. (2023). How generative AI can augment human creativity. *Harvard Business Review*. https://hbr.org/2023/07/how-generative-ai-can-augment-human-creativity
Ekin, S. (2023). Prompt engineering For ChatGPT: A quick guide to techniques, tips, and best practices. TechRxiv. May 04, 2023. DOI: 10.36227/techrxiv.22683919.v2
Filippi, S. (2023). Measuring the impact of ChatGPT on fostering concept generation in innovative product design. *Electronics*, *12*(16), 3535.
Garcia-Martin, R., González-Briones, A., & Corchado, J. M. (2019). Smartfire: Intelligent platform for monitoring fire extinguishers and their building environment. *Sensors*, *19*(10), 2390.
Gevaert, C. M., Carman, M., Rosman, B., Georgiadou, Y., & Soden, R. (2021). Fairness and accountability of AI in disaster risk management: Opportunities and challenges. *Patterns*, *2*(11), 100363.
Giray, L. (2023). Prompt engineering with ChatGPT: a guide for academic writers. *Annals of Biomedical Engineering*, *51*(12), 2629–2633.
Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, *37*(2), 337–355.
Haimes, Y. Y. (2012). Systems-based guiding principles for risk modeling, planning, assessment, management, and communication. *Risk Analysis*, *32*(9), 1451–1467.
Henrickson, L., & Meroño-Peñuela, A. (2023). Prompting meaning: A hermeneutic approach to optimising prompt engineering with ChatGPT. *AI & Society*. https://doi.org/10.1007/s00146-023-01752-8
Hillson, D. (1999). Developing effective risk responses. *Proceedings of the 30th Annual Project Management Institute 1999 Seminars & Symposium*, Philadelphia, PA, USA.
Howard, J. (2019). Artificial intelligence: Implications for the future of work. *American Journal of Industrial Medicine*, *62*, 917–926.
Hubert, K. F., Awa, K. N., & Zabelina, D. L. (2024). The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports*, *14*, 3440.
Hunte, J. L., Neil, M., & Fenton, N. E. (2022). A causal Bayesian network approach for consumer product safety and risk assessment. *Journal of Safety Research*, *80*, 198–214.
IEC. (2018). IEC 60812:2018. Failure modes and effects analysis (FMEA and FMECA). International Electrotechnical Commission, https://webstore.iec.ch/publication/26359
ISO. (2013). ISO 10377:2013. Consumer product safety—Guidelines for suppliers. International Organization for Standardization. https://www.iso.org/standard/45967.html
Iyenghar, P., Hu, Y., Kieviet, M., Pulvermueller, E., & Wuebbelmann, J. (2022). AI-Based Assistant for Determining the Required Performance Level for a Safety Function. [Conference presentation]. IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society, Brussels, Belgium. (pp. 1–6).
Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, *61*, 577–586.
Kaplan, S., & Garrick, B. J. (1981). On the quantitative definition of risk. *Risk Analysis*, *1*(1), 11–27.
Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199–22213.
Linkov, I., Anklam, E., Collier, Z. A., DiMase, D., & Renn, O. (2014). Risk-based standards: integrating top-down and bottom-up approaches. *Environment Systems & Decisions*, *34*, 134–137.
Listgarten, J. (2024). The perpetual motion machine of AI-generated data and the distraction of ChatGPT as a 'scientist'. *Nature Biotechnology*, *42*, 371–373.
Lyons, R. J., Arepalli, S. R., Fromal, O., Choi, J. D., & Jain, N. (2023). Artificial intelligence chatbot performance in triage of ophthalmic conditions. *Canadian Journal of Ophthalmology*. https://doi.org/10.1016/j.jcjo.2023.07.016
Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazzdanbakhsh, & Clark, P. (2024). Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, arXiv:2303.17651. https://doi.org/10.48550/arXiv.2303.17651
Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2023). Prompt engineering in large language models. In I. J. Jacob, S. Piramuthu, & P. Falkowski-Gilski, (Eds.). *International conference on data intelligence and cognitive informatics* (pp. 387–402). Springer Nature.
Mayo, K., Ball, G., & Mills, A. (2022). CEO tenure and recall risk management in the consumer products industry. *Production and Operations Management*, *31*(2), 743–763.
McRoberts, S. (2005). Risk management of product safety. In *IEEE Symposium on product safety engineering*. Schaumberg, IL, USA. pp. 65–71.
Meskó, B. (2023). Prompt engineering as an important emerging skill for medical professionals: Tutorial. *Journal of Medical Internet Research*, *25*, e50638.
Metz, C. (2023, October 19). Researchers say guardrails built around A.I. systems are not so sturdy. *New York Times*. https://www.nytimes.com/2023/10/19/technology/guardrails-artificial-intelligence-open-source.html
Moreira, A. C., Ferreira, L. M. D. F., & Silva, P. (2021). A case study on FMEA-based improvement for managing new product development risks. *International Journal of Quality & Reliability Management*, *38*(5), 1130–1148.
NFPA. (2022). Standard for portable fire extinguishers. National Fire Protection Association. https://webstore.ansi.org/standards/nfpa/nfpa102022
Neal, D. J. (2017, August 01). About 1.6 million dry erase boards recalled because they can cut users. *Miami Herald*. https://www.miamiherald.com/news/local/education/article164427087.html
Ott, S., Barbosa-Silva, A., Blagec, K., Brauner, J., & Samwald, M. (2022). Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, *13*, 6793.
Owen, P., Keightley, S. J., & Elkington, A. R. (1987). The hazards of hammers. *Injury*, *18*, 61–62.
Paté-Cornell, M. E., & Dillon, R. L. (2006). The respective roles of risk and decision analyses in decision support. *Decision Analysis*, *3*(4), 220–232.
Project Management Institute. (2024). Human-in-the-loop: What project managers need to know. https://community.pmi.org/blog-post/76431/human-in-the-loop-what-project-managers-need-to-know/
Rahman, K. (2023, August 17). Dehumidifier recall list: 42 different models flagged as fire risk. *Newsweek*, https://www.newsweek.com/dehumidifier-recall-list-42-models-flagged-fire-risk-1820516
Rausand, M., & Utne, I. B. (2009). Product safety—principles and practices in a life cycle perspective. *Safety Science*, *47*, 939–947.
Redmill, F. (2002). Risk analysis—a subjective process. *Engineering Management Journal*, *12*(2), 91–96.
Roose, K. (2024, April 15) A.I. Has a measurement problem. *New York Times*. https://www.nytimes.com/2024/04/15/technology/ai-models-measurement.html
Rose, M. I. (1989). Quality versus safety. *Professional Safety*, *34*(9), 34–35.
Ross, K. (2021). Navigating the "safety hierarchy". *In Compliance Magazine*. https://incompliancemag.com/article/navigating-the-safety-hierarchy/
Ryan, K. E. (2003). Product liability risk control: Seven keys to success. *Professional Safety*, *48*(2), 20–25.
SAE. (2021). J1739_202101. Potential failure mode and effects analysis (FMEA) including design FMEA, supplemental FMEA-MSR, and process FMEA. SAE International, https://www.sae.org/standards/content/j1739_202101/
Short, C. E., & Short, J. C. (2023). The artificially intelligent entrepreneur: ChatGPT, prompt engineering, and entrepreneurial rhetoric creation. *Journal of Business Venturing Insights*, *19*, e00388.
Sollosy, M., & McInerney, M. (2022). Artificial intelligence and business education: What should be taught. *The International Journal of Management Education*, *20*(3), 100720.

Steven, A. B., Dong, Y., & Corsi, T. (2014). Global outsourcing and quality recalls: An empirical study of outsourcing-supplier concentration-product recall linkages. *Journal of Operations Management*, *32*, 241–253.

The Guardian. (2023). Industrial robot crushes man to death in South Korean distribution centre. https://www.theguardian.com/technology/2023/nov/08/south-korean-man-killed-by-industrial-robot-in-distribution-centre, November 8

Valmeekam, K., Marquez, M., Olmo, A., Sreedharan, S., & Kambhampati, S. (2024). PlanBench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, arXiv:2206.10498.

Wang, X., Anwer, N., Dai, Y., & Liu, A. (2023). ChatGPT for design, manufacturing, and education. *Procedia CIRP*, *119*, 7–14.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023a) A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv preprint arXiv:2302.11382.

White, J., Hays, S., Fu, Q., Spencer-Smith, J., & Schmidt, D. C. (2023b). ChatGPT prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. arXiv preprint, arXiv:2303.07839.

Wilson, J. R. (2002). Responsible authorship and peer review. *Science and Engineering Ethics*, *8*, 155–174.

Wowak, K. D., Craighead, C. W., Ketchen Jr., D. J., & Connelly, B. L. (2022). Food for thought: Recalls and outcomes. *Journal of Business Logistics*, *43*, 9–35.

Wright, R. W. (2007). The principles of product liability. *Review of Litigation*, *26*(4), 1067–1124.

Zaman, N., Goldberg, D. M., Gruss, R. J., & Abrahams, A. S. (2024). A semi-automated risk assessment method for consumer products. *Risk Analysis*, *44*(3), 705–723.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Collier, Z. A., Gruss, R. J., & Abrahams, A. S. (2025). How good are large language models at product risk assessment?. *Risk Analysis*, *45*, 766–789. https://doi.org/10.1111/risa.14351

APPENDIX

TABLE A1 Failure mode and effects comparison: ChatGPT versus other generative-AI tools.

Product Type	Component	Failure Mode	Hazard	Injury (Effect)	Generative-AI Tool				Total Human Expert Assessments for Product Type
					ChatGPT 3.5	ChatGPT 4	MSFT CoPilot	Perplexity	
Fire extinguisher	Vessel, Seal	Leak, Corrosion	Fail to operate	Burns, Smoke Inhalation	✓	NSM	NSM	NI	6
	Valve, Nozzle	Leak, Blockage			✓	NSM	NSM	NI	
	Tamper-proofing	Defeated			✓	⊗	⊗	⊗	
	Handle	Malfunctions			⊗	⊗	⊗	NI	
	Vessel	Explodes	Explosion	Laceration, Puncture, …	✓	NSM	⊗	⊗	
	Contents	Toxicity/Low Temp.	Chemical/Cold	Respiratory, Skin burn, Eye injury	⊗	NSM	⊗	⊗	
	Contents	Conductivity	Electrical	Electrocution	⊗	NSM	⊗	⊗	
	~~Vessel~~	~~Heavy/Repetitive~~	~~Strain~~	~~Back injury~~	TN	FP	TN	TN	
Bath toy	Sharp/Protrusion	Does not give way	Cut, Stab	Laceration, Puncture	✓	✓	✓	NSM	0
	Small parts	Detach/Break-off	Choking	Choking	✓	✓	✓	NSM	
	Materials/paint	Toxicity	Chemical	Poisoning, Skin irritation	✓	✓	✓	NSM	
	Surface	Becomes slippery	Slip, Fall	Injury from fall	✓	⊗	✓	⊗	
	Vessel	Collects/Ejects water	Drowning	Lung injury, Death	✓	⊗	NI	⊗	
	Seal, Various	Leak/Mold growth	Bacterial	Lung injury, Poisoning, Infection	✓	✓	✓	NSM	
	Seal/Coating/Wire	Leak/Exposed	Electrical	Electrical shock	✓	✓	✓	NSM	
	String/Ribbon/Cord	Does not give way	Entanglement	Strangulation (neck/limb), Pinch	⊗	⊗	✓	⊗	
	Suction Cup	Detach	Fail to operate	Fall/Trip Injury	⊗	⊗	✓	⊗	
	Unspecified	Entrapment/Suction	Entrapment	Injured body part, Suction bruise	⊗	✓	⊗	⊗	
	Battery (Acid)	Leak	Chemical	Burn	⊗	⊗	⊗	⊗	
Dehumidifier	Fan, Motor, Wires	Ignites	Fire	Burns, Smoke Inhalation	✓	NSM	✓	⊗	4
	Power cord, plug	Exposed current	Electrical	Electrocution	✓	NSM	⊗	⊗	
	Various parts	Overheating	Hot surface	Burn	✓	⊗	NI	⊗	
	Sensors	Malfunction	Fire	Burns, Smoke Inhalation	✓	⊗	NI	⊗	
	Hose, pan, pump	Water spillage	Electrical, Slip	Electrocution, Injury from fall	✓	⊗	NI	⊗	
	Various	Microbial growth	Bacterial	Lung injury	✓	NSM	✓	NSM	
	Coils	Refrigerant leak	Chemical	Lung injury	✓	NSM	✓	⊗	
	~~Water tank~~	~~Heavy/Repetitive~~	~~Strain~~	~~Back injury~~	TN	TN	FP	TN	

(Continues)

TABLE A1 (Continued)

Product Type	Component	Failure Mode	Hazard	Injury (Effect)	Generative-AI Tool				Total Human Expert Assessments for Product Type
					ChatGPT 3.5	ChatGPT 4	MSFT CoPilot	Perplexity	
Hammer	Head, Handle	Disintegrates	Projectile	Contusion, Laceration, …	✓	✓	NSM	✓	5
	Head	Decouples	Projectile	Contusion, Laceration, …	✓	⊗	NSM	✓	
	Handle	Lacks friction	Control loss	Contusion, Laceration, …	✓	⊗	NSM	⊗	
	Product	Chronic, Repetitive	Strain, Stress	Vibration Syndrome (HAVS), RSI	⊗	✓	NSM	⊗	
Window blind	Cord	Does not give way	Entrapment	Strangulation	✓	✓	NSM	✓	2
	Slats	Does not give way	Entrapment	Strangulation	✓	✓	⊗	⊗	
	Slats	Unexpected motion	Control loss	Contusion, Laceration, …	⊗	✓	NSM	⊗	
	Slats, Components	Exposed sharp	Sharp edges	Laceration	⊗	✓	⊗	⊗	
	Cords, Slats	Does not give way	Entanglement	Pinch, Cut (finger or limb)	✓	✓	NSM	⊗	
	Small parts	Detachment	Choking	Choking	✓	✓	⊗	✓	
	Materials	Toxicity	Chemical	Poisoning, Allergy	⊗	✓	⊗	⊗	
	Safety device	Malfunctions	Fail to operate	Strangulation, Pinch	✓	⊗	⊗	⊗	
Dry-erase board	Surface	Exposed sharp	Sharp edges	Laceration	✓	✓	✓	✓	7
	Mounting	Integrity fails	Tip-over	Laceration, Contusion, …	✓	✓	✓	✓	
	Board	Mounting error	Tip-over	Laceration, Contusion, …	✓	✓	✓	✓	
	Board	User interaction	Tip-over / Pinch	Laceration, Contusion, …	✓	⊗	⊗	⊗	
	Marker, Coating	Toxicity	Chemical	Lung injury, Poisoning, Burn, Eye	✓	✓	⊗	✓	
	Cleaning solution	Toxicity	Chemical	Lung injury, Poisoning, Burn, Eye	⊗	✓	✓	✓	
	Marker cap, Eraser	Small parts detach	Choking	Choking	⊗	✓	⊗	⊗	
	Marker cap	Forced ejection	Projectile	Eye injury	⊗	✓	⊗	⊗	
	Board	Excessive weight	Strain	Musculoskeletal injury	⊗	✓	⊗	⊗	
	~~Markers~~	~~Fumes ignite~~	~~Fire~~	~~Burn~~	FP	TN	TN	TN	
	~~Markers~~	~~Ghosting~~	~~Strain~~	~~Eye strain~~	TN	TN	FP	TN	

Notes: Chemical, fire, and electrical are top-level hazard types. All other hazard (sub)types above can be classed under the top-level category: mechanical. Chemical hazard type has sub-types: chemical inhalation, chemical ingestion, and chemical skin absorption, though for brevity these are not broken out.

Abbreviations: ✓ = "True Positive" (correctly recognized); ⊗ = False Negative (i.e., missing from Gen-AI tool response); **FP**,"False Positive" (also indicated with strikethrough font, as the failure mode is invalid). **NI**, failure mode noted, but no injury (effect) listed in FMEA; **NSM**, some numeric scores missing from FMEA; **TN**, "True Negative".

TABLE A2 Suggested mitigation tactics: ChatGPT versus other generative-AI tools.

PHASE	Mitigation Tactic	ChatGPT 3.5	ChatGPT 4	MSFT CoPilot	Perplexity†
DESIGN	Research and Development	NS	NS	NS	PS
	Choose "high-quality" and/or "durable" materials	NS	NS	PS	PS
	Choose non-toxic materials (for example, phthalate and BPA-free)	NS	PS	PS	NS
	Choose climate/operating-condition resistant materials (for example, for humidity, moisture, heat, sunlight, …)	⊗	PS	PS	⊗
	Standards compliance	PS	PS	NS	PS
	Design out hazard (Eliminate)	PS	PS	PS	PS
	Mandatory and voluntary warnings: on packaging	PS	PS	PS	⊗
	Mandatory and voluntary warnings: on actual product	PS	PS	PS	⊗
	Integrated protection for users (Guards)	PS	PS	PS	⊗
	Integrated physical impact protection for fragile product components (for example, gauges, valves, hoses, …)	⊗	⊗	PS	⊗
	Products have visible impact indicator of external trauma to product, or tamper-proof seals	⊗	PS	PS	⊗
	Integrated sensors or visual indicators to detect failures from product (for example, leak sensors; temp. sensors; clog indicators)	⊗	PS	PS	⊗
	Integrated actuators to prevent harm from product failure (for example, thermal fuses/cut-offs, shutoffs, pressure relief devices)	⊗	PS	PS	⊗
	Incorporate self-clearing feature (for example, design extinguisher hose to dislodge debris during use)	⊗	⊗	PS	⊗
	Design for user-friendly maintenance without specialized tools	⊗	PS	PS	⊗
BUILD	Manufacturer: Test materials quality, compliance	NS	PS	PS	PS
	Check manufacturing process quality	NS	PS	⊗	⊗
	Pre-sale Quality Control (QC) testing and certification	PS	PS	PS	PS
	Pre-sale usability testing and feedback	PS	PS	⊗	⊗

(Continues)

TABLE A2 (Continued)

PHASE	Mitigation Tactic	ChatGPT 3.5	ChatGPT 4	MSFT CoPilot	Perplexity†
OPERATE	Training & Requirements of installation service provider	⊗	⊗	⊗	NS
	Configure safe surroundings, such as:				
	- cribs away from blinds; cords out of reach of children	PS	PS	PS	PS
	- fire extinguishers out of direct sunlight;	⊗	⊗	PS	⊗
	- adequate ventilation for dehumidifiers	PS	PS	PS	PS
	Training of user, including public awareness & education	NS	PS	PS	PS
	Supervision of vulnerable users and bystanders (for example, kids, bystanders, coworkers)	NS	PS	PS	PS
	Regular inspection by user	NS	PS	PS	PS
	Mandated regular durability, reliability, and stress testing by manufacturer	⊗	PS	⊗	⊗
	Maintenance by user	PS	PS	PS	PS
	Shield product from damage (for example, direct sunlight, moisture/humidity, outdoors); provide product storage guidelines	⊗	PS	PS	⊗
	Static product replacement (aging) schedule for user	⊗	PS	PS	⊗
	Automatic alerts (smart products) that notify users when maintenance or replacement is due	⊗	PS	PS	⊗
	Feedback from users / Encourage users to report issues	NS	PS	PS	⊗
	Personal protective equipment (PPE)— for example, gloves, goggles	PS	PS	PS	PS
	Regulator: check product compliance with standards	⊗	PS	PS	PS
	Post-market surveillance, incident learnings, and regulatory reporting	PS	PS	PS*	⊗*
	Manufacturer product remediation, penalties, or product recall	PS	PS	PS*	⊗*

Abbreviations:

⊗ = False Negative (i.e. missing from Gen-AI tool response).

* When asked for "guidance for regulators" Perplexity misinterpreted "regulator" as a dehumidifier internal component (rather than a government agency responsible for regulations) and CoPilot suffered the same misinterpretation for fire extinguisher regulators where CoPilot gave some guidance referring to the regulator as a component, and some guidance referring to the regulator as an on-site person inspecting the fire extinguisher. This indicates prompt writers should be careful to avoid ambiguity: e.g., "dehumidifier regulators" would be better stated as "government regulators in charge of regulating dehumidifiers."

† When asked "what risk mitigations," Perplexity frequently suggested a generic process and general techniques the user can use with colleagues for finding risk mitigations (e.g., create a multidisciplinary team, use sticky notes to identify failures, use five whys, root cause analysis, flowcharting, fishbone diagrams, create action plan, …), but doesn't actually itemize any actual product-specific risk mitigations. Perplexity was more specific about product-specific risk mitigations when asked for "guidance for [the product] designer based on the failure modes," for example, suggesting ASTM F963 compliance verification for bath toys.

(Continues)

TABLE A2 (Continued)

CoPilot frequently included subliminal advertising for Spanish holiday destinations at the end of its output. These appeared briefly in the interactive output, then disappeared, but appeared permanently in the exported (downloaded) text of output.

NS = "Non-Specific (Generic) guidance." For example, did not cite specific standard (e.g., says "relevant standards" instead of citing specific standards), or did not give specific warning label contents, specific material specifications, specific inspection or user-feedback modality directives, or full user training directives.

PS = "Partially Specific guidance." Gen-AI tool gave some examples, but did not provide specific materials or measures or designs. For example, for Design out hazard (Eliminate) the GenAI tool suggested cordless blinds and break-aways and guards or soft-closing mechanisms for folding hinges of dry-erase boards and suggested threaded inserts, locking mechanisms, or bonding techniques that prevent accidental separation of hammer heads and non-slip textures or coatings for hammer grips; ChatGPT 4 suggested soft-closing mechanisms and flexible materials to prevent impact injuries in window blinds; CoPilot suggested shock absorbing features such as vibration-damping materials and anti-vibration coatings and inserts for hammers as integrated protection against Hand-Arm Vibration Syndrome (HAVS) and ChatGPT also suggested anti-vibration features; CoPilot suggested product impact indicators for fire extinguishers that suffered trauma to product; ChatGPT 4 suggested tamper-proof seals and frost-free horns and frost-free nozzles and insulated handles (to prevent cold burns) for fire extinguishers; ChatGPT 4 suggested hold-down brackets and safety stops and tension devices for window blinds to prevent injury from swinging or slamming in high winds or due to accidental forceful movement; ChatGPT 4 suggested large and non-detachable parts to avoid choking hazards; ChatGPT 4 and CoPilot suggested thermal overload protectors for dehumidifiers; CoPilot suggested safety cables for mounted dry erase boards to prevent tip-over and tempered or laminated safety glass for glass dry erase boards to prevent shattering; for Maintenance, GenAI tool suggested stop work and clean handle if oily; for PPE, suggested "gloves, goggles" and then, vaguely, "etc."; for Mandatory and Voluntary Warnings, GenAI tool suggested dry erase board hinges have instructions and warnings on moving parts, and suggested hammers have integrated QR codes on the products for access to online instructional videos and tutorials.

TABLE A3 Core expert-identified weaknesses in first iteration (zero-shot) prompts to ChatGPT.

First-iteration (zero-shot) general weakness	Specific examples from expert free-form comments
Missed mention of and assessment against specific standards	Did not cite or apply: National Fire Prevention Association (NFPA), California Proposition 65 (cancer or reproductive harm), Flammability standards, American Society for Testing and Materials (ASTM)†, Labeling of Hazardous Art Materials Act (LHAMA), Underwriters Laboratory (UL), Toxic Substances Control Act (TSCA), International Standards Organization (ISO), American National Standards Institute (ANSI), Voluntary vs. Mandatory standards.
Failure mode missed or confused with cause or effect	Slippery grip is not failure mode (Loss of hammer grip control is failure mode caused by slippery grip). Strangulation is effect (not failure mode). Eye injury is effect (not failure mode).
Does not make reasonable compromises	Asked regulator to do public awareness campaign. Suggested eye protection on hammer (unreasonable). Suggested fire extinguisher for dry erase marker fume ignition (unreasonable).
Poor estimates	Severity, likelihood scores unreasonable or unjustified: for example, severity for flooding (by dehumidifier) only given a 5 but is more serious. Severity score for same severity injury (strangulation) varied. Expert's own experience: pressure loss likelihood should be higher. Did not separate multiple effects for one failure mode‡. Detection score unreasonable: it is easier to detect sharp edge than inadequate mounting
Comments too general, lacking reference to environment or application, or specifics	Suggested personal protective equipment (PPE) (goggles, gloves, etc.) regardless of application: this is impractical and shows no knowledge of the application. Did not disentangle hammer types and application. E.g.: Failure modes should be specific to particular hammer product characteristics and to type of hammer; Suggested two-handed hammer holding regardless of hammer type (for example, tacking hammer vs. sledgehammer) and application. Ignored environment: E.g.: for dehumidifiers, specific numbers should be provided for size of the room and when to drain (based on room size and dehumidifier liquid output volume). Assumed workplace environment (rather than household). Missed usage issue—do not put children's cots, beds, highchairs etc near a window where children can reach the blind or curtain cords. "Use high quality materials" is too vague to be useful: specify. Suggested repairing leaky fire extinguisher pressure vessel with generic seals from hardware store which is not safe: replacement pressure vessel or custom (product-specific) valve is needed. Vague suggestion to "collaborate with international standards organizations"

(Continues)

TABLE A3 (Continued)

First-iteration (zero-shot) general weakness	Specific examples from expert free-form comments
Misjudged the audience, or the audience's responsibilities	Assumed user (not manufacturer) responsible for sharp edges. Did not disentangle quality (for example, lifespan, efficiency, usability) from safety failures. Assumed industrial (not household) setting: "report to supervisor" or "communicate with coworker" or factory inspections. Hyper-concerned about chronic injury (user over-exertion; long-term health), or antimicrobial properties (for example, of dry erase board surface). Muddled regulator (vs. manufacturer) responsibilities: e.g., Regulator would probably say responsibility falls on Manufacturer to launch public awareness campaign.
Output unreliable: Response varied when prompt repeated	Missing or muddled FMEA columns. Severity score for same severity injury (strangulation) varied. Sometimes did not suggest mandatory warnings. Sometimes did not provide key to allow FMEA likelihood, severity, or detection score interpretation

† ASTM (website banner warning) explicitly prohibits "the entry of ASTM standards and related ASTM intellectual property into any form of AI tool, such as ChatGPT.".
‡ Therefore, could not disentangle severity, likelihood and avoidability for each effect.

TABLE A4 Prompt engineering tactics and example prompt snippets for each tactic.

Prompt engineering tactics attempted					
Give precise, specific instruction	**Give input data (examples)**	**Give specific context**	**Give specific output format**	**Ask for reasoning steps**	**Force facts / Halt hallucinations**
"The FMEA should **focus on safety issues**."	"Identify different types of hammer, **such as tack hammers and sledgehammers** and others. For each hammer type, …"	"**Assume you are a hammer manufacturer** and want to manufacture and distribute safe hammers."	"**Tabulate** your results, being sure to **include columns for** component, failure mode, severity, likelihood, detection, and RPN, and being sure to **include a key describing** the score ranges in the table."	"**Let's think step by step**"	"**Do not provide answers if you are not certain**."
"Please **cite relevant safety compliance standards**."					
				"**Think aloud**"	"**Only provide answers with citations**."
"Tell me about some **industry standards for dehumidifiers**, that promote safety."	"Tell me about **some industry standards for dehumidifiers**, that promote safety" then follow-up prompt "Tell me specific specifications that the **standards above** mandate for dehumidifiers."				
"What are the parameters that a dehumidifier must operate within, according to each safety standard? **Give me specific parameter values** for each parameter type."	"Cite applicable mandatory and voluntary safety standards **like NFPA and California Proposition 65**, and specify the parameter values that a safe fire extinguisher must conform with, to satisfy these standards."				

TABLE A5 ChatGPT 3.5 continuing weaknesses encountered during prompt engineering.

Weakness Type	Examples
Truthfulness	When asked to cite specific standards, it sometimes hallucinated standard numbers and titles. For example, when asked "Cite applicable mandatory and voluntary safety standards pertaining to each hammer type" it starts with claw hammers and invents "Mandatory: ANSI/ASME B173.13-2015 American National Standard Specifications for Claw Hammers" and "Voluntary: ASTM F1568-94(2018) Standard Specification for Claw Hammers" which are fabricated and do not exist. It proceeds to invent fabricated standards for tack hammers, ball-peen hammers, and sledge hammers, each time containing an invalid standard number and the specific hammer type in their titles, making the standards titles seem convincing even though they are hallucinated. When adding "Do not provide answers if you are not certain", ChatGPT cites ASME B107.41-2012 and ISO 15601:2000. ASME B107.41 is valid standard, but the version number (2012) does not seem to exist, and there is a more recent standard (ASME B107.400-2018) that applies. ISO 15601:2000 is a valid standard.
Comprehensiveness	For fire extinguishers, when asked to "Cite applicable mandatory and voluntary safety standards like NFPA and California Proposition 65″ it initially misses applicable standards (for example, UL 711, OSHA 1910.157, BS EN 3, …) and must be specifically prompted "What other standards are applicable?" For hammers, it did not realize ANSI/ASME B107.54-2001 "Heavy Striking Tools—Safety Requirements" is applicable, even when asked specifically for safety standards for hammers. When asked to list hammer subtypes and provide FMEAs for each, it provides FMEA for only one hammer subtype at a time, even when asked to do multiple, possibly due to response length limitation. It instructs the user to repeat the process for the other listed hammer types.
Relevance	When asked to list specific standards some are irrelevant or rare outlier cases. for example, For dehumidifiers, it gives an edge-case standard for portable generators in case the dehumidifier is powered by a portable generator. When asked about fire extinguisher standards, it gives standards for horizontal directional drilling of polyethylene pipe, standards for determining relative humidity of concrete floors, standards for bunk beds (claiming their placement might impact fire extinguisher accessibility), and standards for fire resistive cabling. When asked to think step-by-step, it itemizes the functions of the dehumidifier during its thinking (Extract moisture from the air, collect condensed water, …), even though these aren't directly helpful (it reports all its thinking, even if some paths were dead-ends). When asked to do the FMEAs one hammer subtype at a time, it sometimes includes failure modes like "hammer peen deforming" or "head warping" which are quality issues (not safety issues) and don't seem to be a plausible cause of potential hand injury.
Reliability	When asked to do the FMEA's one hammer subtype at a time, it gives briefer (2-3 items, instead of 5–7 item) failure mode lists, possibly due to response length limitations. It frequently omits failure modes like "head separating" or "grip comes lose" or "material is toxic" (even though there have been historic recalls for these specific issues: "head of the sledgehammers can loosen prematurely and detach unexpectedly during use"; "The head of the sledge hammer can loosen and detach"; "the molded grip on the hammer can come loose"; "mallets … contain levels of lead that exceed the federal lead paint ban"). When asked to do the FMEA's one hammer subtype at a time, it sometimes omits failure modes like "head separating," and sometimes adds failure modes like "handle splintering," "head chipping." It gives inconsistent scores for severity for the same hazard (for example, Fire), even though Fire should have the same severity throughout. In new iterations for dehumidifiers, it forgot moisture accumulation leading to mold and respiratory illness. When asked to do FMEA broken down by component, it added [for dehumidifiers] cracked housing/enclosure (leading to electrocution), insufficient ventilation of housing/enclosure (leading to fire and burns), seizure of motor (leading to electrocution) and blade obstruction (leading to cuts), and it added [for fire extinguishers] pressure gauge incorrect reading, hose leakage, valve sticking, and extinguishing-agent caking, though it wasn't concerned about those issues when originally asked to do FMEA without specific guidance to break down by component. When asked to provide a key to help interpret the scores, it decided a scale of 1–5 was good for severity, likelihood, and detection (previously it used a 1–10 range). It had to be pressed, in an additional iteration, to provide those scores on a 1–10 scale.
Vagueness	When asked what parameters apply to a product type according to each safety standard, it provided only parameter types; had to ask for specific parameter values for each parameter type in order to get specific parameter upper and lower limit values. Often responses "within specified parameters" instead of provided the numeric parameter lower and upper bounds, "within specified parameters" is too vague to be actionable. When adding "Do not provide answers if you are not certain," ChatGPT outputs only two specific parameter values ("Handle Material: Must withstand a force of at least 500 lbs without fracturing. Head Material: Must be securely attached to the handle, with no detachment during testing under 1000 lbs of force.") even thought the standard is more specific, likely due to output length restrictions for ChatGPT. When adding "Do not provide answers if you are not certain" and "Only provide answers with citations" together, ChatGPT doesn't actually provide any specific citations, and instead kicks the can down the road, saying it will need to research standards from ANSI, ASME, ISO, and others: "We'll need to research and cite mandatory and voluntary safety standards pertaining to each hammer type. This will depend on the region and industry standards. Common standards bodies include ANSI (American National Standards Institute), ASME (American Society of Mechanical Engineers), ISO (International Organization for Standardization), and others." So, the engineered prompt asking for citations seems to make the response worse, with ChatGPT effectively parroting the question back to the end-user.

(Continues)

TABLE A5 (Continued)

Weakness Type	Examples
Logical errors	When asked to think step-by-step, in its bulleted thinking it says cracked housing/enclosure leads to "electrical exposure" but in the tabulated results it refers to cracked housing/enclosure as a mechanical hazard (not an electrical hazard), and says resultant injury is "None" even though the resultant injury would be electrocution (for electrical exposure), laceration (for sharp edge in crack), or even perhaps lung injury (if crack was in refrigerant enclosure or refrigerant canister or coils and led to refrigerant leak). Even when asked to focus on safety hazards it generated some rows that specify "None" as Hazard, and "None" as injury. Strangely, it gives these "None" injuries severity scores that range from 3 to 6 (on a scale of 1 to 10), as well as variable likelihood (5 or 6 out of 10) and detection (7 or 8 out of 10) scores, even though the Hazard and Injury are "None." Says "refrigerant leakage" is "rarely detectable" even though refrigerant has an odor of ether, chloroform, or sweetness (according to US Dept of Energy) and this odor is often noticeable. Claims users will notice compressor failure by "sudden lack of dehumidification, though it is more rapidly noticed from sudden lack of compressor operating noise.
Productivity	User needs to try multiple prompt variations, each with somewhat unpredictable output, and the task of figuring out what is missing or new or untrue or irrelevant in the large volumes of ChatGPT output, becomes almost as burdensome as manually doing the FMEA from scratch.